NNT : 2016SACLE013

THESE DE DOCTORAT
DE L'UNIVERSITE PARIS-SACLAY,
Préparée à "l'Université d'Evry Val d'Essonne"

ÉCOLE DOCTORALE N° (577)
Structure et dynamique des systèmes vivants

Spécialité de doctorat : Bioinformatique

Par

Mme Rim ZAAG

Enrichissement de profils transcriptomiques par intégration de données hétérogènes : annotation fonctionnelle de gènes d'*Arabidopsis thaliana* impliqués dans la réponse aux stress.

Thèse présentée et soutenue à Orsay le 20 Juin 2016:

Composition du Jury :

Mme F. Monéger	Directrice de recherche, CNRS	Présidente du Jury
Mr M. Elati	Maître de conférence, Université d'Evry	Examinateur
Mme B. Mangin	Directrice de recherche INRA	Rapporteur
Mr P. Meyer	Professeur, Université de Liège	Rapporteur
Mme ML. Martin-Magniette	Directrice de recherche, INRA	Directrice de thèse
Mr E. Delannoy	Chargé de recherche, INRA	Co-directeur de thèse

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Remerciements

Mes remerciements s'adressent à toutes les personnes qui m'ont aidée à faire avancer mon projet de thèse, mais aussi à toutes celles qui m'ont soutenue, qui m'ont encouragée et m'ont aidée à surmonter les moments difficiles.

Mes premiers remerciements vont à mes directeurs de thèse :

D'abord Marie-Laure Martin-Magniette qui m'a accompagnée tout au long de cette expérience. Merci pour tout le temps que tu m'as consacré, tes conseils, ton écoute et ton dynamisme. Merci de m'avoir poussée toujours de l'avant, de m'avoir aidée à dépasser mes limites, mes peurs avec tous les exercices que tu m'as fait faire et notamment toutes les présentations un peu partout ! Ce n'était pas un exercice facile pour moi au début mais grâce à ton obstination à m'envoyer présenter mes travaux de thèse à toutes les conférences possibles, je vais pouvoir être (plus) sereine à ma soutenance de thèse :)

Un grand merci à Etienne Delannoy, qui a co-encadré ce travail depuis ma deuxième année de thèse. Merci pour ta patience, tes conseils, ta disponibilité et pour toutes les discussions enrichissantes. C'est toujours un réel plaisir de discuter avec toi :) Merci pour ta bonne humeur et pour les gâteaux aux moments difficiles !! Merci à toi et à Marie-Laure pour les innombrables relectures à la fois promptes et minutieuses que ce soit pour le manuscrit de thèse ou pour les résumés de soumissions!! Désolée de vous avoir fait travailler pendant vos vacances et vos trajets de déplacement :)

Je remercie également Sébastien Aubourg, mon directeur de thèse pendant la première année. Merci pour la confiance que tu m'as accordée, d'avoir cru en moi et de m'avoir proposé cette thèse avec Marie-Laure. Merci aussi de m'avoir accueillie au sein de l'équipe et de m'avoir encadré depuis mon stage de master 2. Merci pour tes conseils et la personne sincère que tu es.

Je tiens à exprimer mes remerciements aux membres du jury, qui ont accepté d'évaluer mon travail de thèse. Je remercie Madame Brigitte Mangin, Directrice de recherche au Département Mathématique et Informatique Appliquées à l'INRA et Monsieur Patrick Meyer, Professeur à l'université de Liège de m'avoir fait l'honneur d'être les rapporteurs de cette thèse. Merci à Madame Françoise Monéger, Directrice de recherche CNRS et Monsieur Mohamed Elati, maître de conférence à l'université d'Evry, qui ont accepté de faire partie du jury. Je vous remercie tous de m'avoir fait l'honneur d'assister à ma soutenance.

Je remercie aussi les membres de mon comité de thèse Madame Gaelle Lelandais, Monsieur Eric Ruelland et Monsieur Jean Colcombet avec qui j'ai eu des discussions très intéressantes et très enrichissantes. Merci d'avoir contribué à l'amélioration de cette thèse par vos remarques et suggestions pertinentes.

Je remercie toutes les personnes que j'ai pu côtoyer lors de mes activités d'enseignement en tant que moniteur à l'université d'Evry, en particulier Madame Valérie Chaudru. Je remercie

les collaborateurs avec qui j'ai eu la chance de travailler, je remercie en particulier Nicolas Frei Dit Frey pour la collaboration fructueuse que nous avons eu.

Un grand merci à tous les membres de l'équipe Genomic Networks pour leur accueil chaleureux depuis les premiers jours de mon stage de master 2, il y a maintenant quelques années :) eh oui depuis Janvier 2011 ! Merci à tous pour votre soutien et votre amitié. Merci Cécile d'avoir pris soin de mon bébé GEM2Net, pour ton aide afin de dénicher les fautes d'orthographe dans mon manuscrit de thèse, et pour le soutien technique lors de la rédaction des mails sérieux :) Merci pour ta bienveillance! Tu sais que je te redoute beaucoup aux moments où le moral n'est pas au top car tu me démasques dès le premier coup d'œil !!! Merci pour les séances de soutien moral. De même pour Véro, merci pour ton soutien, tes encouragements et tes conseils. Merci pour ton humour, le bureau n'est pas le même quand tu n'y es pas !! Merci à toi et à Jean-Philippe pour vos relectures de mon manuscrit de thèse :) Eh oui j'ai fait relire ma thèse à toute l'équipe :) Merci Jean-Philippe pour tout: pour FLAGdb, pour tes plats si délicieux !! Merci Zakia, d'abord pour tout le travail que tu as fait pour l'interface GEM2Net, je t'ai bien pris la tête avec tous les détails sur lesquels je pinaillais :) merci pour ton aide et ton appui technique lors de la mise en page du manuscrit !! Je n'oublierai jamais :) pour tes gâteaux et surtout le roulé :) bon courage pour ton parcours professionnel. Merci Christine pour ton soutien et pour les discussions sur le trajet de retour en covoiturage :) merci Philippe pour ton soutien précieux, merci Guillem pour tes conseils statistiques avisés, merci Nathalie pour ta gentillesse. Merci Nadia pour ton soutien même à distance!! Merci à toute l'équipe, travailler avec vous tous est un réel plaisir tant au niveau humain que scientifique.

Ma plus profonde gratitude va à mes parents qui m'ont toujours soutenue et cru en moi. Qu'ils trouvent dans ce diplôme qui était leur rêve avant qu'il soit le mien, l'aboutissement de leurs efforts ainsi que l'expression de ma très grande reconnaissance. Merci à toute ma famille, mes frères, mon mari et surtout ma fille. Ma chère et douce Fifi !! C′est sans doute la personne qui m'a le plus soutenu dans cette expérience et surtout les derniers mois !! Elle a spontanément et volontairement pris beaucoup de responsabilité pour combler mon absence auprès de son petit frère. Elle s'est substituée en une formidable petite maman pour lui!! Merci Firyel pour tout cela, pour les câlins et les bisous en guise de soutien et de réconfort, pour les massages avec tes petites mains lorsque maman est fatiguée et a mal au dos !!! Merci pour les gâteaux :) Je suis vraiment fière de toi !!

« Qui n’aime pas gravir la montagne,
vivra éternellement au fond des vallées »
Abou El Kacem Chebbi

Sommaire

Table des tableaux

Table des figures

Introduction

I. Les annotations d'un génome

La molécule d'ADN ou Acide Désoxyribonucléique, découverte en 1953 par Watson et Crick, est le support de l'information génétique. Les avancées technologiques notamment en séquençage à haut débit (Schena *et al.* 1995 ; Wang *et al.* 2009) ont permis le séquençage complet d'un grand nombre de génomes. Cependant le séquençage n'est pas un objectif en lui-même mais un moyen pour déchiffrer les secrets du vivant. En effet, un des défis majeurs de la génomique est la traduction de ces séquences brutes en informations utiles autour des entités génétiques clés responsables des propriétés biologiques de l'organisme. C'est ce qu'on appelle l'annotation du génome. Elle est constituée de deux tâches principales. La première consiste à localiser topologiquement les séquences informatives notamment les gènes et les signaux régulateurs associés. Ce processus est appelé annotation structurale ou syntaxique. La deuxième tâche correspond à l'identification ou la prédiction de la fonction des produits de ces séquences. Cette étape correspond à l'annotation fonctionnelle.

La quantité de données et la complexité de l'organisation des génomes notamment chez les eucaryotes rendent ces annotations fastidieuses et quasiment impossibles manuellement. L'annotation des génomes repose donc essentiellement sur des méthodes d'annotation *in silico* qui permettent d'automatiser les différentes opérations requises. Les méthodes d'annotation structurale peuvent être classées en deux types d'approches: les méthodes dites *ab initio* ou *intrinsèques* qui reposent sur la recherche de signaux intrinsèques à la séquence étudiée et les méthodes comparatives ou extrinsèques qui reposent sur la recherche de similarité à d'autres séquences. Certains outils tels qu'EuGene (Sallet *et al.* 2014) intègrent les deux approches pour améliorer la qualité de leurs prédictions. Ces méthodes peuvent être complétées par une expertise manuelle afin d'éviter la propagation d'erreurs.

En partant de l'information statique donnée par l'analyse structurale des séquences génomiques, l'annotation fonctionnelle analyse l'aspect dynamique de cette information telle que la transcription des gènes, la traduction et les interactions entre leurs produits ou leur activité enzymatique. L'objectif principal de l'annotation fonctionnelle d'un génome est donc l'identification ou la prédiction de la fonction des produits de ses gènes afin de comprendre leurs rôles dans les

phénotypes observés. Les produits des gènes tels que les protéines ou les enzymes sont les agents actifs responsables d'une tâche particulière. Cependant, par souci de simplicité, dans la suite de ce manuscrit nous parlerons de la fonction des gènes au lieu de la fonction des produits des gènes.

La principale difficulté de l'annotation fonctionnelle réside dans le fait que la fonction d'un gène est dépendante du contexte et de la question biologique posée. Un biochimiste par exemple cherche à caractériser les protéines à travers leurs activités et leurs quantités dans la cellule, quand les généticiens s'intéressent aux phénotypes associés aux mutations des gènes étudiés. Ainsi, un gène peut assurer une activité biochimique bien précise au sein de la cellule telle qu'une activité kinase et en même temps, il peut avoir une localisation subcellulaire ou des propriétés physiologiques particulières lui permettant de participer à un processus biologique bien déterminé tel que la photosynthèse. Ainsi le terme fonction fait référence à un ensemble de fonctions moléculaires, de localisations cellulaires, de domaines fonctionnels, de voies métaboliques, de signaux de localisation, ou toute autre caractéristique et il est donc plus pertinent de parler des fonctions d'un gène.

Ce genre de problématique a aidé à entreprendre des efforts de standardisation afin d'avoir un langage uni ou un vocabulaire contrôlé qui permet de décrire les gènes et les informations associées de la même manière entre les différentes communautés scientifiques et pour les différentes espèces étudiées. C'est le cas du projet Gene Ontology ou GO (Ashburner *et al.* 2000) qui met à disposition un ensemble de termes décrivant les gènes et leurs produits grâce à trois domaines ou ontologies indépendantes qui sont : « Fonction Moléculaire » décrivant les activités biochimiques des protéines telles qu'une activité kinase, « Processus Biologique » décrivant le processus cellulaire dans lequel s'inscrit l'activité de la protéine autrement dit l'objectif de cette activité tel que la réponse au stress, et enfin « Localisation Cellulaire » déterminant le compartiment subcellulaire dans lequel se trouve la protéine majoritairement. Dans la suite du manuscrit, ces trois ontologies seront notées MF, BP et CC.

L'annotation fonctionnelle continue d'être un défi majeur de la génomique. En effet, selon différentes estimations, 20 à 40 % des gènes prédits des organismes eucaryotes dont le génome est complètement séquencé, n'ont aucune fonction attribuée (Wortman *et al.* 2003 ; Gollery *et al.* 2006, 2007 ; Hanson *et al.* 2010). Concernant la plante modèle *Arabidopsis thaliana,* séquencée en 2000, la fonction d'environ 5 000 gènes soit environ 19% des gènes annotés (Zaag *et al.* 2015) reste totalement inconnue selon la dernière version de l'annotation officielle TAIR10 de novembre 2010 (www.arabidopsis.org) (Garcia-Hernandez *et al.* 2002). Ces gènes pour lesquels on ne dispose

d'aucun indice autour de leurs fonctions potentielles sont appelés gènes orphelins de fonction (Domazet-Lozo and Tautz, 2003; Fukushi and Nishikawa, 2003) et sont systématiquement mis de coté dans les analyses de génétique inverse ou autres approches fonctionnelles dirigées. A côté de ces gènes orphelins il existe aussi un nombre très important de gènes partiellement connus c'est-à-dire dont on ne connait qu'une partie de leurs fonctions ou annotations GO. C'est le cas par exemple des gènes ayant une localisation cellulaire ou une annotation concernant le processus biologique dans lequel ils sont impliqués mais dont on ne connait pas leurs fonctions moléculaires. Chez *A. thaliana,* sur 34 042 gènes prédits, le nombre de gènes sans annotation GO pour les ontologies BP, MF ou CC est respectivement de 8 211, 8 644 et 7 273, auxquels s'ajoutent respectivement 7 790, 8 725 et 2 516 gènes dont l'annotation est « unknown ». Ces gènes orphelins ou mal caractérisés constituent un verrou important dans la compréhension et la reconstruction des réseaux génomiques mis en place pour assurer les fonctions de base et les réponses à des conditions spécifiques en vue de s'adapter à des modifications de leur environnement.

L'annotation fonctionnelle à haut débit repose essentiellement sur la bioinformatique et les méthodes automatiques. En effet, les approches expérimentales dans un cadre « gène par gène » telles que la génétique inverse sont fastidieuses et demandent beaucoup de temps. Par exemple, il est estimé que la caractérisation expérimentale de nouveaux gènes chez *Escherichia coli* s'effectue à un rythme de 20 à 30 gènes par an (Kolker *et al.* 2004).

Les méthodes d'annotation fonctionnelle *in silico* se déroulent généralement en deux étapes principales. La première correspond à la recherche de gènes proches, ayant potentiellement des fonctions similaires, et ensuite une étape d'inférence de fonction qui correspond à la propagation des fonctions des gènes connus aux gènes inconnus étudiés. Pour la recherche des gènes proches, les méthodes d'annotation fonctionnelle peuvent être classées en deux catégories principales : d'une part les méthodes fondées sur la recherche de similarité de séquence et d'autre part les méthodes fondées sur la recherche des partenaires des gènes étudiés. Dans ce dernier cas, l'obtention des groupes de gènes partenaires peut passer par l'intégration de plusieurs types de données biologiques qui sont en général hétérogènes. Pour les deux catégories, nous décrirons le principe, les méthodes et leurs limites.

II. Annotation fonctionnelle par similarité de séquence

1. Principe

Les méthodes fondées sur la recherche de similarité de séquences cherchent à identifier des similarités structurales susceptibles d'apporter des informations précieuses autour des relations entre les séquences, leurs structures mais aussi leurs fonctions. Le principe est de comparer les séquences des gènes à annoter à celles des gènes dont on connait la fonction afin d'identifier une similarité de séquence, de structure ou de motifs qui suggère que les gènes partagent la même fonction biologique. La similarité peut être calculée entre les séquences d'espèces différentes révélant la présence d'un ancêtre commun qui est à l'origine des séquences en question (Dolinski et Botstein 2007; Hulsen *et al.* 2006). La similarité peut aussi être calculée au sein d'une même espèce. Dans ce cas il s'agit généralement d'un processus de duplication à l'origine de ces séquences structuralement proches qui conservent plus ou moins la même fonction.

2. Méthodes

a. *Similarité de séquence primaire*

Des algorithmes de programmation dynamique tels que l'algorithme de Needleman et Wunsch d'alignement global (Needleman et Wunsch 1970) ou l'algorithme de Smith et Waterman d'alignement local (Smith et Waterman 1981) permettent la comparaison d'une séquence d'intérêt avec toutes les séquences connues dans les banques de données. Par contre ces algorithmes d'alignement optimal sont très couteux en temps de calcul. Des algorithmes heuristiques c'est-à-dire qui donnent des solutions approchées ont été alors développés afin de réduire les temps de calcul. L'idée est de passer par des étapes sélectives qui permettent de restreindre le calcul de l'alignement optimal à seulement des régions dans les séquences présentant les plus fortes similarités, ou contre seulement un sous-ensemble de séquences de la banque potentiellement plus significatif. Un des outils les plus utilisés est le logiciel BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool, Altschul et Gish 1996), qui permet l'alignement et la comparaison d'une séquence nucléotidique ou protéique contre toutes les séquences d'une banque de données choisie. Un score et une e-value sont fournis pour donner une indication sur le degré et la fiabilité de la similarité avec la séquence retrouvée.

D’autres méthodes se focalisent sur la recherche de similarité de courtes séquences au niveau de la séquence protéique, appelées domaines fonctionnels. Ces domaines sont répartis en familles et les membres d’une même famille partagent souvent au moins une fonction biochimique commune. Les différentes combinaisons et coopérations de ces domaines sont responsables de la variabilité des fonctions biologiques (Bashton et Chothia 2007). Plusieurs outils et bases de données sont disponibles pour le stockage et la recherche de domaines tels que Pfam (Bateman *et al.* 2000), PRODOM (Corpet *et al.* 2000), Prosite (Sigrist *et al.* 2010) et SCOP5 (Andreeva *et al.* 2004).

b. *Similarité de structure*

L’activité biochimique d’une protéine est directement liée au repliement de sa chaîne polypeptidique dans l'espace. En effet, par différents types de liaisons, les chaînes polypeptidiques se replient sur elles-mêmes et adoptent ainsi une conformation secondaire puis tertiaire qui leur est propre et qui est dépendante de leur séquence. Le repliement de la protéine peut engendrer plusieurs types de structures secondaires dont les principales sont les hélices a et les feuillets p. Nous pouvons distinguer également certaines configurations types de la structure tertiaire telles que les doigts de zinc, caractéristiques des protéines fixant l’ADN comme les facteurs de transcription. L’identification et la localisation de ces structures dans la séquence des protéines apportent des indices sur la fonction de la protéine étudiée car des protéines présentant des structures tertiaires similaires partagent souvent la même activité biochimique.

Ces structures sont déterminées expérimentalement par cristallographie ou spectrométrie qui sont des méthodes très laborieuses et coûteuses. Plusieurs méthodes bioinformatiques ont été proposées afin de prédire les structures secondaires et tertiaires d'une protéine sur la base de sa séquence en acides aminés. Bien que la structure tertiaire d'une protéine soit très informative sur sa fonction, sa prédiction est très difficile à cause du nombre considérable de possibilités de repliement à partir de la séquence primaire. De plus, la quantité des données expérimentales de structures reste très limitée surtout pour les génomes de plantes (Rhee et Mutwil 2014). La base de données PDB (Bernstein *et al.* 1977) est considérée comme l’unique référence pour les structures tridimensionnelles des protéines (Velankar *et al.* 2012). Elle permet l’accès aux descriptions et aux structures tridimensionnelles de macromolécules essentiellement des protéines identifiées expérimentalement et déposées directement par les scientifiques.

3. Limites de l'analyse par similarité de séquence

Les procédures d'annotation fonctionnelle fondées sur la similarité de séquences ont atteint leurs limites pour de nombreuses raisons. Premièrement la détermination de la similarité de fonction s'appuie sur la détermination de ressemblance ou le pourcentage d'identité entre les séquences comparées. Mais le seuil de similarité à considérer peut être ambigu (Médigue *et al.* 2002). Plusieurs études considèrent que le seuil minimal d'identité est de 30% alors que d'autres prennent un seuil à 40%. La considération de seuils limites peut conduire au transfert de la fonction à une autre séquence à tort, au risque de propager encore cette erreur à d'autres séquences en se basant sur le même principe.

Deuxièmement, une forte similarité de séquence ne garantit pas forcément une similarité de fonction. Tian *et al.* (2003) montrent que même avec une identité de séquence à 60%, dans 10% des cas la fonction inférée est incorrecte. L'étude de Rost (2002) montre également que des séquences très proches peuvent ne pas avoir la même fonction malgré une e-value significative de BLAST à 10^{-50} ou moins. A l'inverse des protéines ayant des séquences très éloignées peuvent partager la même fonction (Galperin *et al.* 1998).

Troisièmement une étude comparative des méthodes existantes d'annotation fonctionnelle a permis de montrer que les méthodes qui utilisent uniquement l'information de la séquence ont une faible précision et dépendent de la qualité des séquences disponibles (Radivojac *et al.* 2013). Les auteurs attirent l'attention également sur le fait que ces méthodes telles que BLAST renseignent généralement sur la fonction moléculaire des protéines mais sont inefficaces quant à la prédiction des termes fonctionnels reliés à l'ontologie BP. Ils expliquent que cela est peut-être dû au fait que des homologues ayant la même fonction moléculaire peuvent être impliqués dans des rôles biologiques différents dans différents tissus et organismes (Nehrt *et al.* 2011). Il est donc supposé que la similarité de séquence informe sur la fonction biochimique mais pas sur la fonction biologique.

Enfin, les gènes orphelins sont caractérisés par l'absence d'homologues ayant des fonctions connues. Sans référence, l'inférence fonctionnelle par similarité de séquence ne peut être exploitée.

Cela conduit donc à la conclusion qu'il est nécessaire de se tourner vers d'autres méthodes d'annotation fonctionnelle.

III. Annotation fonctionnelle par recherche de partenaires

1. Principe

Afin de compléter l'information et de palier aux limites de l'annotation fonctionnelle par similarité de séquence, d'autres méthodes ont été développées en se fondant sur le concept de voisinage (Dandekar *et al.* 1998). En effet, généralement pour assurer une fonction particulière un gène nécessite un ensemble de partenaires qui vont répondre de manière coordonnée. Ces groupes de gènes peuvent être impliqués dans un même complexe protéique ou une même voie métabolique (Mushegian et Koonin 1996 ; Tamames *et al.* 1997) et peuvent être corégulés afin d'assurer un processus moléculaire. L'objectif est donc d'identifier et de regrouper les gènes partenaires en cherchant des caractéristiques les reliant. Une fois les partenaires identifiés, et étant donné que les gènes du même groupe sont supposés partager la même fonction biologique, la fonction des gènes orphelins au sein d'un groupe est prédite sur la base de la fonction des gènes connus. C'est le principe de culpabilité par association connue généralement sous le terme anglais « guilt by association ».

Nous pouvons classer les méthodes de recherche de groupes de gènes partenaires en deux catégories. La première exploite les données structurelles et notamment la proximité physique des gènes. La seconde est basée sur la recherche de partenaires reliés par des interactions de causalité que ce soit de manière directe ou indirecte. Nous détaillerons pour chaque catégorie, les principales méthodes, les données sur lesquelles elles s'appuient ainsi que leurs limites.

2. Méthodes fondées sur l'exploitation des données structurelles

a. *Fusion des gènes*

La fusion des gènes est un évènement d'évolution par lequel deux protéines distinctes A et B dans une espèce fusionnent en une seule protéine AB dans une autre espèce (Marcotte *et al.* 1999a). Ces 2 gènes ont généralement des fonctions très proches (Marcotte et Marcotte, 2002) et peuvent même interagir physiquement (Enright *et al.* 1999). Une méthode d'annotation dite méthode de la pierre de Rosette, repose sur l'identification de ces gènes ayant fusionné (Suhre et Claverie 2004). Malgré le

potentiel de ces informations pour l'amélioration des connaissances fonctionnelles des gènes, leur disponibilité représente la limite majeure de leur utilisation.

b. *Localisation chromosomique : synthénie*

L'observation de l'organisation des gènes procaryotes en opérons est à l'origine du transfert de fonction entre les gènes proches physiquement (Overbeek *et al.* 1999). En effet les gènes au sein du même opéron sont co-transcrits grâce au même facteur de transcription qui vient se lier à une séquence régulatrice d'un promoteur commun. Ces gènes sont donc corégulés afin d'assurer un processus biologique particulier et la fonction d'un gène inconnu au sein d'un opéron est inférée à partir de la fonction des autres gènes connus de cet opéron.

L'étude de la conservation de la proximité des gènes a conduit à définir la notion de synthénie. Elle est liée à trois critères principaux : similitude de séquence, conservation de l'ordre des gènes sur les chromosomes des différents espèces et distance entre les gènes (Von Mering *et al.* 2003). La caractérisation des groupes de synthénie a été utilisée pour proposer des fonctions aux gènes inconnus au sein de ces groupes, sur la base de la comparaison de la conservation, de la localisation et de l'ordre de ces gènes dans d'autres espèces mieux caractérisées (voir figure1). Cette propriété a été exploitée dans certains outils tels que WIT (Overbeek *et al.* 2002) et STRING (Snel *et al.* 2000).

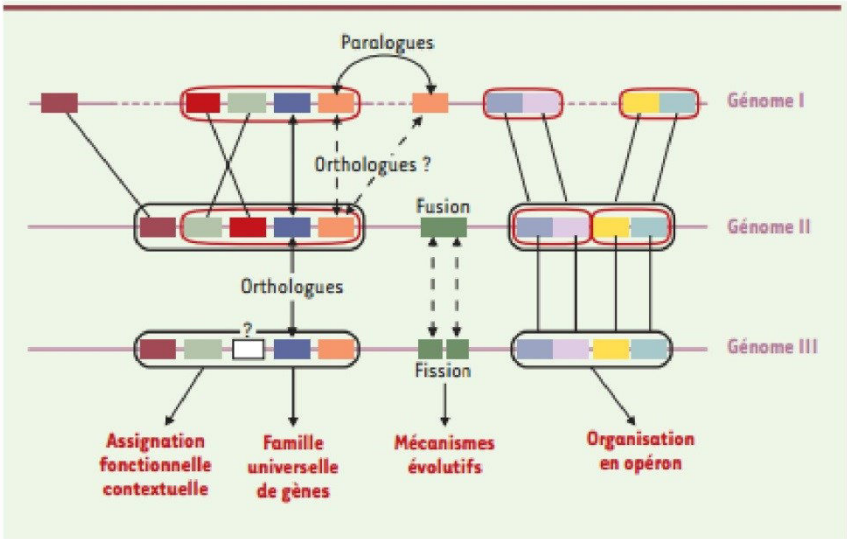


Figure 1 : Comparaison de la conservation de la localisation et de l'ordre des gènes entre les espèces.
La figure illustre la comparaison du génome III aux génomes I et II. Cette comparaison révèle que le premier ensemble constitue de cinq gènes contigus du génome III, contient 4 gènes dont les orthologues ont une même organisation locale dans le génome II, et une organisation proche dans le génome I. Les auteurs supposent donc que cet ensemble de gènes représente une famille de gènes universelle et que le gène situé au centre de cette famille (en

blanc) a probablement une fonction proche du gène rouge situe au même endroit dans le génome II. Figure extraite de Medigue et al. 2002.

c. *Recherche de motifs régulateurs*

En étendant aux eucaryotes le principe d'opéron, on peut supposer que les gènes impliqués dans les mêmes processus biologiques sont généralement régulés par les mêmes facteurs de transcription qui sont capables de reconnaitre et de se fixer sur une région précise du promoteur de ces gènes. Ces régions définissent des motifs régulateurs bien conservés bien qu'ils soient souvent dégénérés c'est-à-dire admettant une variabilité à certaines positions. Ce sont généralement des séquences de 4 à 15 paires de bases localisées dans la région promotrice des gènes, préférentiellement en amont du site du début de transcription (TSS) et sont susceptibles d'être impliqués dans la régulation de l'expression des gènes. La recherche de tels motifs au sein des séquences étudiées est très complexe à cause de la dégénérescence de ces motifs et leur nombre mais très répandue, car ils permettent de donner une indication sur la fonction ou le processus biologique dans lequel ces gènes sont impliqués. En effet, les gènes ayant les mêmes motifs au sein de leurs promoteurs sont supposés être des partenaires fonctionnels puisqu' ils sont régulés par les mêmes facteurs de transcription. Les gènes inconnus au sein de tels groupes se voient alors attribuer la ou les fonctions associées à ces motifs. Par exemple les séquences promotrices ayant les motifs cis-régulateurs contenant le cœur invariant TGAC sont connues sous le nom des éléments Wbox, essentiels pour la liaison et la fonction des facteurs de transcription WRKY. Les gènes associés à de tels motifs chez les plantes sont associés à la défense et à la réponse aux infections par des pathogènes et aux stress (Eulgem *et al.* 2000). Les bases de données PLACE (Higo *et al.* 1998) et AGRIS (Davuluri *et al.* 2003) sont les plus utilisées pour recenser et décrire les motifs connus chez les plantes.

3. Méthodes fondées sur l'exploitation des données d'interactions moléculaires

Les approches fondées sur la recherche de partenaires fonctionnels exploitant le voisinage structural ont permis de renforcer la validité des prédictions fonctionnelles dans plusieurs études (Marcotte *et al,* 1999a). Cependant, ces approches restent dépendantes de la qualité des séquences analysées et la disponibilité des données sur lesquelles s'appuient ces approches est dépendante des organismes étudiés. C'est pourquoi d'autres méthodes ont cherché à exploiter les interactions moléculaires pour la recherche de partenaires.

Le fonctionnement de la cellule nécessite l'interaction des différents objets génomiques (protéines, métabolites, enzymes, éléments de régulation etc.) qui permettent la mise en place des réseaux biologiques et la formation de complexes protéiques responsables des réponses adaptées aux stimuli de l'environnement et aux besoins vitaux. L'identification des interactions entre les acteurs renseigne donc sur le lien fonctionnel entre eux et sur le processus moléculaire et biologique dans lequel ils sont impliqués (Vidal *et al.* 2011). Ainsi la fonction d'un gène inconnu peut être étudiée à travers l'identification de ses voisins ou partenaires par les différentes interactions moléculaires qu'il peut avoir dans la cellule. Ces interactions peuvent être de nature différente : des interactions directes qui correspondent à des relations physiques de type interactions protéine-protéine ou protéine-ADN et des interactions indirectes qui correspondent à des relations fonctionnelles telles que des gènes impliqués dans une même voie de signalisation.

a. *Réseaux d'interactions protéine-protéine*

Les interactions protéine-protéine (PPI) correspondent à l'interaction physique entre les protéines pour former des complexes protéiques assurant une fonction moléculaire bien spécifique. Des protéines appartenant au même complexe protéique participent donc très probablement au même processus biologique.

Ces interactions sont identifiées par des méthodes expérimentales *in vivo* ou in *vitro* telles que : les puces protéiques, la technique du double hybride dans la levure (Yeast-two hybrid Y2H), le TAP-tag (Tandem Affinity Purification), la surface plasmon resonance, PCA (Protein-fragment Complementation Assay) et de nombreuses autres techniques expérimentales. La revue de Lalonde *et al.* (2008) permet de détailler toutes ces techniques ainsi que leurs avantages et limites.

Des outils bioinformatiques existent aussi pour la prédiction des PPI à partir d'informations telles que la proximité des gènes sur le chromosome, la recherche de l'existence d'un évènement de fusion entre deux gènes, la coexpression des gènes ou encore la co-citation des gènes dans la littérature. La majorité de ces outils utilise des méthodes d'apprentissage telles que les réseaux de neurones, les méthodes de classification bayésiennes, les régressions linéaires et les SVM (Séparateurs à Vastes Marges).

Toutes ces techniques expérimentales et méthodes bioinformatiques ont permis d'augmenter de manière importante la quantité de données PPI disponible, d'où la mise en place de nombreuses bases de données dédiées à la mise à disposition, le partage et la visualisation de ces informations

par l'ensemble de la communauté telles qu'IntAct (Kerrien *et al.* 2012), DIP (Xenarios *et al.* 2000), BIND (Willis et Hogue 2006), MINT (Licata *et al.* 2012) et BioGRID (Stark *et al.* 2011). D'autres bases de données se sont spécialisées dans les données prédites telles que les bases STRING (Szklarczyk *et al.* 2011) et PAIR (Lin *et al.* 2011). La revue de Braun *et al* (2013) permet de détailler les principales approches de construction et d'analyse des réseaux d'interactions PPI ainsi que les ressources disponibles.

Malgré l'importante quantité de données d'interactions protéine-protéine disponible, seulement une petite proportion de l'interactome est testée réellement. Chez la plante modèle *A. thaliana* par exemple, cette proportion est quantifiée à hauteur de 2 % selon l'étude de Stark *et al* (2011) et à hauteur de 10% selon l'Arabidopsis Interactome Mapping Consortium (2011). De plus, il est supposé que parmi ces données la proportion des interactions détectées mais non avérées *in vivo* (faux positifs) et celle des interactions non détectées alors qu'elles existent (faux négatifs) sont très élevées (Ito, *et al.* 2001). Huang H *et al.* (2007) estiment que le taux de faux positifs est compris entre 25 et 45% pour la levure, le nématode et la drosophile et que le taux de faux négatifs varie de 75% pour le nématode à 90% pour la drosophile. L'Arabidopsis Interactome Mapping Consortium (2011) quant à lui, estime que 20% des 6000 interactions PPI identifiées chez *A. thaliana* sont probablement des faux positifs et estime le taux de faux négatifs à 84%. Cette faible qualité des données PPI pourrait être due aux propriétés transitoires de certains complexes protéiques ou également à la couverture et la fiabilité des technologies à haut-débit employées pour l'identification des interactions (Schachter 2002; von Mering *et al.* 2002). Enfin les techniques d'identification sont pour la plupart des techniques *in vitro* donc les protéines testées sont sorties de leur contexte naturel et certaines de leurs interactions sont susceptibles de ne pas exister réellement dans la cellule.

En conclusion, les données PPI correspondent à des interactions physiques. Elles reflètent donc une preuve forte de liens fonctionnels entre les protéines partenaires. Cependant ces données souffrent de plusieurs limites qui doivent être prises en considération lors de leur utilisation pour l'inférence fonctionnelle.

b. *Réseaux d'interactions TF-cibles*

Les réseaux de régulation génétique décrivent les interactions qui peuvent être établies entre les gènes et la manière avec laquelle l'expression de certains affecte l'expression des autres. Un

exemple classique d'interaction génétique est la relation d'activation ou d'inhibition des facteurs de transcription sur leurs cibles.

L'identification des interactions facteurs de transcription-cibles (TF-cibles) peut être considérée comme un moyen utile pour l'identification de groupes de gènes corégulés car les membres de tels groupes sont susceptibles de partager une fonction biologique commune. La technique de Chromatine Immuno Précipitation (ChIP) (Collas 2010) est largement utilisée pour l'identification et la prédiction des liaisons de facteurs de transcription au sein des régions régulatrices des gènes cibles. Il est également possible d'étudier les réseaux génétiques à partir des données d'expression (puces ADN, RNA-seq) (D'Haeseleer *et al.* 2000) en supposant que les niveaux d'expression des régulateurs sont corrélés avec leur niveau d'activité de régulation. La base de données AGRIS (Arabidopsis Gene Regulatory Information) représente la source de référence des données autour des séquences promotrices, des motifs cis-régulateurs et des interactions TF-cibles chez la plante modèle *Arabidospis thaliana.*

Bien que le potentiel de ces interactions physiques à refléter le lien fonctionnel soit grand, la qualité de ces données même expérimentales a été largement discutée. En effet, des études (Farnham 2009; Moreno-Risueno *et al.* 2010) montrent que seulement 1 à 10 % des gènes cibles impliqués dans des interactions identifiées par ChIP répondent significativement à une altération du niveau d'expression du facteur de transcription correspondant. La qualité des données prédites est également discutée par l'étude d'Elnitski *et al.* (2006) qui pointe les limites des techniques de prédiction des interactions de type protéine-ADN et la complication de leur application à des organismes complexes tels que les mammifères.

c. *Réseaux de coexpression des gènes*

La coexpression des gènes est analysée à partir des données transcriptomiques. Le transcriptome correspond à l'ensemble des transcrits d'un organisme exprimés dans un contexte expérimental donné. L'étude du transcriptome a pris un tournant important grâce aux technologies à haut-débit telles que les puces ADN ou encore le RNA-seq (RNA Sequencing) qui permettent d'étudier l'ensemble des transcrits accumulés dans un tissu à un moment donné et sous des conditions données. Ces techniques ont permis l'accumulation d'une quantité de données importante pour un grand nombre d'organismes.

Les études fonctionnelles fondées sur l'analyse du transcriptome s'intéressent essentiellement à l' information concernant le profil d' expression de chaque gène défini par le vecteur de mesures d' expression associé à ce gène pour toutes les conditions testées. Le défi est de trier les profils en fonction de leurs ressemblances afin de créer des groupes de gènes ayant des profils d'expression similaires et un comportement coordonné sous un ensemble de conditions. Les membres de ces groupes sont de bons candidats pour être impliqués dans le même processus biologique (Eisen *et al.* 1998; Wolfe *et al.* 2005 ; Schoner *et al.* 2007). Les réseaux de coexpression synthétisent donc la similitude de l'expression entre les gènes. Les méthodes les plus couramment utilisées sont fondées sur le calcul d'une distance à partir des données transcriptomiques, telles que les k_means, la classification hiérarchique ou encore l'utilisation d'un seuil sur la valeur de la corrélation de Pearson ou Spearman. Elles permettent de grouper les gènes selon leur profil d'expression (Stuart *et al.* 2003). D'autres, moins utilisées, sont les méthodes probabilistes de classification non- supervisée, telles que les modèles de mélange qui permettent de considérer l'ensemble des données simultanément (Yeung *et al.* 2001). Pour gérer la quantité de données sans cesse croissante, plusieurs bases de données dédiées au stockage et la gestion de ces données transcriptomiques ont été développées. Parmi les bases de données les plus utilisées, nous pouvons citer Gene Expression Omnibus de NCBI (Barrett *et al.* 2007) ou encore la base ArrayExpress de l'EBI (Parkinson *et al.* 2007). D'autres sont plus orientées vers la mise en place d'outils d'analyse et d'affichage des données telles que Genevestigator (Zimmermann *et al.* 2005) ou aussi Stanford Microarray Database (Demeter *et al.* 2007). Quant à la base de données CATdb (Gagnot *et al.* 2008), elle est dédiée aux données transcriptomiques d'une seule plateforme transcriptome, celle de l'unité où j'ai réalisé ma thèse. Initialement dédiée à la plante modèle *A. thaliana* elle contient désormais des transcriptomes de 20 plantes (http://urgv.evry.inra.fr/CATdb).

Les données de coexpression correspondent à des données contextuelles c'est-à-dire que les interactions mises en évidence sont dépendantes d'un contexte particulier (mutation, stress, stade de développement etc). Cependant, contrairement aux données d'interactome, les données d'expression mesurent l'accumulation des transcrits *in vivo*. En plus, parmi toutes les ressources de données disponibles, la ressource de données transcriptomiques est considérée comme la plus abondante et la plus complète, ce qui offre à la génomique fonctionnelle la possibilité d'étudier le comportement coordonné des gènes à l'échelle du génome entier.

IV. Annotation fonctionnelle par intégration de données biologiques

1. Limitations dans l'exploitation d'un seul type de données

Chaque type de données présente des limitations de disponibilité, de qualité ou de biais techniques, qui restreignent son utilisation dans le cadre de l'annotation fonctionnelle. Après plusieurs études il semble que chaque type de données n'est susceptible d'éclairer qu'une seule partie de la fonction d'un gène ou un seul des trois domaines d'ontologie GO. La similarité de séquence est plus informative sur la fonction moléculaire des gènes (Tian et Skolnick 2003). L'analyse de l'interactome et la coexpression sont plus adaptées pour l'inférence des processus biologiques et de la localisation cellulaire (Vazquez *et al.* 2003 ; Persson *et al.* 2005 ; Ryngajllo *et al.* 2011).

Il est évident que l'utilisation d'un seul type de données fournit seulement une dimension de la machinerie qui contrôle le comportement des cellules. Il est donc indispensable de confronter toutes les données et les connaissances disponibles afin de palier aux limites de chaque type de données pris à part, de renforcer la fiabilité des données sur lesquelles se fondent les hypothèses émises pour la compréhension du fonctionnement de la cellule, et d'avoir enfin tous les angles d'analyse nécessaires pour compléter nos connaissances fonctionnelles des gènes étudiés.

2. Le défi de l'intégration des données biologiques

Face à la complexité du monde vivant (Blagosklonny et Pardee 2002), les quantités exponentielles de données biologiques à tous les niveaux organisationnels de la cellule et les limites et spécificités de chaque type de données, la bioinformatique est obligée de se tourner vers leur intégration pour en extraire de la connaissance biologique. La recherche en bioinformatique ne doit pas se contenter de collecter et d'organiser cette masse de données, elle doit être capable de mettre en place des stratégies et des outils d'intégration pour une représentation unifiée afin d'aider à l'interprétation et l'exploitation optimales des données. Cependant l'intégration est une tâche complexe (Chung and Wooley, 2003) qui se heurte à plusieurs difficultés notamment l'hétérogénéité des données qui se manifeste à plusieurs niveaux :

Type des données

Il s'agit de l'hétérogénéité des types de données qui peuvent refléter chacun un aspect différent de l'information biologique. Ainsi les données disponibles sont des données quantitatives (exemple :

niveau d'expression des gènes) ou des données qualitatives (exemple : localisation cellulaire d'une protéine). L'intégration des deux types de données nécessite la mise en place de stratégies qui tiennent compte de ces différences.

Objets renseignés

Les données décrivent également des objets ou des entités différents : d'un côté les entités biologiques (gènes, protéine, enzymes, métabolites, etc.) et de l'autre les relations qui lient ces différentes entités biologiques (PPI, coexpression, régulation, réaction enzymatique, etc.).

Niveaux moléculaires

Les données génomiques, transcriptomiques, protéomiques ou métaboliques ne reflètent pas le même niveau organisationnel de la cellule ni le même niveau de régulation. Pourtant ces données sont complémentaires et la compréhension des fonctions physiologiques d'un organisme passe forcément par la compréhension des mécanismes moléculaires via la superposition des différents niveaux d'organisation.

Techniques d'acquisition

Les techniques d'acquisition et leurs qualités sont également différentes. Les données disponibles sont issues de différents laboratoires et différents protocoles expérimentaux. La confiance et le poids qu'on peut leur accorder peuvent être très variables et représenter un défi supplémentaire. Les données PPI par exemple peuvent être acquises par différentes techniques tels que : Y2H, TAP-tag, PCA etc. De la même manière les données transcriptomiques peuvent être produites par plusieurs techniques à haut-débit et plusieurs protocoles d'hybridation dans différents laboratoires qui utilisent différentes méthodes d'analyses et outils statistiques pour leur normalisation et correction (Horan *et al,* 2008).

Hétérogénéité sémantique

Il s'agit de l'hétérogénéité sémantique liée aux bases de données et systèmes d'informations. Cela peut être au niveau des schémas et structures conceptuels des bases de données utilisées pour représenter, stocker et gérer les données biologiques ce qui rend l'accès aux données issues de ces systèmes d'informations encore plus difficile. L'hétérogénéité sémantique apparaît également au niveau des données elles-mêmes qui peuvent être représentées par différents vocabulaires dans chacune de ces bases de données alors qu'il s'agit de la même donnée.

L'utilisation des ontologies, définies comme une spécification explicite et formelle d'une conceptualisation commune par Thomas Gruber (Gruber 1993), aide à réduire cette hétérogénéité et

facilite ainsi l'intégration. En effet, pour pouvoir intégrer toutes ces données il est indispensable qu'elles soient facilement comparables et échangeables et donc décrites sous forme standard. Plusieurs représentations unifiées de l'information ont été proposées telles que l'annotation Gene Ontologie (GO) ou les ontologies PSO (Plant Structure Ontology) (Ilic *et al.* 2007) qui permettent d'unifier la manière de décrire l'anatomie et la morphologie des plantes à fleurs. Le consortium MGED (Microarray Gene Expression Data, www.mged.org) a permis la mise en place de trois systèmes principaux pour la représentation et la mise en ligne des données de puces à ADN : le système MIAME (le Minimum Information About A Microarray Experiment) définit le minimum d'informations requises pour la description d'un échantillon. MAGE-OM (MicroArray Gene Expression Object Model) spécifie le modèle pour représenter les informations collectées et MGED ontology définit les termes à utiliser pour désigner les différents concepts.

3. Etat de la situation

a. *Apport de l'intégration de données*

Plusieurs études ont été proposées pour l'intégration de données afin d'améliorer le pouvoir prédictif des réseaux fonctionnels. Il est considéré que des interactions moléculaires validées par différentes approches expérimentales permettent d'augmenter la précision de l'inférence fonctionnelle (Lee *et al*. 2011 ; Lee *et al.* 2010; Karaoz *et al.* 2004 ; Troyanskaya *et al.* 2003 ; Marcotte *et al.* 1999a). Un projet d'évaluation des méthodes d'inférence « critical assessment of protein function annotation (CAFA) » montre que les méthodes intégrant plusieurs types de données ont de meilleurs résultats quant à l'inférence de fonction (Radivojac *et al.* 2013).

b. *Données intégrées*

Différents types de données ont été intégrés et notamment les interactions moléculaires qui suscitent un grand intérêt de la part des chercheurs pour leur disponibilité et leur importance pour la détection de partenaires fonctionnels. En se référant à certaines analyses qui montrent que les gènes partenaires d'interactions physiques de type protéine-protéine tendent à avoir un profil d'expression similaire (Ge *et al*. 2001 ; Hahn *et al*. 2005), plusieurs études se sont donc intéressées à l'intégration de la dynamique de l'expression des gènes avec les réseaux d'interactions protéiques : l'utilisation de ces données hétérogènes peut être différente en fonction de la confiance accordée à chaque type de données, de la question biologique posée et de leur intégration simultanée ou non (Heyndrickx et

Vandepoele, 2012; Ideker *et al*. 2002 ; Cabusora *et al.* 2005 ; Segal *et al.* 2003). Certaines études proposent aussi d'étudier les interactions protéine-protéine des groupes de gènes ayant les mêmes profils d'expression sous un ensemble de conditions (Balazsi *et al.* 2005 ; de Lichtenberg *et al.* 2005 ; Wachi *et al.* 2005 ; Luscombe *et al.* 2004). D'autres proposent une approche inverse qui permet d'étudier la cohérence de l'expression de voies ou complexes connus (Zien *et al.* 2000; Jansen *et al*. 2002; Tornow et Mewes 2003; Simonis *et al*. 2004). Des données d'interactions protéine-protéine, d'interactions protéine-ADN et voies moléculaires ont été également intégrées ensemble (Hanisch *et al.* 2002). AraNet (Lee *et al.* 2010), GeneMania (Warde-Farley *et al.* 2010), STRING (Szklarczyk *et al.* 2011), CORNET (De Bodt *et al.* 2012) et kiwi (Leif Varemo *et al.* 2014) sont des exemples d'outils permettant la confrontation de données hétérogènes (profils phylogénétiques, fusion de gènes, coexpression, cocitation etc) et la visualisation des liaisons entre ces données via des interfaces graphiques.

c. *Méthodes d'intégration proposées*

Une multitude de méthodes d'intégration a été proposée, à commencer par des méthodes simples qui se limitent à collecter plusieurs types de données, à mettre à disposition un système de visualisation et des outils d'analyse notamment les tests d'enrichissement afin de mettre en évidence des caractéristiques des groupes de gènes étudiés telles que l'enrichissement en termes GO (Khatri et Draghici 2005; Wrobel *et al*. 2005). D'autres méthodes sont plus sophistiquées et se fondent sur des modèles mathématiques et statistiques. Je vais exposer dans les sous-paragraphes suivants de cette section, l'état de l'art des différentes méthodes d'intégration de données actuelles. J'ai classé les méthodes en fonction des données considérées et du type d'approche utilisée. Ainsi je les ai regroupées en deux classes principales. La première classe correspond aux méthodes d'intégration de données d'interactions moléculaires et de construction de réseaux hétérogènes. La deuxième classe correspond aux méthodes permettant d'intégrer différents types de données via l'utilisation des modèles mathématiques et des méthodes d'apprentissage automatique.

i. Méthodes d'intégration de réseaux moléculaires

Ces méthodes tirent parti de la disponibilité des réseaux d'interactions et la majorité de ces méthodes tente de combiner les différents types de réseaux d'interactions moléculaires pour former un seul et unique réseau composite. Les arêtes du réseau obtenu peuvent être pondérées en fonction du type de données et de la confiance qu'on lui accorde. Le réseau composite obtenu est alors utilisé pour la

prédiction de fonction (Tsuda *et al*. 2005 ; Mostafavi *et al.* 2008 ; Wang *et al.* 2009; Mostafavi et Morris 2010 ; Guoxian Yu *et al.* 2015) via des méthodes d'inférence discutées dans la prochaine section « V. Inférence de fonction ». Il y a trois types d'approches pour la construction du réseau composite selon Karaoz *et al.* (2004) : (i) une méthode simple mais stringente inclut uniquement les arêtes d'intersection entre les réseaux individuels (Marcotte *et al.* 1999b), l'inconvénient de ce type d'approche est le taux élevé de faux négatifs. (ii) une autre méthode est de considérer l'union des réseaux individuels. Cette approche est donc plus permissive mais tend à faire augmenter le taux de faux positifs. (iii) La troisième approche consiste à calculer la probabilité de l'existence d'une arête dans le réseau composite à partir des arêtes existant dans les réseaux individuels (Pavlovic *et al.* 2002 ; Troyanskaya *et al*. 2003).

En opposition aux approches qui combinent les réseaux individuels pour former un réseau composite, d'autres approches proposent d'utiliser des classifieurs ou des modèles pour chaque réseau source séparément et de combiner ensuite les différents classifieurs ou décisions prises à part en une seule décision finale grâce aux techniques d'apprentissage. La revue de Tian *et al.* (2011) permet de détailler les méthodes d'intégration de données omiques.

ii. Méthodes d'intégration de tout type de données

Ces méthodes sont fondées sur l'utilisation et le développement d'outils mathématiques et de méthodes d'apprentissage automatique permettant d'intégrer différents types de données et pas seulement des données d'interactions puisque les données disponibles peuvent être également des caractéristiques décrivant les entités biologiques (gènes, protéines, enzymes etc.). L'ensemble des données est représenté sous la forme d'un vecteur d'attributs donné en entrée aux algorithmes d'apprentissage (Boser *et al*. 1992; Burges. 1998; Cristianini and Shawe-Taylor 2000). Ces méthodes peuvent être classées en deux types d'approches.

Le premier type d'approche correspond aux méthodes qui intègrent d'abord tous les attributs par source d'information disponible, puis effectuent l'analyse fonctionnelle (Glenisson, Mathys *et al*. 2003; Lanckriet *et al*. 2004). Chaque type de données est représenté grâce à une fonction à noyau. Toutes les matrices sont ensuite intégrées pour former une matrice noyau composite utilisée pour l'inférence.

Le deuxième type d'approche permet de modéliser chaque source de données séparément, d'apprendre de chaque source indépendamment et de combiner ensuite ces modèles. Sass *et al.* (2013) ont développé un algorithme appelé « MONA : Multi-level ONtology Analysis » fondé sur

les modèles bayésiens et qui permet de modéliser chaque source indépendamment. Yang *et al.* (2005) ont développé un algorithme appelé « MSC : Multi-Source Clustering » qui permet la classification de plusieurs types de données tout en construisant de manière stochastique des modèles pour chaque source de données.

Pavlidis *et al*. (2002) étudient trois manières d'intégration de données d'expression avec des profils phylogénétiques dans un SVM pour prédire la fonction de gènes. Ces trois manières correspondent à trois moments différents pour l'intégration des données : (i) Intégration simultanée en combinant tous les vecteurs d'attributs des différentes sources pour chaque gène en un seul vecteur. (ii) Intégration parallèle où une fonction noyau est calculée séparément pour chaque source de données, les noyaux résultants sont ensuite additionnés en une seule matrice. (iii) Intégration a posteriori où les résultats finaux de l'apprentissage de chaque source de données sont combinés. Plusieurs résultats intéressants sont issus de leur étude. D'abord ils confirment que le SVM apprend mieux de la combinaison des deux types de données que d'un seul type de données. Ils ont révélé également que les différentes sources de données ont des qualités différentes, et que l'établissement de pondérations pour les différents réseaux peut améliorer la précision de l'inférence fonctionnelle. Finalement leur étude comparative a révélé que les intégrations simultanée et parallèle montrent une meilleure performance que l'intégration a posteriori. Cependant cette dernière conclusion est remise en cause par Fujishima *et al.* (2007) dont l'étude de comparaison montre que l'intégration a posteriori a la meilleure performance en termes de sensibilité, alors que l'intégration parallèle a la meilleure performance en termes de spécificité et de précision. Ils expliquent cette contradiction de résultat par le fait que les performances de prédiction peuvent être impactées par de nombreux facteurs tels que les critères et les indices d'évaluation, les types de données, les paramètres des fonctions de noyaux ou encore les algorithmes de sélection d'attributs. Ainsi ces études comparatives mettent le doigt sur la complexité de l'utilisation de ces méthodes d'intégration et la difficulté de l'évaluation de leurs performances prédictives.

d. *Limites de la littérature*

Horan *et al.* (2008) soulignent le fait que certaines études souffrent de l'hétérogénéité de l'origine des données utilisées qui peuvent être issues de différents laboratoires et différents protocoles expérimentaux. Bassel *et al.* (2012) discutent le manque de spécificité de contexte de certaines de ces études qui profitent de la disponibilité des données d'expression à haut débit pour intégrer des transcriptomes de différents types de tissus, différents types de cellules et différentes conditions, ce

qui ne permet de capturer que les processus communs dans les échantillons utilisés tels que la photosynthèse. Bassel *et al.* (2011b) suggèrent que les analyses dans un contexte spécifique c'est-à-dire condition-dépendant permettent d'augmenter la spécificité des processus et interactions fonctionnelles identifiés. En conclusion, La qualité des prédictions fonctionnelles est étroitement liée à la sélection des données brutes utilisées pour l'analyse.

V. Inférence de fonction

L'intégration des données biologiques a pour objectif de collecter toutes les informations susceptibles d'éclairer la fonction des gènes inconnus analysés, notamment en identifiant leurs partenaires potentiels. Cette tâche correspond à la première étape de l'annotation fonctionnelle. L'étape suivante correspond au développement de méthodes permettant l'exploitation de ces informations afin de proposer une fonction aux gènes inconnus ou mal caractérisés. L'inférence ou la prédiction de fonction constitue un défi de taille. Les protéines sont souvent multifonctionnelles (Jeffery, 1999) et promiscuitaires (Kersonsky et Tawfik, 2010). Ainsi l'hypothèse « un gène-une enzyme » (Beadle et Tatum, 1941) semble maintenant être une simplification excessive. L'étude de Clark et Radivojac (2011) attire l'attention sur le fait que 30% des protéines de la base de données Swiss-Prot (Boeckmann *et al.* 2003) sont associées à plus d'un terme de l'ontologie « Fonction Moléculaire » et 60% pour l'ontologie « Processus Biologique ». Pour toutes ces raisons, l'utilisation des réseaux d'interactions moléculaires issus des données omiques est utile car ils ont la capacité de révéler le caractère modulaire de l'activité cellulaire et de délimiter les processus biologiques, les voies et complexes auxquels les protéines sont impliquées. De nombreux travaux étudient donc la fonction des gènes dans le contexte des réseaux d'interactions en supposant que les protéines partenaires sont susceptibles de partager les mêmes fonctions. Les paragraphes suivants de cette section sont dédiés à la description et l'état de l'art des différentes méthodes actuelles de l'inférence fonctionnelle. Nous pouvons classer ces méthodes en deux classes principales. D'une part les méthodes fondées sur l'analyse de la structure et de la topologie des réseaux. D'autre part les méthodes qui tirent profit des méthodes d'apprentissage automatique.

1. Inférence par analyse topologique des réseaux d'interactions moléculaires

Les interactions moléculaires directes ou indirectes qui relient les entités biologiques font partie du processus d'annotation fonctionnelle des génomes puisque d'une part ils peuvent renseigner sur le lien fonctionnel entre les gènes et d'autre part sur l'organisation fonctionnelle globale des cellules et des organismes. Ainsi les réseaux construits à partir de ces interactions permettent de refléter l'architecture modulaire de l'organisation des processus biologiques au sein de la cellule.

Les données d'interactions sont représentées et visualisées naturellement comme un graphe où les nœuds sont les objets génomiques et les arêtes sont les interactions. Cependant, la quantité de données d'interactions étant très importante, la simple représentation visuelle n'est pas suffisante pour l'exploration des données. Les études d'analyses fonctionnelles ont donc utilisé la théorie des graphes pour analyser les propriétés des réseaux obtenus. Ainsi l'étude de la topologie de ces réseaux permet de caractériser leur structure, leur stabilité, les réponses dynamiques et la fonction des gènes (Boccalettia *et al.* 2006; Zhu et Qin 2005 ; Bergmann *et al.* 2004 ; Wuchty *et al.* 2006). La topologie d'un réseau désigne son architecture physique. Les propriétés topologiques étudiées sont généralement la distribution des degrés des nœuds, des coefficients de clusterisation et la densité de distribution de distances entre les nœuds (Wagner, 2001; Watts et Strogatz, 1998).

Le degré des nœuds (Bader et Hogue, 2003) est la propriété la plus étudiée. Elle correspond à la connectivité d'un nœud autrement dit le nombre de voisins auxquels il est connecté. L'étude des réseaux réels a permis de révéler que la distribution du degré des nœuds de la majorité de ces réseaux est gouvernée par une loi de puissance. Dans ces réseaux la majorité des nœuds ont un faible degré c'est-à-dire très peu connectés alors qu'une très faible minorité de nœuds sont fortement connectés. Cette propriété dite structure « scale free » ou sans échelle a été observée dans des réseaux non biologiques (Barabasi et albert, 1999) tels que le graphe des appels téléphoniques ou aussi le World Wide Web et dans les réseaux biologiques tels que les réseaux métaboliques (Jeong *et al*. 2000) ou les réseaux d'interactions protéine-protéine (Maslov et Sneppen, 2002). Cette propriété confère la robustesse et la stabilité aux réseaux car si on supprime un nœud pris au hasard il y a de forte chance que ce soit sans influence sur sa structure et son fonctionnement puisqu'il y a de forte chance qu'il s'agisse d'un nœud peu connecté (Albert *et al.* 2000). Les gènes fortement connectés, appelés des « hubs », ont été largement étudiés (Rhee et Mutwill, 2014 ; Ning *et al.* 2010 ; Wu et Qi 2010 ; Han *et al.* 2004 ; Fox *et al.* 2011). Il semble que ces gènes soient essentiels (Jeong *et al.* 2001), qu'ils aient tendance à être conservés entre les espèces et à évoluer lentement (Fraser *et al.*

2002). L'inactivation de ces gènes conduit à l'apparition de nombreux phénotypes distincts liés probablement à l'activité pléiotropique de ces gènes (Yu *et al.* 2008).

D'autres propriétés sont étudiées notamment le coefficient de clustering défini comme la probabilité que deux voisins d'un nœud soient connectés entre eux. Typiquement, les réseaux réels ont un coefficient de clustering élevé traduisant l'existence de zones avec une densité locale élevée par rapport à la densité globale. Ces zones correspondent à des modules de gènes fortement connectés entre eux et suggèrent que les membres de tels modules participent aux mêmes processus.

Une autre mesure calculée pour la caractérisation des réseaux est le « betweenness centrality » qui correspond pour un nœud au nombre de chemins entre deux nœuds qui passent par celui-ci. Elle peut être considérée comme une mesure de la résistance physique du réseau. Cette mesure peut être utilisée également pour la détection de modules ou communautés (la notion de communauté sera expliquée dans le paragraphe suivant). Par ailleurs la distance moyenne entre les nœuds dans les réseaux biologiques est faible. Cette caractéristique est appelée effet « petit monde ».

Ces critères topologiques reflètent l'organisation globale des processus biologiques au sein de la cellule mais beaucoup d'études ont utilisé la topologie des réseaux pour la prédiction de la fonction des gènes inconnus au sein de ces réseaux sur la base de l'hypothèse que les partenaires interagissant ensemble partagent probablement la même fonction.

Sharan *et al*. (2007) résument les méthodes d'inférence exploitant la topologie des réseaux d'interactions essentiellement en deux groupes : les méthodes d'annotation directes qui prédisent la fonction d'une protéine sur la base de ses connexions dans le réseau (Chua *et al*. 2006 ; Nabieva *et al*. 2005 ; Deng *et al.* 2004) et les méthodes dites « module-assistées » qui identifient en premier des motifs ou des modules de protéines apparentées puis utilisent ces modules pour l'annotation des protéines inconnues.

a. *Méthodes d'inférence directes*

Il a été observé que les gènes reliés par des interactions dans les réseaux fonctionnels tels que les réseaux d'interactions protéine-protéine, partagent souvent la même fonction. Ainsi plusieurs méthodes proposent la propagation des informations fonctionnelles des gènes bien caractérisés dans ces réseaux vers les gènes inconnus sur la base de leurs interactions. Dans un réseau d'interactions protéine-protéine, Schwikowski *et al.* (2000) ont utilisé une approche qui assigne la fonction majoritaire des voisins directs aux gènes inconnus. Ils classent les fonctions de tous les voisins d'un gène inconnu. Ils attribuent ensuite les trois fonctions les plus fréquentes au gène étudié. Ils montrent

que cette approche est capable de prédire correctement la fonction de 72% des gènes de leur jeu de données. Titz *et al*. (2004) montrent également que les paires de protéines interagissant ensemble ont 70 à 80% de chances de partager au moins une fonction. Cependant, de nombreux gènes peuvent être reliés uniquement à des gènes non annotés dans le réseau ou peuvent être reliés à plusieurs partenaires tous de fonctions différentes. Afin de palier à ces limites, des études (Hishigaki *et al.* 2001 ; Chua *et al*. 2006) ont proposé de considérer les voisins indirects correspondant à des nœuds situés à une distance 2 ou 3. Chua *et al.* (2006) ont montré que ce type de voisinage montre une similarité de fonction sensiblement supérieure à celle attendue par hasard. D'autres approches toujours fondées sur l'hypothèse que les partenaires d'interactions sont probablement impliqués dans les mêmes fonctions, permettent de tenir compte de la totalité de la topologie du réseau. Les algorithmes de théorie des graphes tels que les approches exploitant la notion de flux dans les réseaux (en anglais « flow-based algorithm ») ou les approches exploitant la notion de coupe dans les réseaux (en anglais « cut-based algorithm ») ont été utilisés dans ce contexte. Nabieva *et al.* (2005) simulent le flux fonctionnel dans le réseau entre les gènes, où chaque gène de fonction connue est une source de flux pour les autres gènes. Vazquez *et al* (2003) et Karaoz *et al* (2004) ont cherché à optimiser le poids des arêtes cohérentes c'est-à-dire où les deux partenaires d'une arête partagent la même fonction et à minimiser le poids des arêtes incompatibles. D'autres études (Letovsky et Kasif, 2003; Deng *et al.* 2003) suggèrent des approches probabilistes exploitant le champ de Markov. Une étude plus récente (Clark et Kalita, 2014) propose d'aligner les réseaux d'interactions protéine-protéine entre plusieurs espèces afin d'identifier des parties conservées entre ces réseaux et de transférer ainsi les fonctions des gènes connus dans une espèce aux gènes inconnus de l'autre espèce.

b. *Analyse locale fondée sur la détection de motifs et modules*

L'observation de la topologie des réseaux biologiques a conduit à l'identification de structures intéressantes notamment les motifs et les modules. Ces structures locales sont à l'origine de plusieurs hypothèses qui ont été émises pour expliquer la structure et le fonctionnement des réseaux étudiés et les processus biologiques sous-jacents. Un motif dans un réseau est défini comme un patron d'interactions ou un sous-graphe significativement sur-représenté par rapport à un réseau aléatoire de même taille. Ces structures ont été observées lors de l'analyse du réseau de régulation transcriptionnelle chez *E. coli* (Milo *et al.* 2002, Shen-Orr *et al.* 2002) et ont été utilisées par Berg et Lassig (2004) pour émettre des hypothèses quant à l'histoire évolutive des réseaux à travers la

comparaison des motifs dans les réseaux de régulation de plusieurs espèces. Il s'est avéré que ces structures sont conservées du point de vue évolutif dans les réseaux d'interaction (Wuchty *et al.* 2003). La conservation de plusieurs motifs de deux, trois ou quatre nœuds dans les réseaux de régulation avec une description cohérente telle que les structures appelées « feed forward loop », montre que ces motifs ont un rôle dynamique bien spécifique (Alon, 2007 ; Shoval and Alon, 2010).

Une autre structure a également beaucoup suscité l'intérêt de la communauté scientifique biologique, ce sont les modules appelés aussi communautés c'est-à-dire des régions densément connectées entre elles et faiblement connectées aux autres. A l'origine, le terme « communauté » a été introduit pour les analyses des réseaux sociaux avec son interprétation naturelle. Dans les réseaux biologiques on utilise plutôt le terme « module » qui peut correspondre à des fonctions biologiques, complexes protéiques ou un groupe de gènes corégulés. De nombreux travaux ont révélé l'organisation en modules des réseaux étudiés tels que les réseaux métaboliques (Ravasz *et al.* 2002), ou aussi les réseaux d'interactions protéine-protéine (Sharan *et al.* 2005) et les réseaux de coexpression (Atias *et al.* 2009). La détection de modules dans les réseaux repose sur les méthodes de classification y compris des méthodes de clustering hiérarchique (Arnau *et al.* 2005 ; Rives et Galitski, 2003; Maciag *et al.* 2006 ; Brun *et al.* 2003; Samanta et Liang, 2003) et des méthodes appelées méthodes de clustering de graphe (Spirin et Mirny, 2003; King *et al.* 2004 ; Pereira-Leal *et al.* 2004 ; Przulj *et al.* 2004 ; Dunn *et al.* 2005 ; Adamcsek *et al.* 2006). Bader et Hogue (2003) ont développé un algorithme appelé MCODE (Molecular Complex Detection) pour la détection des complexes moléculaires en exploitant la connectivité du réseau avec le coefficient de clustering. Cet algorithme a été largement utilisé dans plusieurs études d'annotation fonctionnelle (Atias *et al.* 2009; Xia *et al.* 2006). L'algorithme MCL (Markov Cluster algorithm) a été adapté pour le clustering dans un graphe en considérant les arêtes comme étant une mesure de similarité entre les protéines (Enright *et al.* 2002). Nepusz *et al.* (2012) ont développé un outil appelé ClusterONE qui permet le chevauchement des gènes dans la détection des modules, ce qui est plus intéressant du point de vue biologique puisque les gènes peuvent assurer plusieurs fonctions à la fois dans une cellule (pour une revue sur les méthodes de classification exploitant les réseaux d'interaction voir Schaeffer, 2007).

L'hypothèse sous-jacente sur laquelle se fonde ce type de travaux est que les membres de tels groupes sont liés physiquement ou fonctionnellement afin d'accomplir une fonction (Hartwell *et al.* 1999, Barabasi et Oltvai 2004). Cela leur permet d'émettre des hypothèses concernant la fonction potentielle des gènes orphelins au sein de tels groupes sur la base de l'information des fonctions de

ses voisins. Plusieurs études assignent la fonction majoritaire du module aux gènes non-annotés du même groupe (N.Armstrong et M. van de Wiel, 2004).

2. Inférence par les algorithmes d'apprentissage automatique

En plus des méthodes d'inférence fondées sur l'analyse topologique des réseaux d'interactions, une deuxième classe de méthodes existe. Cette classe correspond aux algorithmes d'apprentissage automatique qui interviennent dans des processus de décision pour la découverte de la fonction des gènes tout en permettant de prendre en compte des données d'interactions mais aussi d'autres types de données. Toutes les données caractérisant les gènes ou protéines considérées sont représentées comme un ensemble d'attributs pris en entrée par ces algorithmes.

Les méthodes d'apprentissage automatique sont utilisées pour l'intégration de données (comme expliqué dans le paragraphe « Méthodes d'intégration de tous types de données » de la section « IV. Annotation fonctionnelle par intégration de données biologiques »), mais aussi pour l'inférence de fonction. En effet comme montré précédemment, certaines de ces méthodes permettent d'effectuer l'intégration des données et l'attribution de fonction aux gènes dans une même analyse.

Lors de ce paragraphe nous aborderons la définition et le principe général des méthodes d'apprentissage, une vue globale du large éventail des méthodes et approches proposées dans le cadre de la prédiction de fonctions des gènes, l'apport de ces approches et les critères de leur évaluation.

a. *Définition et Principe*

L'apprentissage automatique connu sous le terme anglais « Machine Learning » correspond à un des champs d'étude de l'intelligence artificielle. Cette discipline scientifique fait référence, selon la définition de Wikipedia, au développement, à l'analyse et l'implémentation de méthodes automatisables qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques.

L'apprentissage automatique est considéré comme une vraie révolution dans le domaine informatique puisqu'il permet de franchir une nouvelle étape qui correspond à « apprendre » des données et pas seulement « calculer ». En effet, il s'agit d'automatiser la prise de décision, habituellement effectuée par des experts et donc d'en exclure toute intervention humaine à travers la

mise en place d'un certain nombre de règles apprises directement des données sur la base d'exemples déjà traités. L'apprentissage automatique est très utile dans le cadre d'application sur des données volumineuses.

Je me suis intéressée naturellement dans le cadre de ce travail, à l'application de l'apprentissage supervisé au champ de la prédiction de la fonction des gènes inconnus. Les gènes dont la fonction est connue sont le support de base du système pour la mise en place d'un modèle qui décrit un ensemble de règles pour la prédiction de fonction aux gènes non annotés. Le processus se passe en deux phases : la phase d'apprentissage pendant laquelle le système apprend des données pour construire un modèle. Ensuite, la phase de test où le système prédit la fonction d'un nouveau gène à partir du modèle construit. De nombreux algorithmes d'apprentissage supervisé ont été proposés pour la construction du modèle, y compris les arbres de décision, les SVM, la classification Bayésienne, la classification des k plus proches voisins, les réseaux de neurones ou encore les forêts d'arbres décisionnels (Random Forest).

Dans ce contexte, les méthodes d'apprentissage automatique peuvent apporter des éléments de réponse autour de la fonction des gènes de deux manières possibles. Des réponses binaires c'est-à-dire si oui ou non une étiquette ou une fonction est assignée au gène étudié ou des réponses qui peuvent être aussi sous une forme discrète à plus de deux valeurs. Nous pouvons classer ainsi les approches selon ces deux types de réponses.

i. Méthodes de classification binaires

Ce type d'approche est appelé classification binaire car elle permet d'évaluer, pour chaque fonction considérée, si le gène étudié peut être annoté par cette fonction. La fonction des protéines est ainsi étudiée comme une série de classificateurs binaires en considérant que chaque classe ou fonction est indépendante des autres classes.

Plusieurs études sont fondées sur cette approche. Ryngajllo *et al.* (2011) ont développé un outil appelé SLocX afin de prédire les localisations subcellulaires des protéines chez *A. thaliana* via un classifieur SVM binaire appliqué à chaque compartiment cellulaire. La classification binaire a permis également à Lan *et al*. (2007) d'identifier les gènes qui répondent au stress chez *A. thaliana* à partir de données d'expression et via l'utilisation de plusieurs méthodes d'apprentissage. Ils proposent ensuite une méthode de combinaison des décisions des différentes méthodes d'apprentissage en une décision finale pour améliorer le pouvoir prédictif. Peng *et al*. 2014 ont développé un algorithme qui calcule la similarité des domaines Pfam (Protein family) entre les

protéines partenaires dans un réseau PPI. Pour chaque fonction ils vérifient si elle est assignée au partenaire identifié comme étant le plus proche de la protéine étudiée, alors cette fonction lui sera attribuée. La prédiction d'interactions pour l'inférence de réseau peut aussi être considérée comme un problème de décision binaire. L'algorithme d'inférence joue dans ce contexte, le rôle d'un classifieur pour chaque paire de gènes afin de décider si une étiquette positive (arête) ou une étiquette négative (pas d'arête) sera attribuée au couple (De Bodt *et al*. 2009 ; Meyer *et al*. 2008).

ii. Méthodes de classification à plusieurs classes

Une protéine est généralement associée à plusieurs fonctions. De nombreuses approches traitent ainsi le problème de la prédiction en utilisant une classification à plusieurs classes ou multi-étiquettes. Barutcuoglu *et al.* (2006) proposent de combiner les prédictions produites par des classifieurs binaires pour chaque fonction, grâce à un modèle Bayésien qui tient compte de la structure hiérarchique des termes GO. En effet, il est connu que les ontologies d'annotations telles que FunCat (Ruepp *et al.* 2004) ou GO sont organisées de manière hiérarchique où les termes généraux incluent les termes les plus spécifiques. Plusieurs autres études ont introduit une structure hiérarchique afin de tenir compte de cette spécificité (Cheng *et al.* 2014 ; Schietgat *et al.* 2010; Valentini, 2011). Pandey *et al.* (2009) ont introduit la notion de corrélation de fonctions dans les approches multi-étiquettes en exploitant la similarité de Lin (1998) et la méthode du plus proche voisin KNN créant ainsi un classifieur appelé LKNN. Wang *et al.* (2013) ont également étudié la corrélation entre les fonctions et proposent un algorithme appelé FCML fondé sur les graphes et les fonctions à noyaux.

La revue de Zhang *et al.* (2013) permet de détailler les méthodes d'apprentissage pour la classification multi-classes.

3. Performance des méthodes d'inférence actuelles et défis

Comme décrit précédemment, il existe une panoplie de méthodes et d'approches d'inférence de fonctions possibles exploitant différents types de données. Il est donc nécessaire de pouvoir évaluer la performance de ces méthodes. Dans cet objectif, un projet mondial appelé CAFA (Critical Assessment of protein Function Annotation), visant à analyser et évaluer les méthodes de prédiction de la fonction de la protéine à large échelle a été réalisé (Radivojac *et al*. 2013). Ce projet a permis l'évaluation de 54 algorithmes et méthodes d'annotation représentant l'état de l'art mais dont la majorité est fondée sur l'analyse de similarité de séquence. Cette analyse a permis de montrer une

différence significative de la capacité des méthodes testées à prédire les termes des deux ontologies GO : MF et BP. Les auteurs ont considéré que la performance globale de ces méthodes pour la prédiction des termes de l'ontologie MF est acceptable, par contre les résultats sont en dessous de leurs attentes quant à la prédiction des termes de l'ontologie BP. Selon cette étude les deux méthodes les plus performantes pour prédire les ontologies MF et BP sont Jones-UCL (Cozzetto *et al*. 2013) et Argot2 (Falda, *et al*. 2012). Gillis et Pavlidis (2013) présentent une étude indépendante des résultats sortis par le projet CAFA. Ils montrent qu'en utilisant d'autres critères d'évaluation ils arrivent à des conclusions différentes. En effet, contrairement à Radivojac *et al.* (2013) ils concluent que les méthodes simples reposant sur la similarité de séquence sont hautement concurrentielles. Cette analyse met ainsi le doigt sur l'impact des critères d'évaluation sur le calcul du pouvoir prédictif des méthodes et leur comparaison. Cette problématique a été discutée également dans d'autres études (Elkan et Noto, 2008 ; Rider *et al*. 2013) qui soulèvent aussi l'impact de la qualité des données disponibles et la fiabilité des étiquettes positives et négatives qui sont utilisées lors de l'apprentissage. En effet pour entrainer un algorithme discriminatoire, le jeu de données d'apprentissage doit correspondre à un ensemble d'objets (en occurrence des gènes) positifs (associés à une fonction étudiée) et un ensemble d'objets négatifs. Mais en réalité pour une fonction étudiée, nous disposons d'un ensemble de gènes possédant la fonction dont la qualité est parfois discutable et d'un ensemble de gènes non étiquetés mais sans savoir si cette fonction n'est pas assignée à ces gènes négatifs parce qu'ils n'assurent pas cette fonction réellement ou parce qu'ils n'ont pas été testés. L'étude de Rider *et al.* (2013) montre que le comportement des paramètres d'évaluation est instable en présence d'incertitude dans les étiquettes de classe et que la stabilité des paramètres d'évaluation dépend du type de biais dans les données. Jiang *et al*. (2014) discutent également de cette problématique. Leur étude porte sur l'effet des annotations expérimentales incomplètes (à cause des biais tels que ceux des méthodes expérimentales ou l'évolution des termes GO) sur la fiabilité de l'évaluation de la performance des méthodes de prédiction de la fonction des protéines. Ils supposent également un impact non négligeable mais montrent que le système d'évaluation actuel de méthodes à large échelle reste significatif et fiable.

Selon l'état de l'art actuel de ces méthodes nous pouvons relever également deux autres points importants. Le premier point concerne les algorithmes et les méthodes d'inférence sophistiqués notamment ceux qui sont fondés sur l'apprentissage. Malgré le potentiel de ces méthodes et les multiples efforts effectués pour les améliorer en tenant compte de certaines spécificités liées à l'annotation de données telles que la structure hiérarchique des termes d'annotation et leurs

corrélations, ces méthodes ont du mal à dépasser la performance de méthodes simples telles que le vote majoritaire de Schwikowski *et al.* (2000). L'étude de Wang *et al*. (2013) fondée sur l'utilisation des fonctions à noyaux tenant compte de la corrélation entre fonctions et assurant une classification multi-étiquettes, compare la performance de leur algorithme à d'autres algorithmes y compris le vote majoritaire. Selon cette comparaison, bien que leur méthode soit plus performante que les autres, cette amélioration reste modeste. De plus la mise en place et l'utilisation de certaines de ces méthodes sophistiquées pour l'annotation fonctionnelle est difficile et leurs paramètres demandent une certaine maîtrise.

Le deuxième point concerne la performance de la prédiction mais cette fois-ci non pas en fonction des méthodes ou des données utilisées mais en fonction des termes ou classes évaluées. Nous avons remarqué qu'il y a régulièrement des différences de performance en fonction des termes analysés dans la littérature. Par exemple dans l'étude de Radivojac *et al.* (2013), les auteurs montrent qu'il y a une différence de score AUC dans la prédiction des termes liés aux activités catalytique et de transport et les termes reliés aux activités de liaison qui ont un score AUC plus faible. Ils montrent également que les termes spécifiques de l'ontologie BP associés à « adhésion cellulaire », « processus métabolique », « transcription et régulation de l'expression des gènes » sont des termes associés à un haut score AUC, alors que les termes moins spécifiques tels que « locomotion », « processus cellulaire » et « réponse au stress » sont associés à une faible valeur AUC. Cette tendance se confirme également dans l'étude de Ryngajllo *et al.* (2011) où la prédiction de la localisation subcellulaire des protéines est clairement plus performante pour le chloroplaste par rapport aux autres termes. Il semble donc judicieux de rester sur une prédiction et une évaluation par terme.

VI. Contexte et objectifs de la thèse

1. Contexte

Mon projet de thèse s'inscrit dans le cadre du projet de l'équipe « Réseaux Génomiques » au sein de l'Institut des Sciences des Plantes de Paris-Saclay (IPS2), qui vise à développer des méthodes statistiques et bioinformatiques pour améliorer l'annotation fonctionnelle et relationnelle des gènes d'*A. thaliana* et de transférer ces connaissances aux plantes cultivées. Ces méthodes sont appliquées

dans un cadre spécifique qui est la réponse des plantes aux stress. Les mécanismes de défense et d'adaptation à l'environnement impliquent un large panel de changements physiologiques au niveau de la plante et leur compréhension constitue un défi important pour la biologie végétale notamment pour l'amélioration variétale et la résistance durable.

Le projet de l'équipe s'appuie principalement sur les ressources transcriptomiques qui constituent une source très importante de données omiques pour *A. thaliana* et permettent ainsi une analyse globale à l'échelle du génome entier. Notre projet permet de surmonter les limites des études précédentes : le manque de spécificité des données qui est compensé par le choix d'un contexte spécifique qui est la réponse aux stress, leur hétérogénéité en terme de techniques d'acquisition et d'hétérogénéité sémantique qui est compensé par le choix d'un jeu de données contrôlé et homogène issu du même laboratoire et dont la description des expériences respecte les instructions MIAME (Minimum Information About A Microarray Experiment). Il s'agit de la collection de données transcriptomiques produite sur la plateforme de l'unité et stockée dans CATdb (http://urgv.evry.inra.fr/CATdb ; Gagnot *et al*. 2008). Cette collection est obtenue avec les puces CATMA (Complete Arabidopsis transcriptome MicroArray) (http://www.catma.org ; Crowe *et al*. 2003) et a le rare avantage d'être générée en utilisant le même protocole d'hybridation et les mêmes pipelines d'analyses du prétraitement aux analyses différentielles. Cette ressource contient un grand nombre d'expériences concernant des stress biotiques et abiotiques sur lesquels le projet se focalise. De plus, les puces CATMA se distinguent par 5095 gènes qui ne sont pas représentées par les puces Affymetrix « GeneChip ATH1 » couramment utilisées, dont 465 gènes ont été caractérisés à l'aide du logiciel de prédiction EuGene (Aubourg *et al*. 2007) et 609 gènes prédits comme codants pour des petits ARN.

A partir de ces données la liste exhaustive des sondes différentiellement exprimées dans au moins une condition de stress a été établie. La correspondance sonde-gène est effectuée par la suite grâce à la base des données génomiques FLAGdb^{++} (Dèrozier *et al*. 2011) qui contient les différentes couches d'annotation des gènes. Une liste de 17 264 gènes décrits par 387 comparaisons transcriptomiques organisés en 18 catégories de stress (9 biotiques et 9 abiotiques) a été ainsi identifiée. La comparaison de la liste des gènes différentiellement exprimés dans les catégories de stress biotiques à celle des gènes impactés par les stress abiotiques a révélé un large chevauchement des deux jeux de données (figure 2).

L'analyse de la coexpression a été effectuée par la suite par type de stress sur les logs ratios des sondes différentiellement exprimés au moins une fois dans une comparaison à l'aide d'un modèle de

mélange de gaussiennes multidimensionnelles. Le nombre de classes du mélange est déterminé par le critère Bayesian Information Criterion (Zaag *et al*. 2015). Cette étude a conduit à l'identification de 681 clusters de coexpression. Le nombre de clusters obtenus par type de stress, y compris leurs tailles, est résumé dans le tableau 1.

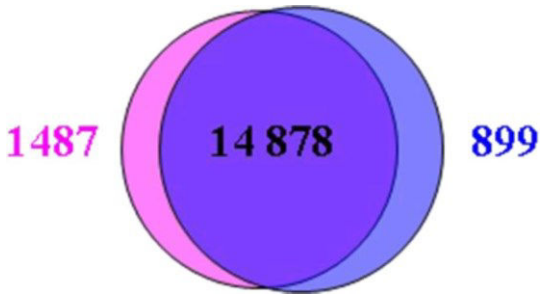


Figure 2 : Comparaison des gènes différentiellement exprimés dans les conditions de stress biotiques (rose) et dans les stress abiotiques (bleu).

Tableau 1 : Taille des clusters obtenus par type de stress.
La 1ère colonne correspond au nom du stress, la 2eme au nombre de conditions testées, la 3ième colonne correspond au nombre de clusters obtenus par les modèles de mélange. Les colonnes 4 et 5 correspondent au nombre total de gènes classes selon deux règles de classification : MAP et MFDR. Avec la règle MAP tous les gènes différentiellement exprimés sont attribués dans un cluster, avec la règle MFDR seuls les gènes ayant une probabilité a posteriori supérieure à un certain seuil sont classés.

	Stress	#conditions	#clusters	#gènes MAP	#gènes MFDR
Biotiques	Bactéries biotrophes	40	56	11 817	6 772
	Bactéries nécrotrophes	26	50	11 030	5 353
	Stifénia	6	17	1 565	821
	Virus	33	54	11 685	6 053
	Fungi	21	51	9 705	3 972
	Rhodococcus	7	13	1 965	1 157
	Nématodes	10	29	7 487	1 526
	Oomycètes	14	31	5 591	2 421
	Autres biotiques	6	20	3 803	1 098
Abiotiques	Sécheresse	17	34	8 167	3 300
	Stress oxydatif	16	52	10 027	3 681

	UV	7	37	7 903	1 004
	Température	45	34	11 199	7 247
	Sel	15	30	5 786	2 106
	Métaux lourds	45	57	10 533	7 138
	Nitrogène	46	60	13 807	8 454
	Gamma	25	32	5 419	3 547
	Autres abiotiques	8	24	3 944	1 941

2. Objectifs

Le point de départ de mon travail correspond à ces clusters de gènes regroupés en fonction de leur profil transcriptomique et qui sont reliés à la réponse aux stress de manière directe ou indirecte. En se fondant sur l'hypothèse que des gènes liés au même processus biologique ont plus de chance d'être coexprimés, mon objectif de thèse est le développement de méthodologies pour l'exploitation biologique de ces clusters de coexpression afin d'améliorer l'annotation des gènes d'*A. thaliana* impliqués dans la réponse aux stress.

Mon projet de thèse s'articule autour de deux axes principaux: premièrement l'intégration de données hétérogènes afin de renforcer la fiabilité de l'analyse de coexpression et d'identifier des groupes de gènes corégulés et deuxièmement l'inférence de fonction aux gènes inconnus au sein de ces groupes sur la base des liens fonctionnels qui les relient. Afin de montrer la démarche scientifique suivie au cours de cette thèse, le manuscrit sera structuré de la manière suivante :

Le chapitre 1 exposera le problème de l'intégration de plusieurs sources de données hétérogènes pour l'annotation fonctionnelle et relationnelle des clusters de coexpression obtenus. Les résultats de cette partie ont été mis à disposition au profit de la communauté scientifique via la conception et le développement d'un nouveau module nommé GEM2Net dans la base de données CATdb.

Le chapitre 2 abordera l'intégration d'un seul type de données à un même niveau moléculaire. Cette étape vise à renforcer la qualité des données de coexpression pour la détection de relations de corégulation entre les gènes et l'évaluation statistique de ces interactions.

Le chapitre 3 sera consacré à la description de la méthodologie développée pour l'exploitation du réseau de corégulation pour la prédiction fonctionnelle par apprentissage supervisée et l'application de cette méthode pour l'inférence de fonction aux gènes inconnus ou partiellement connus au sein de ce réseau.

Enfin je conclurai ce manuscrit en discutant l'apport de cette thèse et ses perspectives à court et long termes.

Chapitre 1 : Intégration d'informations hétérogènes pour la caractérisation fonctionnelle de groupes de gènes

I. Contexte et objectifs

Les données omiques générées à haut-débit offrent la possibilité d'avoir une vision globale des interactions physiques et fonctionnelles qui existent entre les gènes, à condition de mettre en place les outils et les méthodes adéquats pour extraire de l'information de ces données. Comme vu dans l'introduction, la classification non supervisée est considérée comme un moyen efficace pour identifier des groupes de gènes ayant les mêmes profils d'expression. Lors de ce chapitre, je considère la tâche non triviale d'évaluation et d'interprétation biologique des résultats de l'analyse de coexpression qui a été effectuée à l'aide de modèles de mélange et qui a conduit à l'identification de 681 clusters de coexpression impactés par différentes conditions de stress organisées dans 18 catégories (9 stress biotiques et 9 abiotiques).

Ces clusters de coexpression étaient le point de départ de ma thèse et mon premier objectif était leur caractérisation fonctionnelle afin d'étudier leur validité biologique et d'estimer leur potentiel pour l'inférence de fonction. Cette annotation passe par l'intégration d'un ensemble de données hétérogènes et indépendantes. Mon travail vise également à proposer un système qui permet l'exploration et l'interprétation biologique simultanément de toutes ces données à large échelle. Ainsi ce système devait permettre d'une part d'analyser un sous ensemble de clusters impactés par un type de stress afin d'adresser des questions spécifiques autour de la réponse de la plante à une condition de stress particulière, telles que les processus moléculaires mis en place, les gènes impliqués et leurs caractéristiques. D'autre part, il devait permettre d'analyser l'ensemble des clusters obtenus afin d'identifier les processus de réponses globales ou communes aux stress chez la plante étudiée.

Au cours de ce chapitre je décrirai d'abord les données utilisées pour l'annotation des clusters, la spécificité de ces données ainsi que les analyses et procédures mises en place. La section suivante permettra de montrer comment j'ai exploité ces ressources et ces procédures pour l'annotation et l'analyse d'un sous ensemble de clusters dont les gènes sont impactés par une seule condition de stress. Je montrerai par la suite la procédure d'automatisation de ces analyses et le système d'information mis en place pour permettre d'étudier l'ensemble des clusters disponibles et

d'explorer au mieux toutes les ressources. Finalement je conclurai sur l'apport de cette analyse et ses spécificités, ainsi que sur les problématiques rencontrées et les limites de l'analyse.

II. Annotation et caractérisation des groupes de gènes

1. Choix des données et ressources exploitées pour le projet

Généralement l'annotation fonctionnelle d'un groupe de gènes nécessite de rassembler pour chaque gène un faisceau d'informations qui peuvent être relatives à sa structure, sa régulation, sa localisation subcellulaire, son activité biochimique ou biologique, ainsi que les interactions entretenues avec d'autres acteurs moléculaires. Leur annotation passe donc par l'intégration de plusieurs données hétérogènes afin d'avoir une vision globale et de donner des arguments à propos de la fonction biologique dans laquelle ils sont impliqués. Dans la littérature, peu d'études permettent l'intégration de toutes ces ressources simultanément. Parmi celles qui le permettent, rares sont celles qui intègrent les données hétérogènes à large échelle pour caractériser un ensemble composé de milliers de gènes. Beaucoup de ces études souffrent également de l'hétérogénéité de l'origine des données intégrées et du manque de spécificité du contexte étudié.

Afin de pratiquer une méta-analyse pour les caractérisations fonctionnelle et relationnelle des groupes de gènes obtenus, nous avons collecté des informations de différents types en essayant de contrôler au mieux leur qualité. Nous avons utilisé également les ontologies pour nous affranchir de l'hétérogénéité sémantique des données.

a. *Données d'annotation Gene Ontology*

Pour la description des fonctions des gènes, nous avons exploité le vocabulaire contrôlé et standardisé Gene Ontology (GO) associé à la version 10 de l'annotation du génome d'*A. thaliana* (TAIR R10). Les trois branches principales (processus biologique BP, composant cellulaire CC et fonction moléculaire MF) de cette classification ont été utilisées pour caractériser les groupes de gènes.

Nous avons utilisé plus particulièrement la GO Slim. Les termes GO Slim sont une version réduite et générale des termes GO, très utiles pour avoir une vision globale de l'ontologie sans rentrer dans le détail des termes très spécifiques. Le nombre de termes GO Slim par ontologie BP, MF et CC est

respectivement de 13, 14 et 15 termes. Ces termes sont également plus intéressants pour l'inférence de fonction car si on descend trop bas dans la hiérarchie GO, les fonctions des gènes deviennent plus spécifiques et de moins en moins de gènes ont une annotation.

b. *Localisations subcellulaires*

En plus des termes d'annotation GO du domaine compartiment cellulaire CC, nous disposons d'autres données de localisation subcellulaires indépendantes. Ces données sont extraites de FLAGdb^{++} (Dèrozier *et al*. 2011) et correspondent à des données prédites grâce à la détection de signaux d'adressage au chloroplaste, à la mitochondrie, au réticulum endoplasmique ou au noyau. Ces prédictions ont été faites en utilisant un pipeline combinant un panel de logiciels spécifiques : Predotar (Small *et al*. 2004), ChloroP (Emanuelsson *et al*. 1999) et PSORT (Horton *et al*. 2007).

c. *Motifs cis-régulateurs*

Nous nous sommes intéressés également à l'étude de la présence et de la nature des motifs conservés au sein des régions 5' des gènes, appelés motifs cis-régulateurs. Pour cela le jeu de séquences promotrices des gènes des clusters étudiés a été extrait de la base de données FLAGdb^{++}. Il comprend 27 025 séquences d'une longueur de 1 000 paires de bases en amont du site d'initiation de la transcription (TSS) lorsqu'il est connu ou à défaut en amont du codon ATG. Une liste de 400 motifs cis-régulateurs connus chez les plantes a été extraite également des deux bases de données PLACE et AGRIS.

d. *Familles de gènes*

Pour enrichir les métadonnées, j'ai pris en considération deux groupes de familles de gènes significativement impliqués dans la réponse au stress à savoir les facteurs de transcription et les hormones. Les facteurs de transcription jouent un rôle important dans la régulation et le contrôle de plusieurs voies biologiques y compris les voies de réponse aux stress (Horan *et al*. 2008). Les familles de facteurs de transcription ont été obtenues à partir du site REGULATORS (http://urgv.evry.inra.fr/projects/arabidopsis-TF/) (Castrillo *et al*. 2011). Dans ce travail, 2 260 TF ont été classés en 79 familles basés sur la similarité de séquence du domaine de liaison à l'ADN. Une hormone est définie comme un médiateur qui agit à très faible concentration et à distance de son lieu de synthèse pour contrôler une réponse physiologique déterminée. La signalisation hormonale

est ainsi une stratégie importante mise en place par la plante en réponse au stress. L'acide abscissique est un exemple d'hormone de signalisation de stress qui permet de moduler le degré d'ouverture des stomates. Une liste de 695 gènes impliqués dans la réponse aux hormones a été extraite de la base de données AHD2.0 (Jiang *et al*. 2010) et a été utilisée pour l'annotation des clusters.

e. *Interactions protéine-protéine*

La fonction d'une entité biologique résulte de ses relations avec les autres acteurs du système. C'est par cette vue d'ensemble que l'étude fonctionnelle des gènes sera possible. L'objectif de l'intégration de ces données PPI est d'identifier des interactions directes au sein des clusters et de les enrichir avec des partenaires fonctionnels n'ayant pas le même profil transcriptomique, mais du fait de leur interaction physique, ils peuvent être potentiellement impliqués dans un même processus biologique. Cet enrichissement m'a permis de passer d'un groupe de coexpression à un groupe fonctionnel plus pertinent.

Dans ce projet je dispose de trois sources différentes de données d'interactions protéine-protéine.

i. Données AI1 (Arabidopsis Interactome 1)

Les données PPI nommées AI1 sont des données expérimentales de haute qualité. Elles sont issues de la première expérience de criblage (ou mapping) d'interactions protéine-protéine à large échelle chez la plante modèle *A. thaliana* (Arabidopsis Interactome Mapping Consortium, 2011). Pour réaliser ce criblage, 8 750 ORF (Open Reading Frames) d'*A. thaliana*, représentant 32% des gènes prédits comme codant pour des protéines, ont été utilisés. Toutes les combinaisons possibles des couples deux à deux ont été testées (soit 81 millions de tests) en utilisant un pipeline binaire de criblage d'interactome à large échelle basé sur la technique de double hybride chez la levure (Y2H).

Le jeu de données obtenu nommé « AI1 » est composé de 6 475 interactions impliquant 2 836 protéines différentes.

ii. Données LCI (Literature Curated Interaction)

Ce sont des données expérimentales issues de la littérature et collectées de plusieurs sources ou bases de données notamment les bases BioGRId, IntAct, TAIR et BIND.

Etant donné que ces données d'interactions sont issues de sources différentes, naturellement elles ont été détectées par différentes techniques telles que : Double Hybride, Pull down, CoIP, anti-tag CoIP,

études enzymatique etc. Cependant la méthode de détection d'interaction la plus utilisée, reste la méthode de double hybride (Y2H).

L'union de toutes ces données expérimentales collectées mène à la construction d'un jeu de données composé de 7 945 interactions uniques concernant la plante modèle *A. thaliana*.

iii. *Données PP (PAIR Predicted)*

Ces données correspondent à 145 494 interactions prédites et elles ont été extraites de la base de données PAIR (Predicted Arabidopsis Interactome Resource) (Lin *et al*. 2011). Les interactions sont prédites grâce à un système SVM (Support Vector Machine) qui permet l'intégration de données d'homologie, de co-localisation, de coexpression, de partage de domaines d'interaction, de partage d'annotations GO, de similarité de profils phylogénétiques et la présence d'interologues. Le système SVM a été entrainé sur un jeu de données d'apprentissage dont les interactions sont connues. En effet, c'est un ensemble d'interactions nommées GSP (Gold Standard Positives) collecté des bases de données IntAct, BioGrid, TAIR et BIND. Selon leurs estimations, ces données prédites sont censées couvrir 24,47% de l'ensemble de l'interactome d'*A. thaliana* avec une précision de 43,52%.

J'attribue plus de confiance aux données d'interactions expérimentales de type AI1 et LCI qu'aux données prédites (PP) lors des analyses, c'est pourquoi je m'appuie principalement sur ces données expérimentales, qui décrivent 12 741 interactions non redondantes.

f. *Données d'interactions TF-cibles*

Les données de AGRIS (Arabidopsis Gene Regulatory Information Server) ont été utilisées pour lier les facteurs de transcription (FT) à leurs gènes cibles. Parmi les 11 352 interactions directes entre facteurs de transcription et gènes cibles hébergées par la base de données AtRegNet d'AGRIS, seules les 769 interactions confirmées ont été prises en compte. Comme décrit par Yilmaz *et al*. (2011), une interaction est classée comme « confirmée » quand elle est validée par deux ou plusieurs approches expérimentales.

g. *Fouille de littérature*

Une procédure de fouille de littérature a été effectuée afin de déterminer une liste de gènes ayant une preuve expérimentale de leur implication dans la réponse au stress, que nous appelons « liste BiblioStress ». Cette liste a été obtenue à partir du fichier d'annotation GO, de la base de données TAIR, qui a été filtré sur les codes évidences afin de ne conserver que les gènes annotés

expérimentalement par les termes GO Slim « réponse au stress » et « réponse au stimuli biotique ou abiotique » et associés à la littérature scientifique. De cette manière, 1 914 références bibliographiques ont été exploitées conduisant à l'identification de 2 580 gènes connus pour leur implication dans la réponse au stress.

Le chevauchement des clusters de coexpression obtenus avec cette liste est susceptible de conforter d'une part sur la spécificité des données analysées et d'autre part met l'accent sur la présence de partenaires fonctionnels de haute qualité tout en ayant l'information sur leur dynamique de réponse.

h. *Définition des gènes orphelins*

Après 10 versions d'annotation d'*A. thaliana*, une large proportion de gènes code pour des protéines encore inconnues (Swarbreck *et al*. 2008). Ces gènes sont nommés gènes « orphelins » car orphelins de fonction (Domazet-Lozo et Tautz, 2003 ; Fukuchi et Nishikawa, 2003). Parmi ces gènes orphelins, la majorité est qualifiée comme étant des « gènes exprimés » puisque la disponibilité de séquences de transcrits (EST ou cDNA) apparentées représente une preuve de leur transcription. Les autres sont étiquetés comme « gènes hypothétiques » et leurs structures intron-exon viennent des logiciels de prédiction *ab initio* (Haas *et al*. 2005). La proportion de gènes orphelins varie en fonction de la définition elle-même d'un gène orphelin, cette définition variant d'une étude à une autre. L'étude de Horan *et al.* (2008) montre que la définition de ce qu'est un gène orphelin est ambigüe et classe les méthodes définissant les gènes orphelins en deux types d'approches. La première approche de similarité considère une protéine en tant qu'orpheline si elle ne montre aucune similitude de séquence ou de structure avec des protéines caractérisées fonctionnellement dans les bases de données de référence (Boeckmann *et al.* 2003; Leinonen *et al*. 2004). En revanche, l'approche empirique plus conservatrice définit comme orphelines toutes les protéines qui n'ont pas de preuve expérimentale directe comme support de leurs fonctions spécifiques. Selon l'étude de Horan, plus de 8 600 gènes sont encore orphelins de fonction (fonction moléculaire). De plus si on considère non seulement la fonction biochimique des protéines, mais aussi leur rôle physiologique, alors selon certaines définitions les gènes orphelins de fonction deviennent une large majorité puisque chez *A. thaliana*, seulement 14% des gènes ont une fonction biologique caractérisée par des approches expérimentales (TAIR).

Dans cette étude, je donne une définition plus stricte aux gènes orphelins qui permet de restreindre la proportion de ces gènes à ceux qui n'ont aucun indice autour de leur fonction biochimique et biologique et dont la structure ne contient aucun domaine connu défini par InterPro (Hunter *et al*.

2009). Les gènes qui ont une annotation pour une seule branche de l'ontologie GO, par exemple un gène annoté en BP mais pas en MF et CC, ne sont pas considérés comme orphelins mais plutôt comme des gènes mal caractérisés. Selon cette définition, à partir de la description fonctionnelle des 33 602 gènes de TAIR, un gène a été considéré comme orphelin de fonction si sa description est conforme à ces critères : (i) le gène n'est associé à aucune annotation GO ou (ii) le gène est associé aux termes « protéine inconnue » ou « protéine hypothétique » pour les ontologies « processus biologique » et « fonction moléculaire » et (iii) aucun domaine connu défini par InterPro n'est associé à la protéine.

Selon ces critères et cette définition restrictive, 5 105 gènes ont été identifiés comme orphelins de fonction soit environ 15% des gènes d'*A. thaliana*. Le tableau 2 résume le nombre de gènes orphelins dont la fonction pourrait être éclairée grâce à cette ressource. Sur les 5 105 gènes orphelins du génome d'*A. thaliana*, 2 165 (soit 42%) ont été identifiés comme étant impactés par le stress et font partie de notre jeu de données analysées. Cette valeur est proche de la valeur attendue de 50%, compte tenu de la proportion de l'ensemble des gènes (17 264 gènes analysés versus 34 042 gènes dans la référence). De plus, une proportion équivalente de gènes orphelins a été trouvée dans les conditions de stress biotiques, abiotiques et les 18 catégories de stress considérés séparément. La stabilité du nombre de gènes orphelins souligne une distribution régulière du niveau de connaissances dans l'ensemble des catégories de stress, même si le nombre d'expériences n'est pas équivalent.

Tableau 2 : Nombre de gènes orphelins total dans les clusters de coexpression par catégorie de stress.

Catégorie de stress	Nbr gènes	Nbr orphelins	% orphelins
Bactéries Biotrophes	11817	1553	13.1%
Bactéries Nécrotrophes	11030	1395	12.6%
Fungi	9705	1273	13.1%
Nématodes	7487	973	13.0%
Oomycete	5591	682	12.2%
Stifénia	1565	172	11.0%
Virus	11685	1567	13.4%
Rhodococcus	1965	236	12.0%
Autres Biotiques	3803	459	12.1%
Gamma	5419	639	11.8%
Metaux lourds	10533	1383	13.1%
Sécheresse	8167	1051	12.9%
Nitrogène	13807	1818	13.2%
Stress Oxydatif	10027	1321	13.2%
Sel	5786	736	12.7%
Température	11199	1443	12.9%
UV	7903	1004	12.7%
Autres Abiotiques	3944	467	11.8%

2. Description globale des analyses

Après avoir sélectionné les différents types de données informatifs pour l'annotation fonctionnelle avec le soin de contrôler au mieux leurs origines et leurs qualités, l'enjeu consiste maintenant à les utiliser pour annoter fonctionnellement les clusters et à croiser ces multiples sources. La difficulté réside dans le fait qu'elles décrivent des caractéristiques à différents niveaux de l'organisation de la cellule (génome, transcriptome, protéome, métabolome, etc.) et qu'elles caractérisent les gènes mais également la relation entre ces gènes. Comme montré dans l'introduction, l'apport de chaque type de

données par rapport à l'annotation fonctionnelle, leur qualité et disponibilité sont variables. Pour toutes ces raisons et étant donné l'objectif d'évaluer la pertinence biologique des clusters de coexpression, l'idée principale est que ces données supplémentaires et hétérogènes complètent les données d'expression et conduisent à une description plus précise des relations fonctionnelles qui existent entre les gènes d'un cluster. Le choix a donc été porté sur l'intégration de chaque type de données avec les données de coexpression.

L'analyse de l'enrichissement de ces groupes de gènes en certaines catégories d'informations ou caractéristiques peut être un bon moyen pour révéler les biais et les spécificités de ces groupes. Pour cela, des outils et des pipelines ont été mis en place afin de caractériser les clusters en termes de localisation subcellulaire, classification fonctionnelle, présence de gènes connus dans la littérature pour répondre au stress, la présence de facteurs de transcription, le contenu des promoteurs en sites de fixation de facteurs de transcription (TFBS) relatifs à la réponse aux stress (boîtes WRKY, whirly, GCC etc.). J'ai réalisé des tests statistiques pour détecter la présence de biais significatifs dans les caractéristiques mesurées. J'ai mis en place un test de permutation pour tester l'enrichissement des clusters de coexpression en interactions PPI et pour les autres données, j'ai réalisé un test hypergéométrique afin de comparer la proportion de gènes associés à une caractéristique donnée (par exemple un terme GO) dans le cluster analysé par rapport à la proportion de gènes associés à cette même caractéristique dans le génome.

Chacune de ces tâches est décrite dans un paragraphe ci-dessous.

a. *Détection de motifs cis-régulateurs*

Pour l'identification de la présence de motifs au sein des promoteurs des groupes de gènes analysés, j'ai utilisé un outil appelé « PLMdetect » développé par Virginie Bernard durant son projet de thèse au sein de l'équipe (Bernard *et al*. 2006 et 2010). L'avantage de cet outil est de limiter le taux de faux positifs en prenant en considération les positions préférentielles entre les motifs étudiés et la position du TSS. Cette méthode permet de déterminer l'emplacement préférentiel de chaque motif par rapport au TSS ainsi que sa fenêtre fonctionnelle déduite des limites du pic de la région dans laquelle le motif est sur-représenté (figure 3). Un motif identifié par cette méthode est un motif sur-représenté à un endroit précis par rapport au TSS et il est nommé PLM (Preferentially Located Motif). Pour évaluer si un PLM donné est sur-représenté dans un cluster par rapport à tout le génome, un test binomial a été réalisé en comparant le nombre de gènes de ce cluster contenant ce PLM au nombre de gènes contenant ce PLM dans la même fenêtre fonctionnelle à l'échelle du

génome. Les PLM avec une probabilité critique inférieure à 0,01 ont été considérés comme significativement sur-représentés dans le cluster considéré.

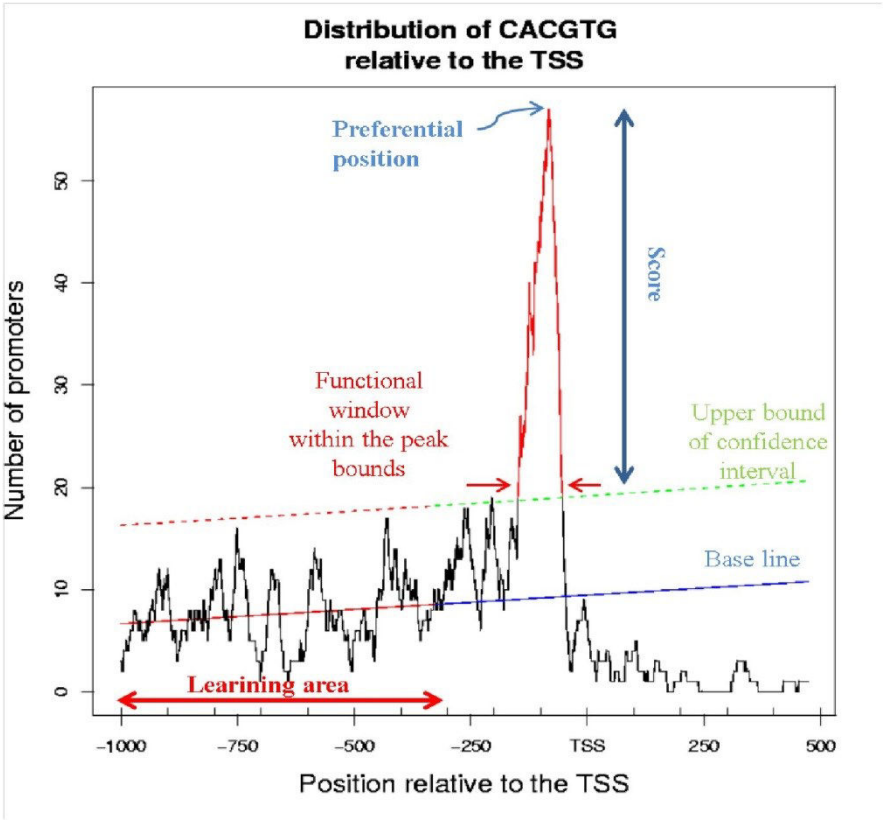


Figure 3 : Détection de motif grâce à l'outils PLMdetect.
Exemple de détection d'un pic correspondant à une localisation préférentielle du motif CACGTG par rapport au site TSS au sein des promoteurs analyses. Figure extraite de Bernard et al. 2010.

b. *Recherche d'interactions PPI au sein de chaque cluster*

J'ai développé un pipeline qui permet d'identifier toutes les interactions PPI impliquant un couple de gènes appartenant à un cluster. Le nombre d'interactions intra clusters trouvées est ainsi comptabilisé. Les interactions trouvées au sein de chaque cluster sont caractérisées selon leurs types: données AI1, LCI ou PP. L'union de ces trois types d'interactions est aussi déterminée. C'est ce qu'on appelle les interactions uniques.

Dans le cas des interactions protéine-protéine (PPI), j'ai utilisé une approche de permutation pour évaluer l'importance du nombre de PPI dans un cluster et vérifier s'il y a un enrichissement du nombre d'interactions au sein de chaque cluster par rapport au hasard. Ainsi, pour chaque cluster, le nombre de PPI impliquant des couples de gènes appartenant au cluster est compté et ce nombre est noté k. Ensuite, 1 000 clusters de la même taille que le cluster analysé sont construits au hasard et le nombre de PPI impliquant des couples de gènes au sein de chacun de ces groupes aléatoires est

compté. J'ai récupéré une probabilité critique en calculant la proportion de groupes aléatoires ayant un nombre de PPI supérieur ou égal à k et j'ai considéré les clusters ayant une probabilité critique inférieure à 5% comme significativement enrichis en interactions PPI.

c. *Analyses d'enrichissement*

Les enrichissements de clusters en termes GO Slim, localisations subcellulaires, facteurs de transcription, hormones ou en gènes de la liste BiblioStress ont été évalués en utilisant un test hypergéométrique. Ce test permet de comparer le nombre de gènes associés à chaque métadonnée étudiée dans le cluster par rapport à sa valeur attendue dans le génome (34 042 gènes). La sur-représentation est déclarée statistiquement significative lorsque la probabilité critique est inférieure à 0,01.

d. *Construction de réseaux d'interactions*

Afin d'enrichir les clusters de coexpression en partenaires fonctionnels qui ne sont pas forcément transcriptionnellement régulés et notamment pour identifier des régulateurs potentiels, nous avons construit un réseau de gènes en croisant les données de coexpression, les données PPI et les données d'interactions TF-cibles. Les interactions prises en considération correspondent uniquement aux données expérimentales.

Les interactions protéine-protéine et les interactions TF-cibles ont été combinées aux clusters de coexpression grâce à un programme Perl qui permet d'identifier toute interaction dans laquelle au moins un gène du cluster analysé est impliqué. Les réseaux de gènes obtenus sont visualisés à l'aide de Cytoscape (Shannon *et al*. 2003). Plusieurs attributs des nœuds (gènes) et des arêtes (interactions) sont intégrés tels que l'annotation GO, le type d'interactions ou le numéro du cluster auquel le gène appartient. A la fin de l'analyse, je génère pour chaque cluster un bilan résumant tous les résultats obtenus.

III. Caractérisation fonctionnelle des clusters de coexpression modulés par la FLAGELLINE

Afin de mettre au point la démarche d'intégration des données hétérogènes pour la caractérisation fonctionnelle et l'interprétation biologique de groupes de gènes, j'ai été impliquée dans une collaboration avec l'équipe MAP-kinases «Mitogen Activated Protein Kinases» de l'IPS2. Les MAP kinases sont d'importants régulateurs de la réponse immunitaire chez les animaux et les plantes et les recherches de cette équipe se concentrent sur l'étude des cascades de MAPK activées en réponse à des stress biotiques et abiotiques. Dans le cadre de ce projet, les collaborateurs s'intéressaient plus particulièrement à l'étude des voies de signalisation MAPK activées dans la plante suite à une attaque microbienne mimée par la reconnaissance de la protéine bactérienne FLAGELLINE (FLG22). Chez *A. thaliana*, la reconnaissance de FLG22 active au moins deux voies de signalisation MAPK. Dans ces deux voies, les MPK3, MPK4 et MPK6 jouent un rôle important mais leurs contributions spécifiques et leur coopération dans ces événements de signalisation restent en grande partie incertaines.

Les données correspondent à des transcriptomes d'*A. thaliana* Col-0 (sauvage) et mutées pour les MAP kinases 3, 4 et 6 en présence et en absence de FLG22. Le support utilisé est la puce deux couleurs CATMAv6.2 et le plan d'expérience comprend 7 comparaisons réalisées en dye-swap avec 3 répétitions biologiques pour chaque comparaison (Figure 4). L'objectif de cette étude était d'élucider les cascades MAPK 3, 4 et 6 ainsi que les fonctions de régulation des gènes dans la réponse au stress induit par les pathogènes. La complexité de l'organisation et le chevauchement des différentes cascades induites par la reprogrammation transcriptionnelle modulées par ces MAPK, nécessite une analyse bioinformatique et la mise en place d'une approche adéquate afin d'identifier les réseaux de gènes modulés par ces MAPK spécifiques et communs ainsi que leurs caractéristiques.

Dans cette section je décriai les méthodes développées pour l'analyse de ces données, ainsi que les résultats obtenus et qui ont été valorisés par la publication d'un article (Frei Dit Frey *et al*. 2014. Annexe A).

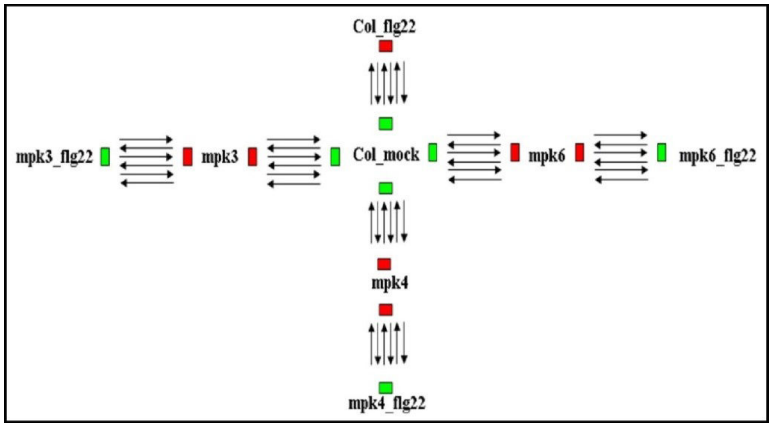


Figure 4 : Plan d'expérience de la réponse au traitement FLG22 chez les mutants *mpk*.
Plan d'expérience pour analyser les changements transcriptionnels induits par la FLG22 chez les mutants mpk3, mpk4 et mpk6. Un total de 21 expériences en dye-swap a été effectué : 7 comparaisons avec 3 répétitions biologiques pour chaque comparaison.

1. Analyse des données transcriptomiques et construction des clusters de coexpression

J'ai considéré les sondes qui ont été exprimées de façon différentielle dans au moins une comparaison parmi les 7 étudiées, en fonction de la probabilité critique ajustée par la procédure de Bonferroni pour limiter le nombre de faux positifs. Cela a conduit à l'identification de 4 378 sondes correspondant à 4 298 gènes, la correspondance sonde-gène ayant été effectuée grâce aux informations contenues dans la base de données FLAGdb^{++}. Le tableau 3 représente le nombre de gènes surexprimés et le nombre de gènes sous-exprimés dans chaque comparaison. Ces résultats montrent que la mutation de chaque *MAPK* impacte de façon différente l'expression des gènes. En absence de FLG22, environ la moitié des gènes qui sont exprimés de manière différentielle dans *mpk3* (51% de gènes sous-exprimés et 60% de gènes surexprimés) affichent une régulation similaire dans *mpk4* (tableau 3, figure 5).

Tableau 3 : Résultats de l'analyse différentielle par comparaison deux à deux des conditions mpk.

Comparaison	Gènes surexprimés	Gènes sous-exprimés
mpk3 vs Col-0	305	191
mpk4 vs Col-0	969	265
mpk6 vs Col-0	45	16
mpk3 + flg vs mpk3	1519	877
mpk4 + flg vs mpk4	1442	634
mpk6 + flg vs mpk6	1468	690
Col-0 + flg vs Col-0	1529	862

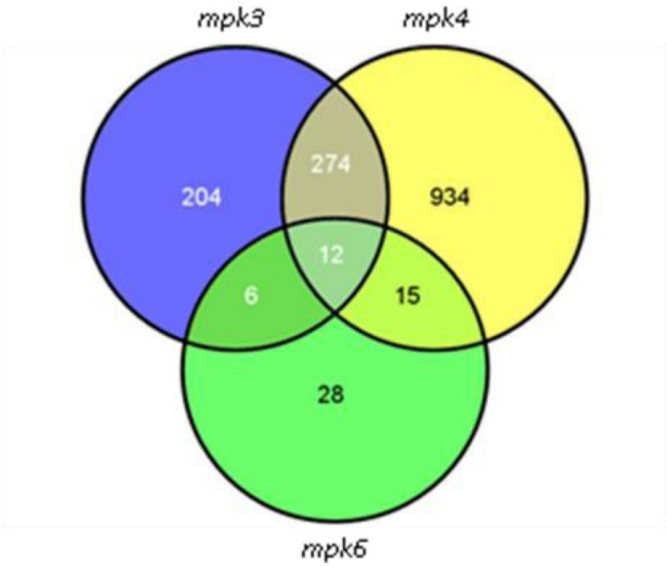


Figure 5 : Représentation du nombre de gènes différentiellement exprimés.
Diagramme de venn représentant le nombre de gènes différentiellement exprimés (induits et réprimés) par mpk3, mpk4 et mpk6.

Pour aller au-delà de cette analyse et afin de tirer plus de connaissances biologiques à partir de ces données transcriptomiques, j'ai effectué l'analyse de coexpression avec un modèle de mélange de gaussiennes multidimensionnelles. Le critère BIC a permis de sélectionner un mélange de 29 composantes ou clusters (figure 6). Les gènes ont été ensuite attribués au cluster pour lequel la probabilité conditionnelle est la plus élevée et l'interprétation a été effectuée seulement pour les sondes pour lesquelles cette probabilité est supérieure à 0,878. Ce seuil est déterminé pour ce jeu de données par la règle de classification MFDR, dont l'objectif est de classer le plus grand nombre de gènes sous la contrainte que l'espérance de la proportion de sondes mal classées, soit contrôlée à un niveau de 5% (figure 7). Ainsi selon ces critères, 1 876 gènes, parmi une liste de 4 298 gènes différentiellement exprimés au moins une fois dans une des 7 comparaisons étudiées, ont été classés dans 29 clusters. La figure 8 montre un exemple de profils d'expressions de gènes regroupés dans le même cluster (cluster numéro 10). Les profils d'expressions des 29 clusters obtenus par cette analyse sont représentés dans la figure annexe B.

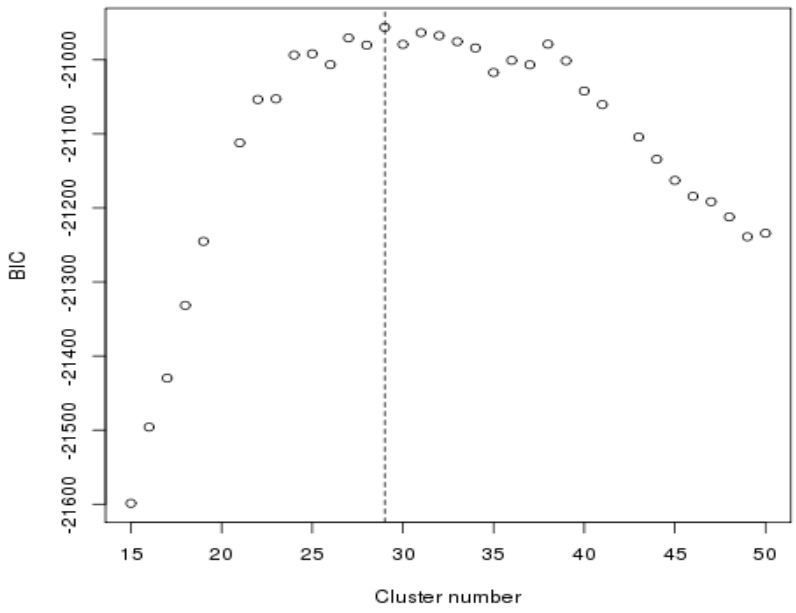


Figure 6 : Evolution du critère BIC en fonction du nombre de composantes.
Le modèle atteint une valeur maximale a un nombre de clusters égal à 29.

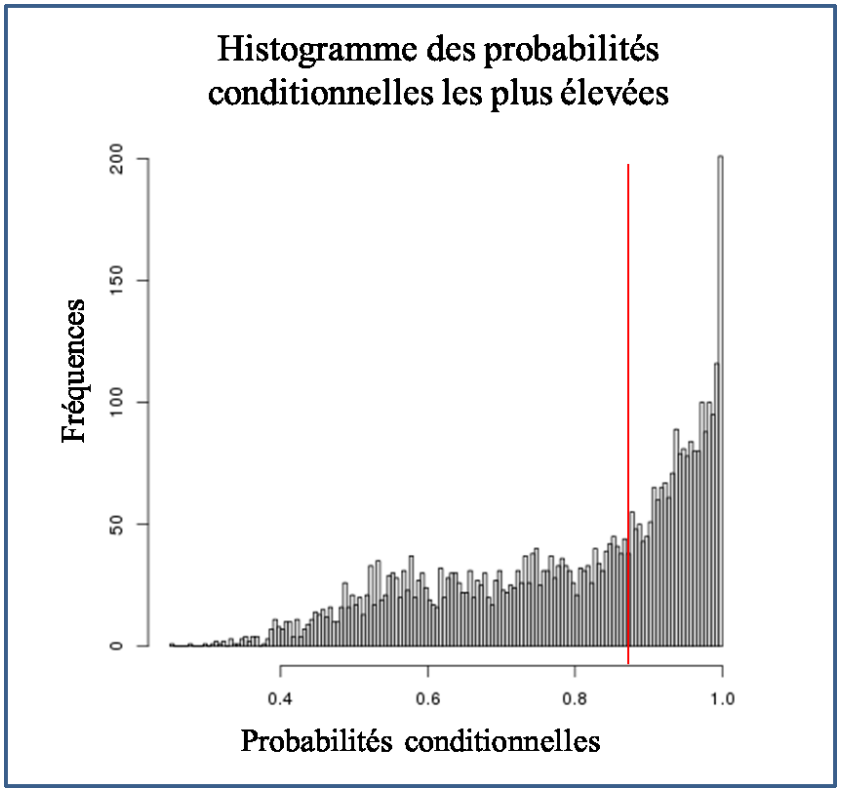


Figure 7 : Histogramme des probabilités conditionnelles les plus élevées des gènes.
La règle de classification MAP permet de classer tous les gènes. En contrôlant le taux de faux positifs à 5%, seuls les gènes pour lesquels la probabilité a posteriori la plus élevée est supérieure au seuil 0.878 sont maintenus dans un cluster.

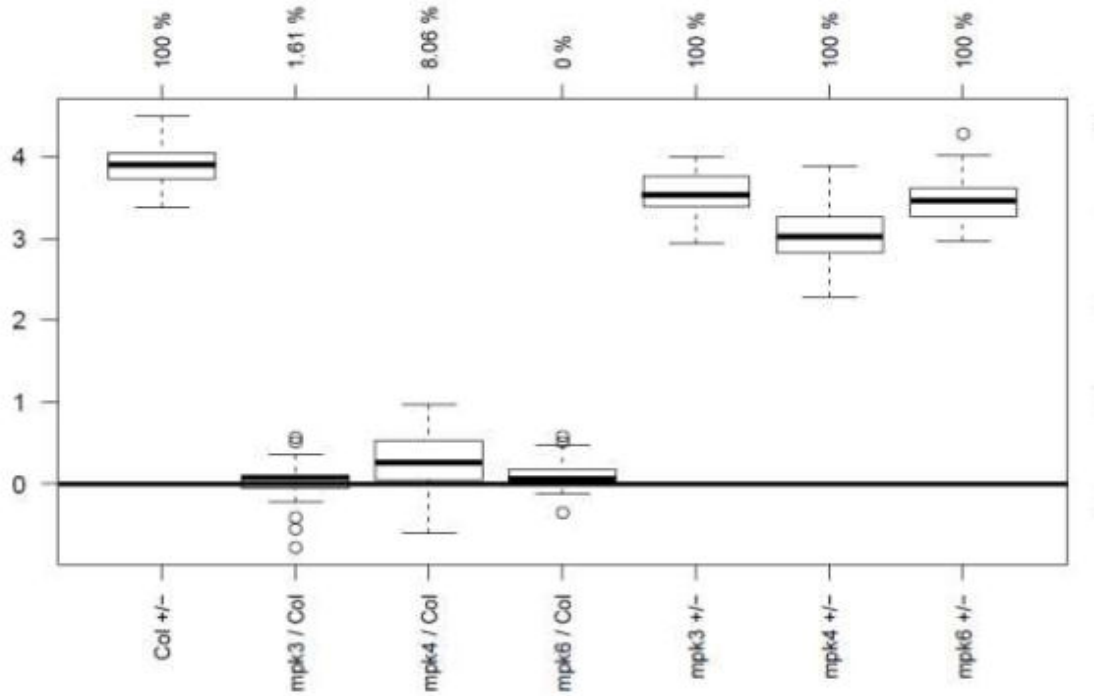


Figure 8 : Exemple de profils d'expression des 62 gènes regroupés dans le cluster numéro 10.
L'axe des ordonnées représente des logs ratios. L'axe des abscisses représente les comparaisons suivantes: col-0 + flg22 vs col-0, mpk3 vs col-0, mpk4 vs col-0, mpk6 vs col-0, mpk3 + flg22 vs mpk3, mpk4+ flg22 vs mpk4, mpk6 + flg22 vs mpk6. Les profils sont représentés comme des boites à moustaches, ou le bas et le haut de la boite sont les premier et troisième quartiles et la ligne à l'intérieur de la boite est la médiane. Les valeurs extrêmes sont représentées par un point. Le pourcentage de gènes du cluster identifiés comme différentiellement exprimés (Bonferroni p-value <0,05) dans chaque comparaison est représenté en haut.

2. Caractérisation fonctionnelle des groupes de gènes

a. *Analyse globale des clusters*

Afin de caractériser les gènes des 29 clusters de coexpression obtenus, les outils et les pipelines mis en place et décrits dans le paragraphe « II. Annotation et caractérisation des groupes de gènes » ont été appliqués. Les tests statistiques décrits précédemment ont été également réalisés pour détecter la présence de biais significatifs dans les caractéristiques mesurées.

Les premiers résultats montrent que 75,8% des clusters ont un biais significatif dans la localisation subcellulaire, 65,5 % sont enrichis en gènes connus dans la littérature pour leur implication dans la réponse aux stress, 86,2 % sont enrichis en terme « Réponse au stimulus stress abiotique ou biotique » ou « réponse au stress » de l'ontologie BP de la classification GO Slim. L'enrichissement des clusters en termes liés au stress, bien que cela ne soit pas étonnant puisque le projet porte sur l'analyse de changements transcriptionnels des gènes en condition de stress biotique, est rassurant vis-à-vis de la qualité des clusters obtenus et de leur lien direct à la réponse au stress. De plus ces

clusters sont également enrichis en termes liés au transport et à la transduction du signal ce qui est très cohérent puisque ces deux termes sont souvent associés au terme de stress dans la littérature. Les enrichissements très homogènes en termes de l'ontologie compartiment cellulaire permettent d'identifier également trois clusters qui correspondent à des clusters de gènes mitochondriaux ou chloroplastiques et qui suscitent l'intérêt de certains biologistes pour l'étude de la dynamique de ces compartiments en réponse aux stress.

L'approche de clustering a permis l'identification de groupes de gènes présentant des profils d'expression similaires. De tels groupes sont supposés être contrôlés par les mêmes régulateurs en amont. Par conséquent, j'ai utilisé les promoteurs des gènes dans chaque cluster pour évaluer l'enrichissement des éléments cis-régulateurs connus au sein de ces groupes. L'analyse des promoteurs via l'outil « PLM detect » a révélé la présence de plusieurs motifs cis-régulateurs et j'ai détecté la sur-représentation de ces motifs par rapport au génome au sein de 22 clusters parmi les 29 analysés. Le tableau 4 résume toutes les données relatives à l'analyse de l'enrichissement des éléments cis-régulateurs, la taille des clusters et le nombre de gènes ayant ces motifs au sein de leurs promoteurs. Beaucoup de clusters des gènes induits par la FLG22 sont enrichis en promoteurs contenant des motifs (TTGAC) ou W-box qui sont les sites de fixation des facteurs de transcription WRKY. Les WRKY sont connus pour être des substrats des MAPK et jouent un rôle important dans la reprogrammation transcriptionnelle des plantes liée au stress (Pandey et Somssich 2009). De nombreux autres motifs sont également identifiés tels que les motifs (CNGTTR) MYB souvent associés avec les motifs W-box, ou encore les sites (GATT) de fixation du facteur de transcription EIN3 retrouvé dans 5 clusters. Ce facteur de transcription est connu pour être régulé par MPK3 et 6 directement par une phosphorylation.

Tableau 4 : PLMs détectés au sein de chaque cluster et déterminés comme sur-représentés par rapport au génome.

Cluster	Taille cluster	Nbr promoteurs	Motif	Fenêtre fonctionnelle		Promoteurs ayant le motif			P-value
						cluster		génome	
				début	fin	nbr	%	%	
Cluster_1	4	Pas de motif significativement sur-représenté							
Cluster_2	14	Pas de motif significativement sur-représenté							
Cluster_3	22	Pas de motif significativement sur-représenté							
Cluster_4	16	13	CAACA	-140	-59	8	61.54	18.94	1.06E-04
Cluster_5	26	Pas de motif significativement sur-représenté							
Cluster_6	32	32	KCACGW	-320	-128	9	28.12	13.96	9.59E-03
Cluster_7	21	21	WGATAR	-305	-266	8	38.1	12.68	5.89E-04
Cluster_8	55	50	GAGAGA	-136	217	35	70	44.38	7.13E-05
			GAGAC	-132	133	26	52	32.1	1.11E-03
			TAACAAR	-359	-256	10	20	8.96	4.14E-03
Cluster_9	42	Pas de motif significativement sur-représenté							
Cluster_10	61	61	ACTTTG	-302	-163	16	26.23	11.32	3.21 E-04
			ACGT	-375	-56	44	72.13	52.15	4.54E-04
Cluster_11	33	32	GATT	-54	-28	15	46.88	26.94	4.55E-03
Cluster_12	47	46	TTGAC	-237	-93	24	52.17	30.89	7.98E-04
			WAACCA	-15	44	11	23.91	8.46	3.38E-04
			CANNTG	-259	-236	12	26.09	9.43	2.37E-04
			YAACKG	-253	-192	7	15.22	6.32	7.64E-03
Cluster_13	38	37	ACACNNG	-191	-14	14	37.84	18.77	1.76E-03
			TTGAC	-142	-24	23	62.16	25.54	5.54E-07
Cluster_14	55	54	ACGTG	-199	4	19	35.19	13.84	2.03E-03
			CAACA	-74	34	24	44.44	23.53	2.06E-04
			GAGAC	-35	42	15	27.78	13.84	2.03E-03
Cluster_15	58	56	TTGAC	-175	-9	34	60.71	33.25	7.07E-06
			WGATAR	-304	-268	15	26.79	11.79	5.52E-04
			CNGTTR	-223	-147	15	26.79	15.48	9.19E-03
			GATT	-153	-143	15	26.79	13.48	2.33E-03
Cluster_16	74	72	ACGTG	-166	53	29	40.28	21.05	6.17E-05
			TTATCC	-102	-22	10	13.89	4.4	3.01 E-04
			AMCWAMC	-6	39	9	12.5	4.94	2.80E-03
Cluster_17	84	Pas de motif significativement sur-représenté							
Cluster_18	74	73	TTGAC	-149	-31	41	56.16	25.91	1.10E-08
			TTATCC	-258	-120	12	16.44	7.08	1.73E-03
			GATT	-249	-235	23	31.51	18.56	2.48E-03
Cluster_19	109	106	GAGAGA	-16	140	46	43.4	31.76	4.44E-03
			CATGTG	-201	-12	22	20.75	8.15	1.25E-05
			CACATG	-206	-10	22	20.75	8.38	1.96E-05
			GCCAC	-96	26	19	17.92	9.55	2.37E-03
			WAACCA	-167	-128	18	16.98	8.77	2.11E-03
			GATAAG	-28	48	12	11.32	3.79	1.96E-04
			ACACNNG	-147	-108	13	12.26	4.51	3.04E-04
Cluster_20	92	89	TTGAC	-180	-31	57	64.04	31.4	6.27E-11
			ACTCAT	-364	-202	16	17.98	8.76	1.72E-03
			ACGT	-277	-255	11	12.36	5.52	3.60E-03
			AMCWAMC	-6	32	11	12.36	4.37	5.01 E-04

Cluster_21	63	Pas de motif significativement sur-représenté							
Cluster_22	66	66	GAGAGA	-124	214	38	57.58	43.7	8.62E-03
			CWWWWWWWWG	-249	-115	21	31.82	15.7	3.36E-04
			TTGAC	-238	-19	54	81.82	41.5	3.72E-12
			CNGTTR	-297	-238	16	24.24	12.0	1.72E-03
Cluster_23	111	107	ACGT	-131	-19	55	51.4	27.6	6.25E-08
			ACGTG	-208	-25	45	42.06	21.1	3.20E-07
			AMCWAMC	-184	75	49	45.79	29.1	9.05E-05
			KCACGW	-165	-97	17	15.89	6.34	1.35E-04
			ACACNNG	-134	-79	14	13.08	6.8	5.95E-03
Cluster_24	91	87	GAGAGA	-118	185	54	62.07	42.6	8.89E-05
			GATAAG	-114	156	25	28.74	10.7	9.60E-07
			GAGAC	53	102	12	13.79	6.38	3.46E-03
			WGATAR	-233	-198	19	21.84	11.5	1.90E-03
			CAACA	-8	14	12	13.79	6.02	2.10E-03
Cluster_25	81	70	CNGTTR	-168	20	33	47.14	32.3	3.46E-03
			GATT	-178	-169	19	27.14	10.6	2.15E-04
Cluster_26	56		GATT	-88	-68	22	40.74	22.5	8.08E-04
Cluster_27	162	160	CGTGTG	-412	-230	11	6.88	3.29	6.97E-03
			ACGTG	-112	-47	27	16.88	10.6	5.83E-03
			TTGAC	-124	-17	52	32.5	22.9	2.19E-03
			ACACNNG	-119	-44	26	16.25	9.51	2.45E-03
			CANNTG	-99	-78	23	14.38	9.02	9.43E-03
			CATGTG	-325	-286	9	5.62	2.13	2.43E-03
			CAACA	-61	-44	17	10.62	4.53	3.72E-04
			MACGYGB	-165	-46	26	16.25	10.3	7.42E-03
			AMCWAMC	-253	-206	19	11.88	6.99	8.43E-03
Cluster_28	146	142	VCGCGB	-169	28	27	19.01	6.53	1.42E-07
			TGTCTC	-15	61	15	10.56	5.21	3.11E-03
			ACACNNG	-70	-7	18	12.68	7.49	9.96E-03
			CNGTTR	-86	-69	13	9.15	4.85	9.43E-03
			WAACCA	-107	-82	15	10.56	5.58	5.97E-03
Cluster_29	153	149	GAGAGAGA	-132	164	42	28.19	17.8	6.33E-04
			VCGCGB	-137	77	31	20.81	6.53	2.14E-09
			GAGAC	-112	130	58	38.93	30.2	9.28E-03
			TTGAC	-49	-32	15	10.07	3.78	1.81 E-04
			CNGTTR	-289	-272	13	8.72	3.71	1.39E-03

b. *Analyse des clusters avec a priori biologique*

Afin de répondre aux critères et aux questions soulevées par les biologistes dans le cadre de ce projet, plusieurs discussions ont eu lieu permettant de classer les clusters en trois types de profils distincts : (i) les clusters avec des gènes modulés par FLG22 dans Col-0 (clusters 2, 5, 10, 11, 15, 18, 20, 22, 28 et 29) correspondant aux clusters de gènes de la réponse à la FLG22, (ii) les clusters avec des gènes régulés uniquement par *mpk4* dans les conditions normales (clusters 4, 6, 7, 12 et 27) et (iii) les clusters avec des gènes présentant une expression différentielle pour les mutants *mpk3, mpk4*

ou *mpk6* avec ou sans traitement par FLG22 mais pas dans Col-0 (clusters 8, 13, 14, 16, 23, 24). Les deux derniers types de profils peuvent paraître surprenants dans un projet s'intéressant à la réponse à la FLG22, cependant les gènes de ces clusters répondent à la FLG22 dans Col-0 mais de manière tardive, 1h à 3h après traitement soit après les analyses effectuées dans ce projet (Denoux *et al.* 2008). La deuxième classe de clusters correspond ainsi aux groupes de gènes spécifiquement contrôlés par MPK4 et impliqués dans la réponse tardive à la FLG22. Plusieurs de ces clusters contiennent les gènes dont l'expression est modulée de manière combinée par plusieurs kinases en réponse à la FLG22. L'analyse des gènes différentiellement exprimés avait permis d'identifier un chevauchement important des gènes impactés par *mpk3* et *mpk4.* Ces clusters contiennent des gènes dont l'expression est maintenue inchangée lors des étapes précoces de la réponse à un traitement par FLG22 due à l'action concertée de ces kinases. Quatre groupes représentatifs des profils intéressants sont représentés dans la figure 9.

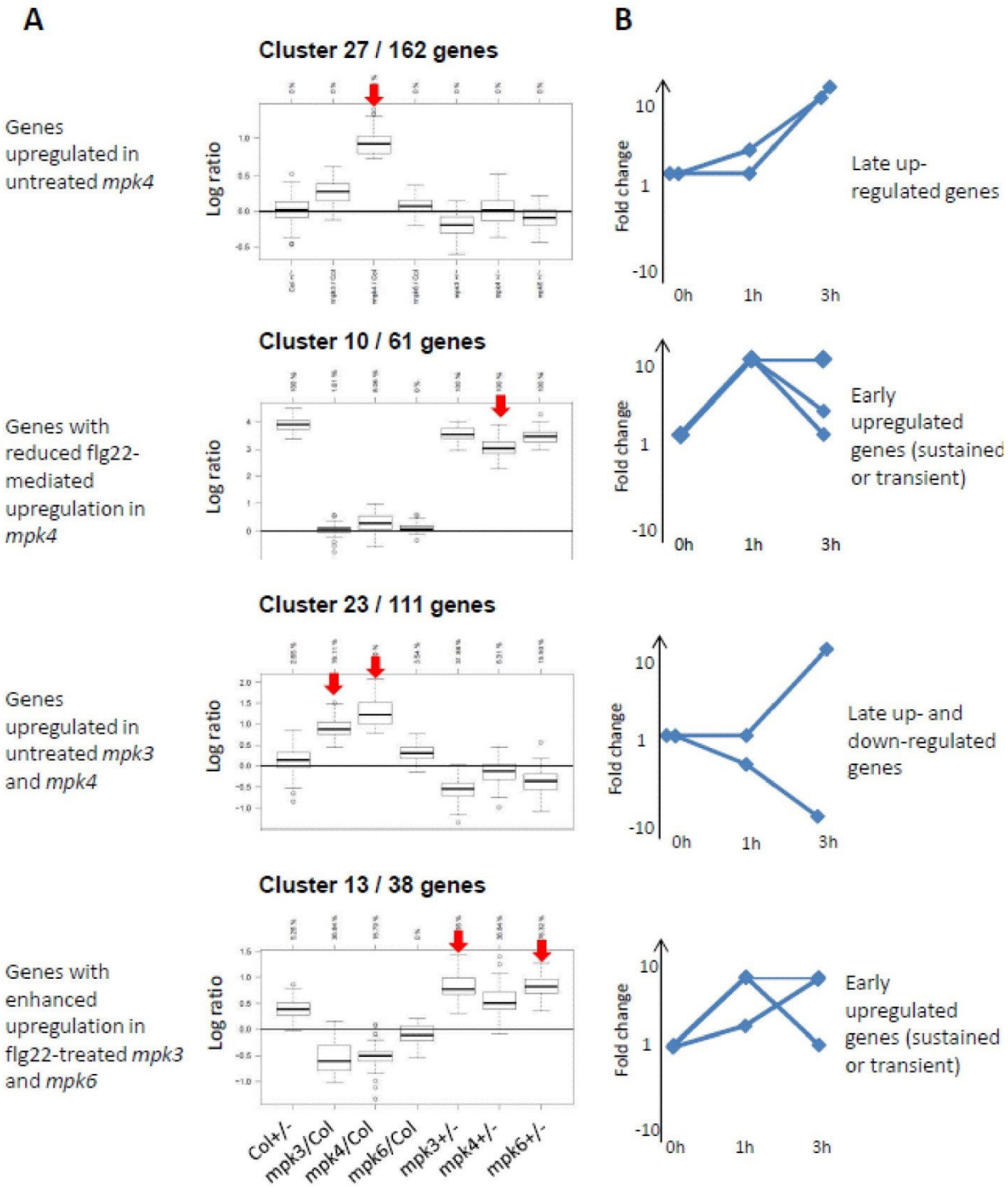


Figure 9 : Figure extraite de Frei Dit Frey *et al*. (2014) représentant une sélection de 4 groupes représentatifs des profils intéressants.

(A) Profils d'expression des clusters. Les commentaires à gauche et les flèches rouges indiquent la dynamique du cluster. (b) Comportement cinétique des clusters comme prédit par la comparaison avec les gènes de réponse « tardive » à la flg22.

Afin de mieux comprendre et analyser la dynamique de réponse de ces différentes classes de profils notamment en termes de liens entre les clusters d'une même classe, tels que des régulateurs communs, j'ai rapproché les clusters en exploitant la similarité des motifs cis-régulateurs. J'ai calculé ainsi une similarité de Jaccard entre les clusters deux à deux en se basant sur la similarité des motifs détectés comme étant sur-représentés dans les différents clusters. La distance calculée m'a permis d'effectuer une classification hiérarchique des clusters (figure 10). Le dendrogramme obtenu montre que nous pouvons classer ces clusters de coexpression essentiellement en deux groupes. Le premier groupe est composé des clusters 11, 26, 22, 29, 18, 15 et 25. Ces clusters appartiennent tous à la première classe de profils à savoir les clusters de gènes impactés par la FLG22 dans Col-0 après 30 minutes d'infection autrement dit les gènes de réponse rapide à la FLG22. Ce groupe de clusters est enrichi en éléments cis-régulateurs correspondants aux motifs W- box, EIN3 et MYB (figure 11). Le motif W-box est lié par les facteurs de transcription WRKY connus pour être des substrats des MAPK. Nous retrouvons également les sites de liaison du facteur de transcription ET (EIN3) qui sont sur-représentés dans cinq clusters contenant des gènes soit surexprimés soit sous-exprimés par le traitement à la FLG22. Ce facteur de transcription est régulé par phosphorylation directement par MPK3 et MPK6 et joue un rôle important dans le contrôle transcriptionnel des composantes de signalisation immunitaire telles que SID2 et FLS2 (Boutrot *et al*. 2010 ; Chen *et al*. 2009 ; Yoo *et al*. 2008). Cette analyse suggère le rôle de MPK 3 et 6 dans les réponses précoces à la FLG22. Cependant l'absence d'effet de la mutation de MPK3 et 6 sur l'expression de ces gènes suggère que ces deux MPK agissent de manière redondante. Le deuxième groupe est composé par le reste des clusters des deux autres classes de profils et sont enrichis en nombreux motifs tels qu'ABRE, RAV1 et ABA. Parmi les clusters présentant des gènes régulés par les mutants mpk3 et mpk4, j'ai trouvé des clusters enrichis par un motif RAV1 généralement lié par un facteur de transcription AP2 / EREBP. Ces clusters regroupent des gènes qui sont surexprimés chez le mutant mpk3 non traité (cluster 14) ou mpk4 (Clusters 4 et 27) ou des gènes qui présentent un profil d'expression impacté chez le mutant mpk3 traité par la FLG22 (Cluster 24). Ces résultats suggèrent que l'activité de ce facteur de transcription est contrôlée par MPK3 et MPK4.

Dans l'ensemble, l'approche de clustering a permis de prédire des groupes biologiquement pertinents de gènes modulés par le stress. Les analyses bioinformatiques que j'ai effectuées d'enrichissements GO et des motifs cis-régulateurs ont permis de mieux comprendre la dynamique de réponse de ces clusters de gènes et d'approfondir les connaissances biologiques autour des voies de régulations

induites par les MAPK en réponse au stress et d'identifier des facteurs de transcription spécifiques dans la régulation de ces groupes.

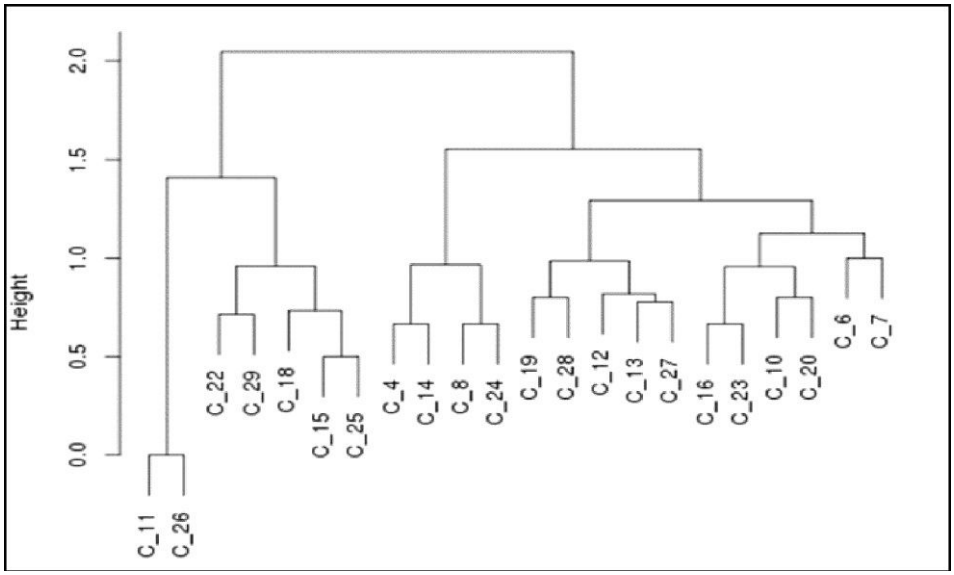


Figure 10 : Dendrogramme du clustering hiérarchique des clusters de coexpression sur la base des motifs sur-représentés qu'ils contiennent, utilisant « la distance de Jaccard » et « ward linkage ».

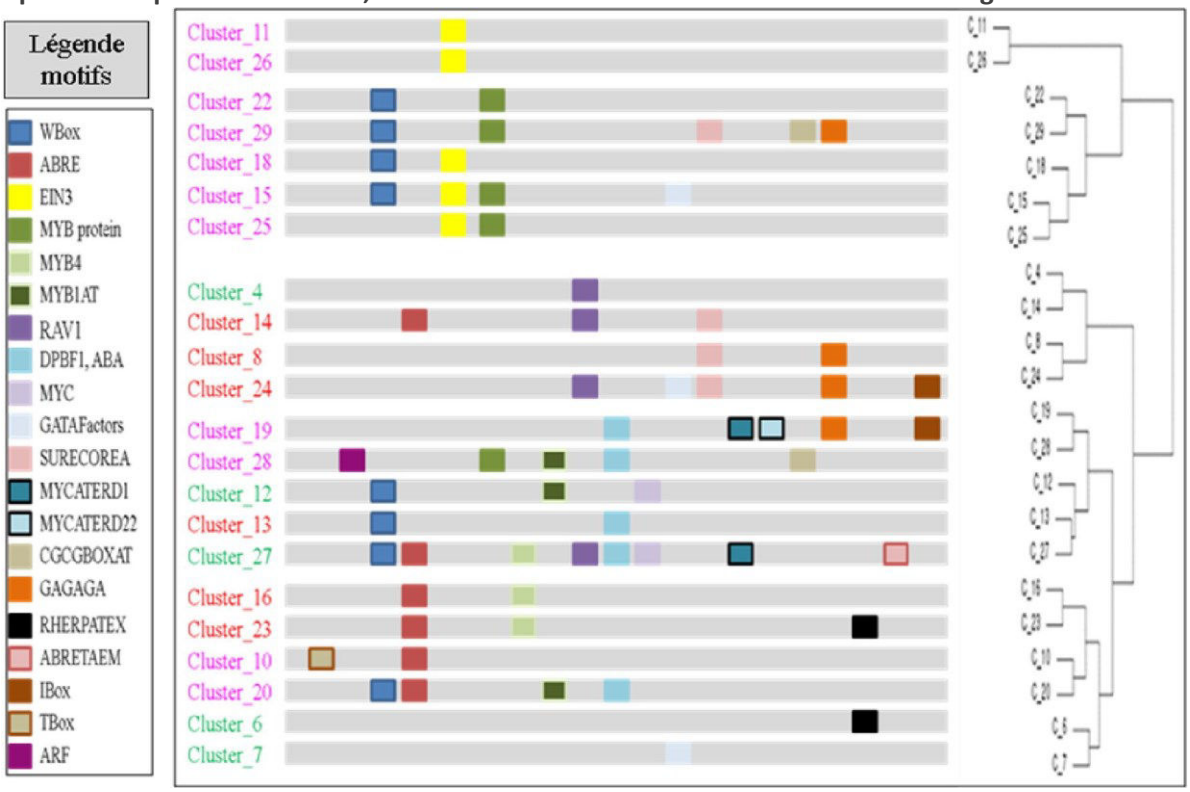


Figure 11 : Représentation schématique de l'enrichissement des clusters en éléments cis-régulateurs.

L'ordre des clusters est déterminé selon la classification hiérarchique des clusters fondée sur la similarité des motifs cis-régulateurs enrichis au sein de chaque cluster. Les clusters écrits en rose appartiennent à la classe de profils de clusters 1 (clusters de gènes de réponse rapide à la flg22). Les clusters en vert appartiennent à la 2ème classe (impactés uniquement par mpk4 et impactés dans la réponse tardive) et les clusters en rouge appartiennent à la 3eme classe (impactés par différents mutants).

L'identification des gènes dans un système est la première étape d'une meilleure compréhension de ce système. L'identification des liens entre ces gènes est l'étape suivante pour mieux l'appréhender en termes d'association fonctionnelle. Ainsi l'objectif de l'étape suivante pour l'analyse de ces données est l'étude du réseau des interactions pouvant exister entre les gènes impactés par les MAPK en réponse au stress par FLG22. Pour cela, nous visons à exploiter simultanément toutes les données d'interactions moléculaires physiques et fonctionnelles disponibles, comme autant d'angles de prise de vue complémentaires décrivant les systèmes biologiques étudiés. Dans notre cas nous utiliserons les données PPI et TF-cibles. La difficulté réside néanmoins dans deux points essentiels. Premièrement, malgré le potentiel de cette approche d'intégration pour la découverte de nouvelles connaissances, l'interprétation biologique pourra être compliquée et délicate notamment en raison de la différence de qualité des données intégrées. La deuxième difficulté correspond à la façon dont elles peuvent être intégrées les unes aux autres, d'autant plus que lors des analyses d'enrichissement j'ai remarqué qu'il y a peu de chevauchement entre les données de transcriptome, interactome et TF-cibles au sein des clusters étudiés, autrement dit que les clusters sont très peu enrichis en interactions PPI ou interactions TF-cibles. Une étude globale de De Bodt et al. (2009) de toutes les PPI identifiées expérimentalement chez A. thaliana montre une faible concordance entre les données de coexpression et les interactions PPI. Une autre étude de De Bodt et al. (2010) montre que le coefficient de corrélation d'expression moyen pour les PPI identifiées expérimentalement est compris entre 0,13 et 0,21. Ils expliquent cette faible concordance de la même manière que ce qui a été mentionné précédemment dans la littérature, à savoir la nature transitoire des PPI. Concernant les interactions TF-cibles, elles sont susceptibles de nous éclairer sur la régulation des groupes de gènes identifiés et sur la réponse en aval de ces gènes. Cependant, les facteurs de transcription qui régulent les gènes d'un cluster de coexpression peuvent ne pas faire partie du cluster lui-même parce que le facteur de transcription peut ne pas être régulé au niveau transcriptionnel, mais au niveau post-transcriptionnel et post-traductionnel. Il pourrait donc avoir un profil d'expression différent de celui des gènes du cluster qu'il régule. Pour ces raisons, l'intégration des données d'interactions à ce stade ne sera pas au niveau de chaque cluster identifié (comme ce qui est fait pour les enrichissements), mais en considérant l'ensemble des gènes modulés par FLG22 à la fois.

Nous espérons ainsi avoir une vue globale sur les réseaux de gènes mis en place en réponse au stress tout en ayant l'information sur la dynamique de cette réponse pour chaque gène.

Dans cette optique, j'ai intégré les données PPI ainsi que les données TF-cibles pour construire un seul réseau hétérogène visualisé sous Cytoscape. Le réseau a été construit en considérant chaque

interaction (PPI ou TF-cible) impliquant au moins un gène appartenant à un cluster. Pour visualiser les données de coexpression avec le réseau obtenu, une solution était de représenter les gènes d'un même cluster avec la même couleur. Par exemple comme le montre la figure 12, tous les gènes du réseau appartenant au cluster 10 sont colorés en rouge. Cela m'a permis d'observer le regroupement de quelques gènes d'un même cluster, et la majorité des autres gènes appartenant à ce cluster sont dispersés dans le réseau. Cela souligne à nouveau la faible concordance entre les données omiques ainsi que la difficulté de l'interprétation biologique de la confrontation de ces données.

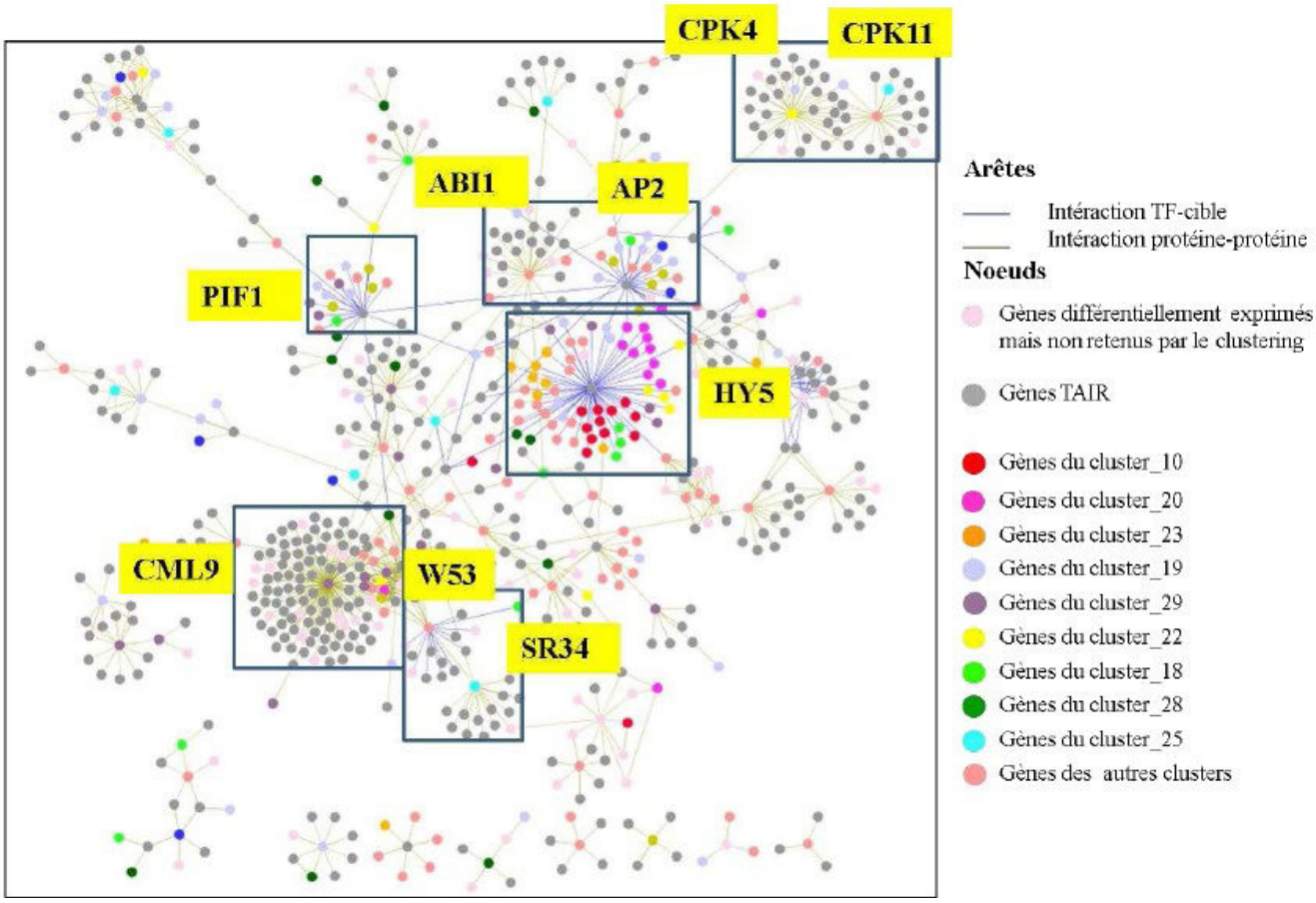


Figure 12 : Réseau d'interactions TF-cible et PPI des gènes différentiellement exprimés dans les conditions du stress flagelline.

Représentation des données TF-cible en bleu et PPI en vert kaki. Les gènes d'un même cluster sont représentés par une même couleur selon la légende à droite. Les cadres entourent les hubs analysés au sein du réseau.

Cependant cette visualisation a mis l'accent sur la présence de plusieurs hubs. Sachant que le degré des nœuds du réseau varie entre 1 et 115, avec une moyenne égale à 1 et un troisième quantile égal à 2, la majorité des gènes sont peu connectés et les dix gènes les plus connectés ont un degré supérieur à 19. J'ai considéré ces 10 gènes comme des hubs : certains sont connus pour être des régulateurs impliqués dans la réponse au stress tels que CML9 ou CPK11, d'autres correspondent à des facteurs

de transcription qui n'ont pas encore été associés à ce processus (figure 12). Ces hubs représentent ainsi des candidats potentiels pour être des régulateurs clés de la réponse au stress et ont suscité l'intérêt des collaborateurs biologistes. J'ai effectué l'analyse des 3 hubs régulant de nombreux gènes des clusters de coexpression : HY5(AT5G11260), AP2(AT4G36920) et PIF1(AT2G20180). Les degrés de ces hubs sont respectivement de 65, 27 et 22. Pour connaître les caractéristiques des gènes régulés par ces 3 hubs, j'ai intégré leurs annotations GO Slim. Plusieurs de ces gènes sont annotés avec le terme « réponse au stress » ce qui indique l'implication de ces facteurs de transcription dans la réponse au stress. J'ai également cherché à savoir si le nombre de gènes appartenant aux clusters et régulés par ces facteurs de transcription est significatif. Pour cela j'ai mis en place un test statistique afin de déterminer si le nombre de gènes régulés par un hub dans un cluster est significatif par rapport au nombre d'interactions du hub dans le génome. Les résultats de ce test sont représentés dans le tableau 5 et montrent par exemple que le facteur de transcription HY5 régule un nombre de gènes significatif dans 8 clusters suggérant ainsi l'importance de ce facteur pour la régulation des gènes impactés par FLG22 et sa relation dans la cascade de régulations induits par les MAPK.

Tableau 5 : Résultat du test statistique mis au point pour déterminer la significativité du nombre de gènes régulés par un hub dans un cluster par rapport au nombre d'interactions du hub dans le génome.

Seuls les clusters pour lesquels cette analyse est significative sont montrés pour les 3 hubs testés. La 1ère colonne correspond à l'identifiant du facteur de transcription (hub) analysé, la 2ème à son nom et la 3ème colonne au nom de la famille à laquelle il appartient. La 4ème colonne correspond au nombre total de gènes dans le génome d'A. thaliana contrôlés par ce TF. Pour chaque cluster analysé (5ème colonne) la taille du cluster et le nombre de gènes de ce cluster contrôlés par le TF sont indiqués. La dernière colonne indique la p-value du test statistique (binomiale) effectué ou le nombre de succès pour le test binomial correspond au nombre d'interactions avec le TF dans le cluster, et la probabilité de succès correspond au nombre d'interactions avec le TF dans le génome divise par le nombre de gènes du génome.

Hub	TF	Famille TF	Nbr ref	Cluster	Taille cluster	Nbr cluster	p-value
AT5G11260	HY5	bZIP	221	cluster_10	61	11	2.4E-13
				cluster_20	92	9	1.9E-08
				cluster_23	111	7	2.4E-05
				cluster_19	109	5	2.0E-03
				cluster_29	153	5	1.6E-02
				cluster_22	66	4	2.0E-03
				cluster_2	14	4	6.7E-07
				cluster_18	74	3	4.2E-02
AT4G36920	AP2	AP2EREBP	165	cluster_19	109	10	2.8E-10
				cluster_24	91	4	2.0E-03
AT2G20180	PIF1	bHLH	189	cluster_19	109	5	1.0E-03

c. *Conclusion de cette analyse*

D'un point de vue biologique, cette collaboration a permis de donner une vue d'ensemble des changements transcriptionnels induits par la FLG22 chez les mutants des MAPK 3, 4, et 6 chez *A. thaliana* et de mettre en évidence des groupes de gènes ayant la même dynamique de réponse dans ces conditions. Les analyses d'enrichissement ont montré que ces groupes sont biologiquement pertinents et cohérents et présentent des biais significatifs en termes d'annotation GO, de localisation subcellulaire, d'hormones ou d'enrichissement en gènes de la liste BiblioStress. Les analyses ont permis également de donner des informations autour des motifs cis-régulateurs présents dans les promoteurs des gènes de ces clusters et de prédire de nouveaux facteurs de transcription impliqués dans la régulation de gènes dépendants de la cascade MAPK et impactés par la FLG22.

Sur le plan technique, cette étude a été l'occasion d'explorer et de mettre en place des méthodes bioinformatiques et statistiques nécessaires à l'analyse et l'interprétation biologique et fonctionnelle des groupes de gènes coexprimés. Les résultats biologiques de ces analyses confortent quant à l'utilisation de ces clusters de coexpression pour l'annotation fonctionnelle. Cette analyse a été dirigée par des questions spécifiques et des *a priori* biologiques ce qui m'a amené jusqu'à l'analyse détaillée de certains clusters et de certains gènes afin d'extraire de nouvelles connaissances biologiques telles que la relation dans un cluster entre les motifs cis-régulateurs enrichis, le profil d'expression et la cinétique de la réponse des gènes au stress.

Toutefois, dans le cadre de ma thèse, je m'intéresse à l'étude de 18 catégories de stress biotiques et abiotiques impactant environ 18 000 gènes classés dans 681 clusters de coexpression. Les méthodes développées pour l'analyse et la caractérisation des 29 clusters modulés par FLG22 devaient être adaptées et automatisées. Pour réaliser ce changement d'échelle, j'ai mis en place un système d'informations décrit dans la section suivante.

IV. GEM2Net : nouveau module de CATdb

La caractérisation biologique et l'analyse fonctionnelle des clusters de coexpression des gènes impactés par une seule catégorie de stress ont montré la validité et la pertinence biologique de ces clusters. L'objectif est d'effectuer ce type d'analyse de manière automatique pour l'ensemble des clusters de coexpression issus de l'analyse des 18 catégories de stress. La finalité de cette analyse est de fournir, pour un ensemble de groupes de gènes ayant la même dynamique de réponse sous une

condition particulière, les éléments nécessaires pour la caractérisation des membres de ces groupes. L'association de gènes orphelins ou mal caractérisés avec des gènes connus au sein de tels groupes est une information considérable autour de leur implication dans la réponse au stress et une première étape pour la compréhension de leurs fonctions potentielles. De plus, cette ressource pourra être le point de départ de nouvelles collaborations avec des scientifiques intéressés par l'étude d'un stress particulier. J'ai donc adapté les méthodes et les outils d'annotation fonctionnelle développés pour l'analyse du stress par FLG22 à l'analyse de l'ensemble des clusters de manière automatique et globale. Au vu du nombre de clusters à interpréter, j'ai également travaillé sur la représentation des résultats de cette annotation fonctionnelle pour les présenter sous une forme globale et compréhensible tout en permettant de faciliter leur interprétation ou de répondre à d'autres questions plus spécifiques.

Dans cet objectif et afin de pouvoir exploiter, organiser et mettre à disposition les résultats de cette analyse à large échelle, la base de données nommée GEM2Net pour « from Gene Expression Modeling to Networks in plants » a été implémentée en tant que nouveau module de la base de données CATdb. Le module GEM2Net permet de fournir un aperçu global de la réponse des plantes aux changements environnementaux ou aux attaques biologiques.

La base a été dotée d'une interface graphique qui permet de résumer graphiquement les biais fonctionnels et les annotations autour des clusters de manière simple et rapide.

Je vais décrire dans cette section la gestion des données ainsi que le système de visualisation mis en place. Je donnerai également une description globale des résultats de la caractérisation fonctionnelle des clusters à l'échelle du projet entier.

1. Gestion des données GEM2Net

a. *Automatisation de l'analyse bioinformatique*

Avec près de 700 clusters obtenus, une étape d'automatisation des analyses bioinformatiques est indispensable afin de caractériser l'ensemble de ces unités de coexpression. J'ai grandement participé à la mise en place des pipelines pour l'automatisation complète des étapes d'analyses et l'organisation des résultats. Pour cela, nous avons mis en place d'abord une certaine arborescence ou architecture de répertoires pour faciliter l'organisation de l'ensemble des résultats produits à chaque étape de l'analyse. Les méthodes implémentées pour l'analyse des clusters du stress par FLG22 ont

été appliquées par type de stress considérés comme des projets indépendants et ont été développées avec le souci de respecter cette arborescence notamment pour l'emplacement des résultats générés.

Les programmes sont implémentés en langage Perl dans un environnement Eclipse sous Unix. L'usage d'un repository sous Subversion (SVN) et le respect de normes de programmation telle que le commentaire de code optimise le partage et la réutilisation des programmes par les autres membres de l'équipe. L'optimisation du temps de calcul a été une priorité. Ces programmes d'analyses sont génériques et totalement indépendants. Ils peuvent être utilisés pour l'analyse de données autres que celles de GEM2Net ou même autres que l'analyse de coexpression comme par exemple des listes de gènes d'intérêt regroupés sur la base de leur co-citation dans la littérature ou autres critères. Les principales étapes d'analyses de GEM2Net sont représentées schématiquement dans la figure 13.

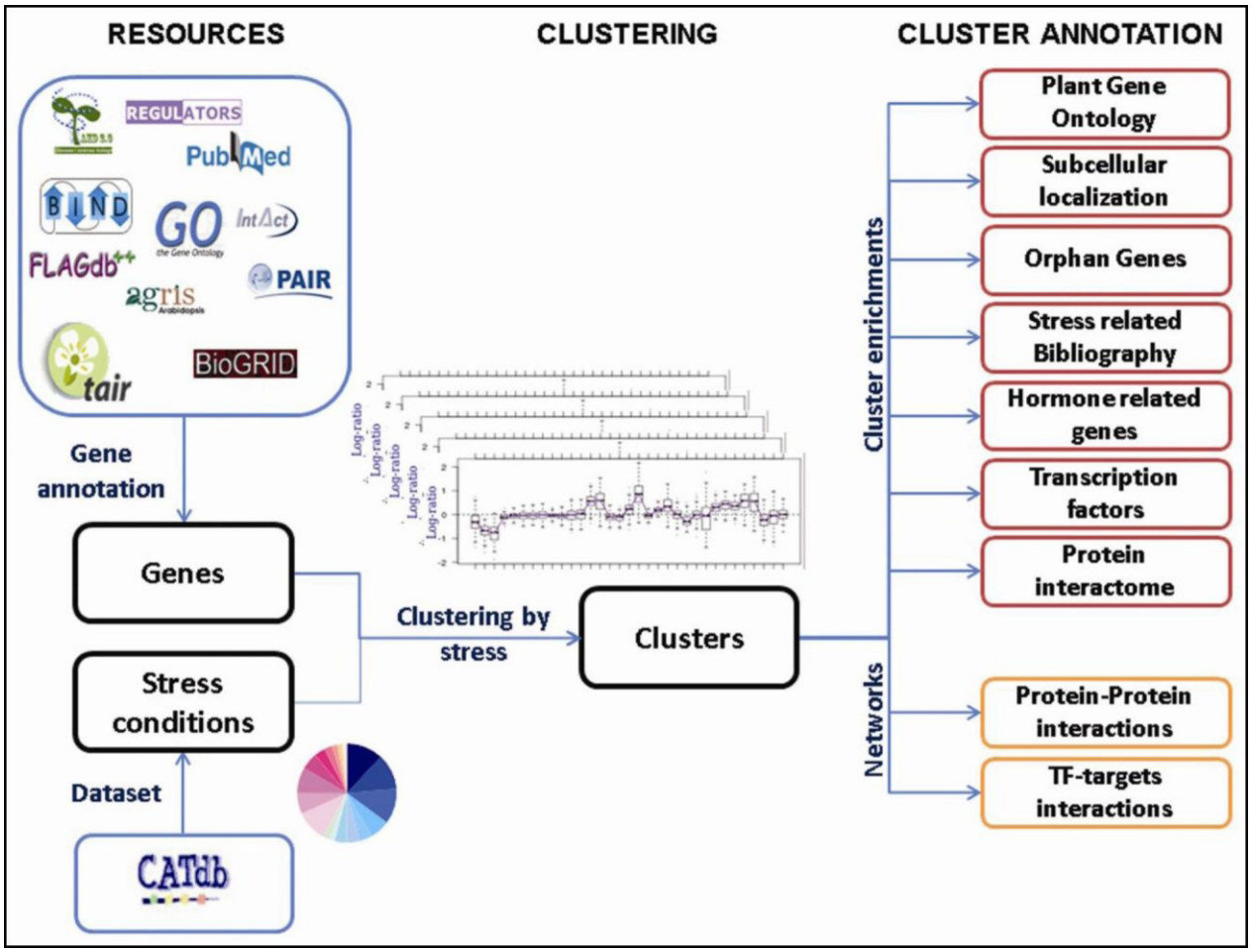


Figure 13 : Organigramme général de GEM2Net (Zaag et al. 2014).

A gauche sont représentées les ressources utilisées pour la caractérisation des gènes et des conditions de stress considérées lors des analyses de clustering. A droite les analyses bioinformatiques effectuées afin de caractériser et de mesurer les biais fonctionnels dans chaque cluster de coexpression.

b. *Conception et implémentation de la base de données*

J'ai travaillé sur la conception et l'élaboration d'une base de données relationnelle sous PostGresSQL. Le modèle permet de représenter les objets manipulés dans le projet (stress, gènes, clusters, enrichissements etc.) et les relations existantes entre ces objets. Pour la conception de la base, les besoins se sont spécifiés et affinés au fur et à mesure de l'avancement de l'analyse du projet FLAGELLINE décrit dans la section précédente. Ce projet a servi à l'enrichissement du schéma conceptuel en fonction des analyses effectuées ainsi que pour les tests de l'interrogation de la base et de l'accès aux données.

Techniquement le module GEM2Net correspond à une base de données indépendante. Elle est composée de 20 tables qui peuvent être découpées en 3 composantes principales. La première permet de stocker les gènes (version TAIR 10) et leurs annotations telles que les termes GO, s'il s'agit d'un gène orphelin, s'il appartient à la liste BiblioStress et les références correspondantes, la relation avec une famille de gènes à laquelle il appartient telle que les facteurs de transcription ou les hormones avec leurs descriptions. Les interactions de type protéine-protéine (PPI) ou facteur de transcription-cibles (TF-cibles) reliant les gènes sont représentées grâce à deux tables reliées à la table des gènes résumant le type de l'interaction (exemple AI1, LCI etc.), sa nature (expérimentale ou prédite), les références bibliographiques associées à ces interactions.

La deuxième composante permet d'organiser le résultat de l'analyse de coexpression à savoir l'organisation des gènes en clusters par type de stress et en fonction du type de classification considéré (MAP ou MFDR). Dans cette composante nous sauvegardons aussi la description des conditions ou expériences transcriptomiques impactées dans chaque clustering.

Enfin la troisième composante concerne les tables résumant les enrichissements trouvés pour toutes les caractéristiques mesurées dans les clusters et les informations associées telles que la probabilité critique ou les références. Le schéma relationnel de la base est détaillé dans la figure 14.

L'insertion des données est assurée par un seul pipeline composé d'un ensemble de programmes qui permettent de récupérer les informations et de les mettre en forme afin de les intégrer dans la base de données en respectant le modèle conceptuel et les règles d'intégrité de la base. J'ai pris en charge le développement de ce pipeline et son application pour insérer les données du stress FLG22 dans la base. Pour le reste des catégories de stress, les membres de l'équipe se sont chargés de l'application du pipeline pour l'insertion des autres données. La mise à jour des données de référence (gènes TAIR, données interactome, motifs cis-régulateurs etc.) sont également pris en charge par l'équipe.

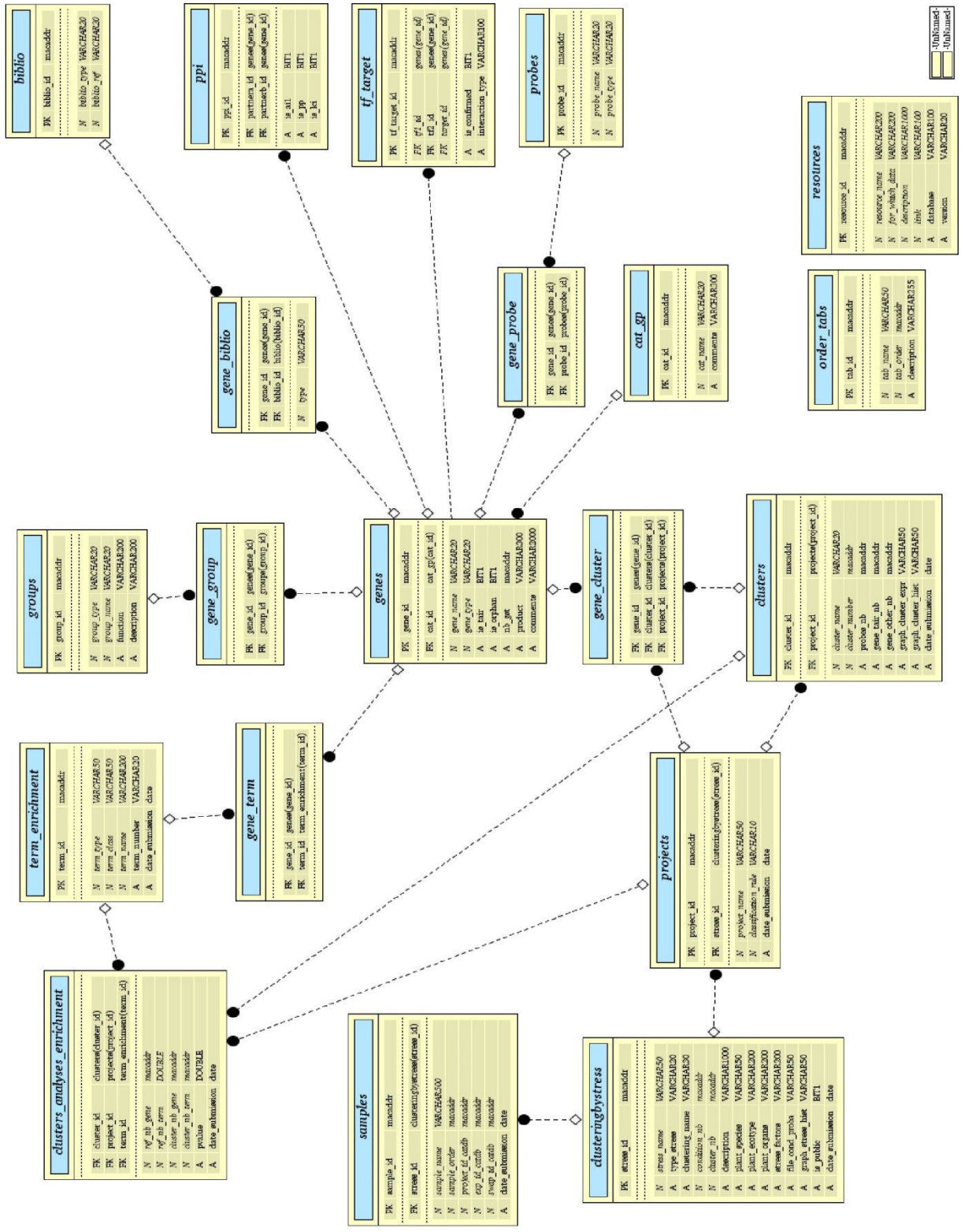


Figure 14 : Schéma relationnel de la base de données GEM2Net.

2. Interface graphique

a. *Implémentation de l'interface graphique*

L'objectif était de mettre en place une interface graphique qui permet de faciliter l'accès aux données stockées dans la base de données GEM2Net et de les représenter de manière simple, compréhensible et globale. Mon expertise lors de l'analyse du projet FLG22 a servi de guide pour définir les besoins et la manière de représenter les clusters de coexpression et les annotations associées à ces groupes. J'ai contribué ainsi à la conception de l'interface et la mise en place des fonctionnalités nécessaires pour une représentation globale des groupes de gènes impliqués dans la réponse aux stress. Cette représentation doit permettre à terme de donner les moyens de pousser plus loin les analyses et d'aider à répondre à des questions spécifiques.

En pratique c'est au niveau de l'interface graphique que la base de données GEM2Net est reliée à la base de données CATdb. L'interface est accessible directement via le lien suivant : http://urgv.evry.inra.fr/GEM2NET ou en passant par l'interface de CATdb à travers le lien http://urgv.evry.inra.fr/CATdb.

Le développement de l'interface a été pris en charge par Zakia Tariq (ingénieur d'étude en CDD au sein de l'équipe) et Jean-Philippe Tamby (ingénieur d'étude titulaire au sein de l'équipe) en utilisant le langage PHP (version 5.3.3) en programmation orientée objet. La visualisation dynamique et la facilité de navigation ont été obtenues grâce à Javascript. L'outil Cytoscape Web (version 1.0.4) a été intégré à l'interface afin de permettre la visualisation des réseaux obtenus à partir des données PPI ou des données TF-cibles.

b. *Système de visualisation*

Le système de visualisation que nous avons mis en place vise à représenter de manière dynamique et parlante les clusters de coexpression et les annotations associées. Ceci a été fait par type de stress. Les résultats des analyses d'enrichissement des ontologies GO, de localisation subcellulaire, d'hormones, de facteurs de transcription, gènes orphelins, gènes de la BiblioStress et enrichissement en interactions PPI, sont tous représentés de la même manière, chacun dans un onglet différent. Dans chaque onglet, chaque cluster de coxpression est représenté par un camembert dont le diamètre est proportionnel à sa taille. Lorsqu'un terme de la caractéristique visualisée, par exemple l'annotation BP, est sur-représenté dans un cluster, une section de taille proportionnelle au nombre de gènes annotés avec ce terme, apparaît dans le cluster avec une couleur associée au terme. Au survol d'un cluster, un tableau apparaît à gauche pour détailler tous les termes reliés à ce cluster ainsi que le

nombre de gènes annotés avec ce terme. Pour les termes sur-représentés la probabilité critique du test hypergéométrique effectué pour tester l'enrichissement est également indiquée (figure 15).

En cliquant sur un cluster, un résumé de toutes les analyses des métadonnées concernant ce cluster, est affiché. La figure 16 montre un exemple pour le cluster 49 du stress Virus, où l'ensemble des biais fonctionnels détectés pour ce cluster sont affichés. Un onglet appelé « Networks » permet la visualisation des interactions impliquant les gènes d'un cluster choisi. Le type d'interaction (PPI, TF-cible) et sa nature (prédite, expérimentale) sont également des paramètres que l'utilisateur peut modifier en fonction de ses besoins. La figure 17 reprend l'exemple du cluster 49 et permet de visualiser les interactions PPI dans lesquelles sont impliqués les gènes du cluster.

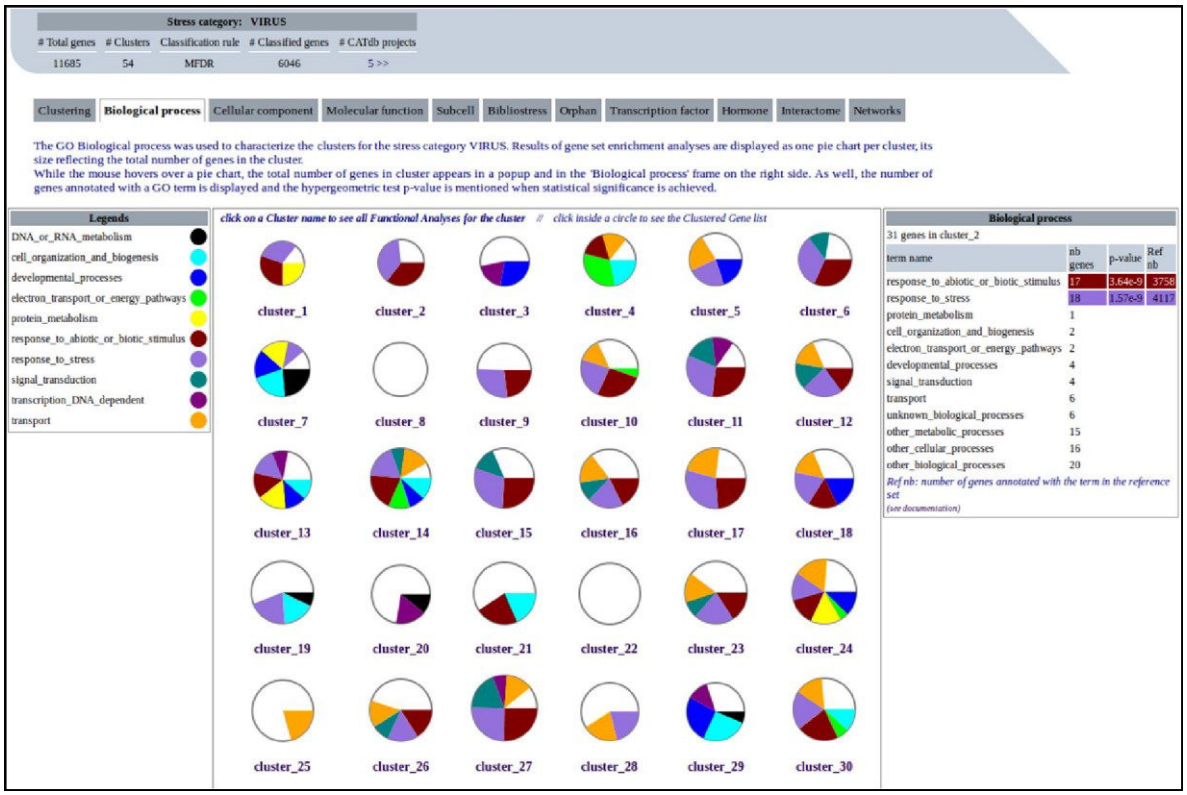


Figure 15 : Capture d'écran de l'interface graphique GEM2Net correspondant à une analyse d'enrichissement des clusters en termes GO de l'ontologie BP pour la catégorie de stress Virus.

Les 30 clusters issus de la classification des gènes impactés par le stress virus sont tous représentés dans la même fenêtre. Chaque camembert correspond à un cluster. Les enrichissements sont représentés par des sections colorées dans le camembert.

Figure 16 : Vue globale de l'ensemble des méta-analyses pour le cluster 49 de la catégorie de stress Virus.
Les camemberts au centre correspondent à chaque fois au cluster 49 et ce pour chaque catégorie de meta-analyse. Le tableau sur le côté supérieur droit permet de donner plus de détails autour des enrichissements détectés. Le tableau dans la partie centrale inferieure représente la liste des gènes impliqués dans les biais mesurés pour une catégorie de méta-analyse choisie (dans cet exemple il correspond à l'ontologie BP). Dans ce tableau, chaque gène est étiqueté avec un point de couleur correspondant a celle du terme auquel il est associé (voir tableau des légendes sur la gauche). Les points bleus correspondent aux enrichissements pour les autres métadonnées.

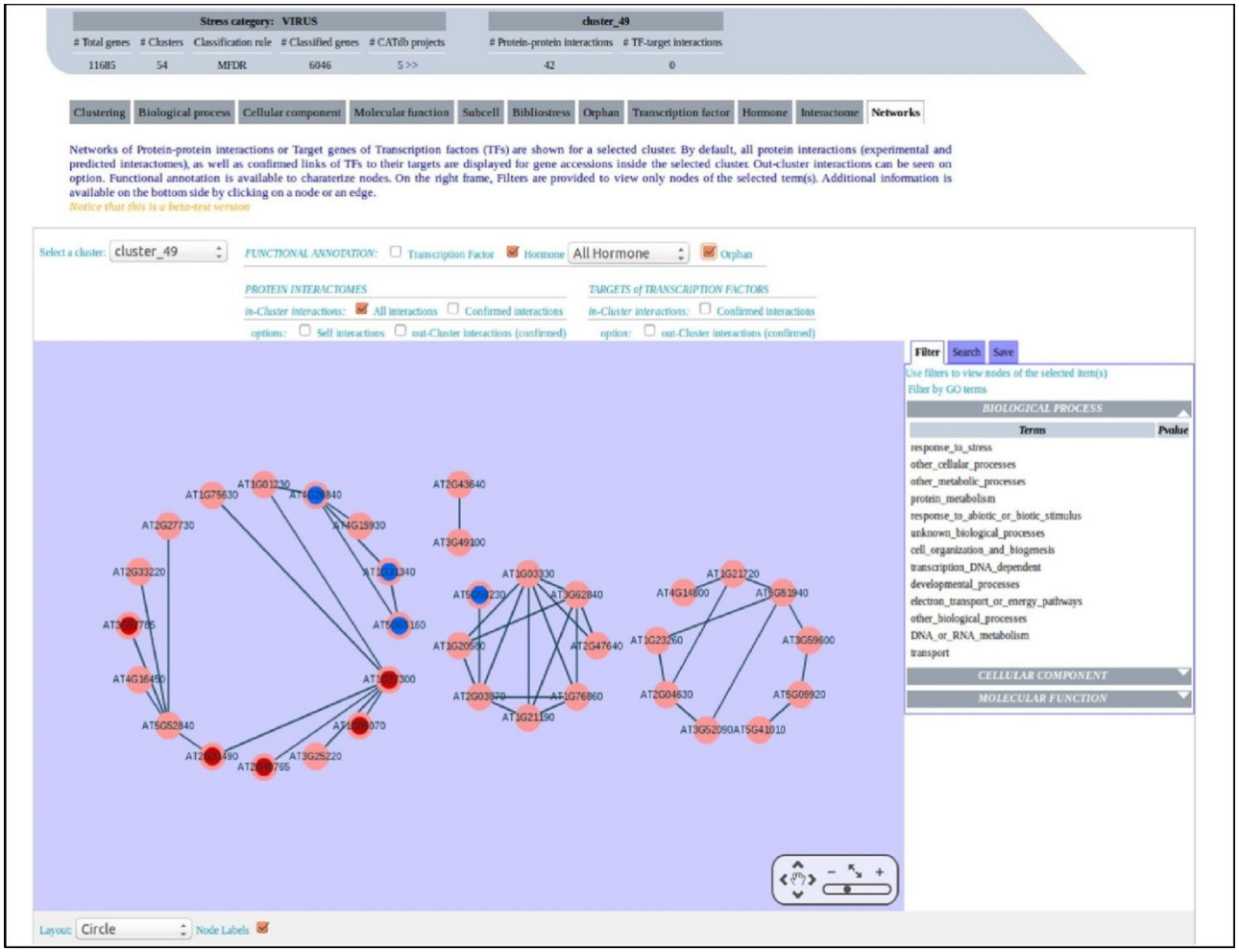


Figure 17 : Visualisation du réseau d'interactions PPI impliquant les gènes du cluster 49 (stress Virus), via l'outil Cytoscape Web.

Dans le panneau central, toutes les interactions PPI (arêtes) entre les gènes du cluster 49 (nœuds) sont représentées par des lignes bleues sombres. L'annotation fonctionnelle des gènes est superposée sur les nœuds du graphe en cochant les cases du panneau supérieur correspondantes aux familles d'hormones (en bleu) et aux orphelins (en rouge). Sur le volet de droite, des filtres de catégories GO peuvent être appliqués sur le réseau pour afficher uniquement les nœuds de l'annotation sélectionnée. De plus, les interactions de type TF-cibles peuvent également être affichées dans le même réseau en cochant la case correspondante dans le panneau supérieur.

c. *Utilisation de l'interface*

La page d'accueil permet de fournir une vue globale du projet, des ressources exploitées et des méthodes appliquées. L'utilisateur peut consulter par la suite les analyses en cliquant sur le lien « Explore DATA & ANALYSES », permettant d'arriver sur une fenêtre avec un tableau résumant pour chaque catégorie de stress, le nombre de clusters obtenus et le nombre de gènes classés selon les deux règles de classification MAP et MFDR. Il faut noter que la règle de classification MFDR est plus stringente et permet une meilleure confiance puisqu'elle ne permet de classer que les gènes ayant une probabilité conditionnelle d'appartenir au cluster plus élevée qu'un certain seuil ce qui

permet de limiter le taux de faux positifs. Avec la règle de classification MAP, tous les gènes différentiellement exprimés sont classés dans un cluster. Nous donnons naturellement plus de confiance aux clusters issus de la classification MFDR qu'à ceux issus de la classification MAP. Les analyses concernant un type de stress sont accessibles en cliquant sur ce nombre de gènes en fonction de la méthode de classification. Un utilisateur intéressé par l'étude d'un type de stress particulier connaitra donc rapidement tous les groupes de gènes impactés par ce stress et partageant des signatures d'expression identiques. Notre ressource lui offre également toute l'analyse d'annotation fonctionnelle de ces clusters et lui permet d'afficher rapidement tous les biais pour lesquels ces clusters sont enrichis.

En plus de cette recherche par type de stress, une autre utilisation de l'interface est possible. L'utilisateur peut être intéressé par un gène ou une liste de gènes identifiés par sa propre étude et souhaite comparer cette liste aux données GEM2Net et rechercher les clusters les plus proches de sa liste de gènes. En soumettant sa liste d'intérêt dans l'interface, l'utilisateur retrouvera tous les clusters dans lesquels se trouvent les gènes de sa liste par type de stress. Ce type de recherche peut donc aider des biologistes à répondre à des questions spécifiques à savoir quels sont les partenaires fonctionnels de ses gènes d'intérêt, comment répondent-ils au stress, quelles sont les caractéristiques biologiques de ce groupe et est ce qu'ils répondent de manière spécifique à un seul stress ou est ce qu'ils sont impliqués dans la réponse à d'autres catégories de stress. Il pourra ainsi accéder aux partenaires ayant la même dynamique de réponse que les gènes dont il dispose et accéder rapidement à l'ensemble des données de caractérisation fonctionnelle de ces groupes. Cette ressource offre aussi tous les moyens nécessaires pour des scientifiques désireux de réaliser une étude de comparaison de stress, en identifiant les groupes de gènes de réponse spécifique et les groupes de gènes de réponse communes aux différentes catégories de stress et d'étudier leurs caractéristiques.

3. Description biologique globale des clusters

Les résultats d'annotation fonctionnelle des clusters étudiés mettent en évidence certaines caractéristiques des données GEM2Net par rapport à tout le génome d'A. thaliana considéré comme la référence pour les tests d'enrichissement. En ce qui concerne l'annotation des gènes liés au stress, un enrichissement attendu est trouvé pour l'ensemble de données GEM2Net avec 23% des gènes annotés avec des termes GO liés au stress (colonne de stress BP) contre 15% dans l'ensemble de référence, et de 12% contre 7,5% pour l'enrichissement en gènes reliés au stress extraits de la

littérature (liste BiblioStress). Une caractéristique originale de l'ensemble des gènes de GEM2Net, qui est visible dans le tableau 6, est l'enrichissement des clusters en facteurs de transcription (9 contre 6,5% pour la référence). Les facteurs de transcription représentent une classe de gènes essentiels pour réguler la transcription d'autres gènes, en particulier dans le cadre de la réponse au stress (Horan *et al*. 2008).

En ce qui concerne les analyses d'enrichissement des clusters en annotation GO, 80% des clusters sont associés aux termes reliés au stress : « réponse au stress » ou « réponse au stress biotique ou abiotique ». Ce biais dans la catégorie de stress est attendu, mais d'autres biais sont également trouvés et permettent de déchiffrer les fonctions des gènes orphelins. Par exemple, 63% des clusters sont enrichis avec le terme GO « transport » et 39% sont enrichis en terme « chloroplaste » pour la localisation subcellulaire des protéines. Globalement 98% des clusters ont un biais fonctionnel dans au moins un terme GO. Ce pourcentage d'enrichissement est supérieur à celui déterminé par Heyndrickx et Vandepoele (2012) où 80% de leurs modules sont enrichis en termes GO.

Tableau 6 : Comparaison du nombre de gènes entre une référence (tous les gènes *d'A. thaliana*) et GEM2Net par métadonnées.

Les métadonnées analysées sont les suivantes: gènes orphelins; termes go reliés au stress («réponse au stress» et «réponse au stress biotique ou abiotique») de l'ontologie BP; liste bibliostress répertoriant les gènes impactés par le stress extraits de la littérature; TF est une liste de gènes caractérisés comme TF d'après le projet regulators; hormone est une liste de gènes ayant un lien avec la réponse aux hormones présents dans la base de données ahd2.0. Les chiffres écrits en caractères gras soulignent des enrichissements significatifs dans gem2net par rapport aux données de référence (test binomial avec pvaleur <0,05).

	Total	Orphelins	« stress » BP	Bibliostress	TF	Hormones
Référence	34 042	5105 (15%)	5106 (15%)	2580 (7.5%)	2260 (6.5%)	695 (2%)
GEM2Net	17 264	2165 (13%)	**4003 (23%)**	**2064 (12%)**	**1578 (9%)**	487 (3%)

Cette homogénéité et cohérence biologique des clusters souligne le potentiel des données transcriptomiques utilisées et de l'approche de clustering pour l'identification de groupes de partenaires fonctionnels et pour l'inférence de fonction. Ces résultats appuient ainsi l'utilisation de ces groupes de gènes pour l'inférence fonctionnelle.

4. Conclusion de ces analyses

Le nouveau module GEM2Net a permis de donner une vue globale des unités de coexpression liées à la réponse au stress chez *A. thaliana*. La caractérisation fonctionnelle de ces unités et les nombreux enrichissements indiquent que cette étude de coexpression à grande échelle génère des clusters biologiquement significatifs. La pertinence biologique de ces clusters présente un cadre rigoureux pour l'étude de fonction des gènes orphelins et mal caractérisés au sein de ces clusters.
La mise à disposition des analyses via un système de visualisation simple et intuitif permettra aux utilisateurs d'exploiter l'ensemble de ces données ou de se focaliser sur l'étude d'un cas particulier grâce aux nombreuses applications développées et pourra être à l'origine de futures collaborations telles que celle qui a été entreprise dans le cadre du projet FLG22.

V. Conclusion et discussion

Les gènes des clusters ont été caractérisés grâce à un ensemble hétérogène mais contrôlé de données, ce qui permet de consolider la pertinence des connaissances biologiques extraites des analyses de coexpression et la confiance qu'on peut leur accorder. J'ai caractérisé les gènes grâce aux ontologies GO, localisations subcellulaires, gènes reliés à la réponse aux stress connus dans la littérature ou encore en familles de gènes telles que les facteurs de transcription et les hormones avec des informations autour de ces familles de gènes et leurs fonctions dans la littérature. Les biais fonctionnels ont été mesurés par la suite grâce à des tests statistiques.
Les méthodes développées afin d'intégrer ces données hétérogènes pour la caractérisation fonctionnelle d'un groupe de gènes ont été appliquées dans le cadre de deux contextes différents. D'une part dans le cadre d'un projet spécifique avec des collaborateurs pour une question biologique précise et d'autre part dans le cadre d'une analyse globale pour caractériser la réponse à 18 catégories de stress différentes. Dans le cadre de l'analyse du projet spécifique, l'exploration des clusters de coexpression a été guidée par des a priori biologiques. J'ai réalisé des analyses au niveau des clusters, des classes de clusters de profils proches et même au niveau des gènes.
Certaines analyses ont été effectuées uniquement pour l'étude de ce projet notamment l'analyse des motifs cis-régulateurs et la construction du réseau d'interactions de gènes. Ceci a permis d'une part de valider la cohérence des unités de coexpressions obtenues, et d'autre part de donner des indices

autour de la fonction et de la régulation de ces gènes. Cette collaboration a donné lieu à la publication d'un article dont je suis deuxième auteur (Frei Dit Frey *et al*. 2014).

Le contexte de l'analyse des unités de coexpressions impactées par 18 catégories de stress était différent. L'objectif était de mettre en place un système de visualisation qui permet d'aider à l'interprétation biologique de ces clusters de manière simple et rapide afin de pouvoir les analyser simultanément. Les priorités et les difficultés de cette analyse à large échelle étaient donc différentes et se concentraient sur l'optimisation et l'automatisation des analyses et également sur la manière de représenter les résultats des biais fonctionnels détectés au sein de ces clusters. Cette ressource a été mise à disposition de la communauté scientifique grâce à un nouveau module GEM2Net dans la base de données CATdb et a été valorisée par un article dans l'édition spéciale de NAR database 2015 (Zaag *et al*. 2015. Annexe C). Elle donne un aperçu global de la réponse des plantes aux changements environnementaux ou aux attaques biologiques et donne les moyens nécessaires aux biologistes intéressés par une catégorie de stress particulière d'approfondir ces analyses avec leurs connaissances. Cette ressource pourra être ainsi à l'origine de futures collaborations pour l'étude spécifique d'un type de stress ou d'une question biologique précise.

Lors de l'étude du stress FLG22 l'intégration des interactions moléculaires a permis la construction d'un réseau et l'identification de certains hubs correspondant à des régulateurs potentiels qui contrôlent un nombre significatifs de gènes. Ces hubs correspondent à des facteurs de transcription qui ne sont pas régulés transcriptionnellement et de ce fait ne sont donc pas identifiés par l'analyse de coexpression. L'intégration des données PPI et TF-cibles a permis ainsi d'identifier ces régulateurs et de faire le lien avec les clusters de coexpression. Cependant la construction de ce réseau a mis en évidence la difficulté de son interprétation car mis à part certains regroupements de gènes appartenant aux mêmes clusters, le réseau a éclaté la majorité des clusters de coexpression. L'analyse des 18 catégories de stress confirme ce résultat puisque très peu d'enrichissement d'interactions PPI et TF-cibles ont été observés parmi les 681 clusters de coexpression. Comme cela a été déjà vu dans la littérature, j'ai constaté ainsi le faible chevauchement entre les données de coexpression, PPI et TF-cibles.

Ce phénomène pourrait provenir d'une certaine hétérogénéité de contexte des données de coexpression. Effectivement ces données sont liées à un contexte environnemental précis tel qu'un stress hydrique ou une attaque bactérienne. Les données PPI ou TF-cibles dont je dispose sont au contraire obtenues pour la plupart indépendamment de toute condition environnementale. Les données PPI sont générées par la technique de double hybride chez la levure qui sort les protéines de

leur contexte cellulaire et élimine les variations environnementales pouvant les influencer. Bien que les interactions TF-cibles soient influencées par le contexte environnemental dans lequel elles ont été identifiées, dans mon projet je n'ai pris en compte que les interactions validées au moins deux fois de manière indépendante. Cette intégration de contexte permet de considérer ces données TF-cibles comme indépendantes de tout contexte. Par la suite je définirai comme données absolues toutes interactions indépendantes du contexte telles que ces données PPI et TF-cibles. Pour associer nos données de coexpression avec ces données absolues, il faut tenir compte de cette hétérogénéité du type des données. Transformer les données absolues en données contextuelles étant impossible, la seule solution est de transformer les données de coexpression en données absolues.

Chapitre 2 : De la coexpression à la corégulation

I. Contexte et Objectifs

Le chapitre précédent a montré la nécessité de transformer les données de coexpression en données absolues pour pouvoir les intégrer avec les données PPI et TF-cibles.

L'analyse de coexpression par les modèles de mélange sur toutes les expériences transcriptomiques simultanément aurait pu être une solution mais elle s'est révélée insatisfaisante. Mon objectif est donc d'effectuer une analyse transversale des différentes catégories de stress. Cette approche me permet d'une part de m'affranchir de la variable contexte environnement tout en tenant compte de la spécificité de la coexpression (quels gènes sont coexprimés dans quels stress). D'autre part, en identifiant les relations de coexpression observées dans plusieurs catégories de stress, cette approche transforme les données de coexpression en données de corégulation ce qui pourra être bénéfique pour l'inférence de fonction. En effet, dans la littérature, certaines études ont déjà remis en question la coexpression car ils considèrent qu'elle est insuffisante pour suggérer la corégulation des gènes et supposer leur implication dans les mêmes fonctions biologiques (D'haeseleer *et al.* 2000). L'inconvénient est que cette approche ne garde que les réponses ubiquitaires en éliminant les unités de coexpression spécifiques correspondant aux signatures de réponse de la plante à un type de stress. Cependant ces informations des gènes spécifiques sont conservées dans l'analyse de coexpression et il a été montré que la majorité des gènes de réponse au stress n'est pas spécifique (Rodriguez et Redman, 2005 ; Kilian *et al.* 2007). Cela semble être le cas de nos données (figure 2 de l'introduction).

Au cours de ce chapitre je détaillerai la procédure d'intégration des données de coexpression mise en place afin d'identifier des interactions de corégulation ainsi que l'évaluation statistique des interactions identifiées. Une section sera consacrée également à l'analyse de la topologie et la caractérisation des réseaux inférés à partir des interactions identifiées.

II. Création du réseau de corégulation

Quelques études d'intégration de la coexpression ont été effectuées dans la littérature. Certaines de ces études collectent un ensemble de jeu des données transcriptomiques du génome humain de différentes expériences et différents laboratoires et effectuent une analyse de coexpression pour chaque jeu de données séparément et intègrent ensuite les résultats pour la construction d'un seul réseau utilisé pour la recherche de modules de gènes ultra-connectés (Lee *et al*. 2004). L'étude de Yan *et al*. (2007) porte sur une approche différente pour l'analyse du transcriptome humain qui vise à identifier des modules conservés dans plusieurs réseaux de coexpression. Atias *et al*. (2009) analysent 43 projets transcriptomiques d'Arabidopsis qui représentent 857 échantillons hybridés issus de 37 laboratoires différents. Ils calculent la coexpression des gènes via la corrélation de Pearson dans chaque projet séparément, puis proposent de calculer un score fondé sur la fréquence des gènes coexprimés dans chaque jeu de données pour intégrer les données de coexpression des différents projets. Dans d'autres études, l'intégration des données de coexpression a été effectuée à partir de plusieurs espèces pour étudier la conservation des gènes coexprimés, en supposant qu'elles indiquent une conservation de la fonction des gènes impliqués dans ces interactions (Stuart *et al*. 2003). De la même manière nous supposons que si les relations de coexpression entre les gènes sont conservées dans plusieurs conditions, ce sera probablement en raison de leur corégulation. La qualité et la pertinence biologique et fonctionnelle de ces interactions seront d'une plus haute valeur et notamment pour l'inférence de fonction. Je définis ainsi un groupe de gènes corégulés comme un groupe de gènes coexprimés dans un large panel de stimuli et de conditions biologiques en l'occurrence ici les conditions de stress.

1. Identification des couples de gènes corégulés

Pour intégrer les clusters de coexpression obtenus pour chaque catégorie de stress, j'identifie les couples de gènes présents dans un même cluster de coexpression pour plusieurs catégories de stress. En partant d'une matrice indiquant, pour chaque gène du projet, le numéro du cluster auquel il appartient pour chaque catégorie de stress, l'identité et l'occurrence des paires de gènes appartenant aux mêmes clusters dans plus de deux catégories de stress, sont identifiées. Le tableau 7 résume les occurrences de ces couples en fonction du nombre de catégories de stress dans lesquelles ils sont conservés. Ces résultats montrent que notre approche permet d'identifier environ 295 000 paires de

gènes qui sont coexprimés dans au moins 2 catégories de stress, et 36 paires de gènes ayant une réponse transcriptionnelle coordonnée dans au moins 12 catégories de stress. Deux paires de gènes sont conservées dans 14 types de stress. La coordination du profil d'expression de ces deux paires de gènes dans 14 catégories de stress n'est manifestement pas due au hasard et souligne probablement la corégulation des gènes en question et leur implication dans le même processus biologique. Ces résultats soulèvent ainsi la question de définir le seuil ou le nombre de catégories de stress minimum à partir duquel la coexpression peut être considérée comme corégulation.

Tableau 7 : Occurrence des paires de gènes conservées dans au moins P conditions de stress.
P allant de 2 à 18 catégories de stress. Par exemple la colonne 2+ désigne le nombre de paires de gènes conservés dans au moins 2 catégories de stress (2 ou plus).

Nbr stress	2+	3+	4+	5+	6+	7+	8+	9+	10+	11+	12+	13+	14+	15+
Nbr paires	295 105	57 833	19 087	7 997	3 782	1 908	903	395	190	83	36	9	2	0

2. Evaluation statistique de la validité des liens

L'objectif de cette étape est d'évaluer statistiquement la validité des relations liant les paires de gènes identifiées dans les mêmes clusters de coexpression à travers de nombreuses conditions de stress et surtout de savoir à partir de quel seuil nous pouvons considérer que ces paires de gènes coexprimés sont en réalité corégulés. Pour cela, j'ai comparé mes résultats aux résultats qui seraient obtenus dans un réseau aléatoire. Cette comparaison a été effectuée grâce à un test statistique décrit dans le paragraphe suivant.

a. *Test statistique*

Afin d'évaluer la pertinence du réseau de corégulation, j'ai utilisé un test de permutation. Soit M, la matrice d'appartenance des gènes aux clusters par catégorie de stress, de taille N x P, où N est le nombre de gènes et P est le nombre de catégories de stress. Dans la matrice M, un élément à la ligne i, colonne j, correspond au numéro de cluster auquel le gène i appartient pour le stress j. J'ai généré 1 000 matrices M' en permutant aléatoirement les éléments dans chaque colonne de M. De cette manière l'appartenance des gènes aux clusters pour chaque catégorie de stress est déterminée au hasard. Pour chacune de ces matrices M', j'ai compté le nombre de paires de gènes qui se trouvent dans le même cluster dans au moins p catégories de stress pour p allant de 2 à 18. Puis un taux

d'erreur a été calculé correspondant à la moyenne d'occurrence des paires dans les matrices aléatoires, divisé par leur occurrence dans la matrice observée.

b. *Résultats*

Les résultats du test de permutation sont résumés dans le tableau 8. Ces résultats montrent que 68 338 paires de gènes peuvent être coexprimées dans au moins 2 catégories de stress par hasard, contre 295 105 paires identifiées par notre approche d'intégration. Le taux d'erreur calculé en divisant l'occurrence des paires de gènes prises au hasard par l'occurrence des paires identifiées à partir de la matrice d'origine, indique que 23,1 % des arêtes du réseau biologique peuvent être dues au hasard à un seuil de 2 catégories de stress ou plus. Au seuil de 3 catégories de stress ou plus, seulement 733 paires de gènes sont conservées par hasard contre 57 833 paires de gènes dans notre réseau biologique, soit un taux d'erreur de 1,2%. Si je considère un taux d'erreur toléré à 5%, ce seuil est donc significatif. De la même manière, seulement 5 paires de gènes peuvent être coexprimées par hasard dans au moins 4 types de stress, alors que notre analyse identifie 19 087 paires de gènes. Le test à ce seuil est significatif puisque j'obtiens un taux d'erreur inférieur à 1% (0,026%). Le tableau 8 montre également qu'à partir du seuil de 5 stress, il est quasiment impossible de trouver des couples de gènes coexprimés par hasard.

Ces résultats suggèrent alors fortement que dans notre réseau biologique les gènes coexprimés dans au moins 4 catégories de stress sont coregulés et pas seulement coexprimés.

Tableau 8 : Résultats du test de permutation.

Comparaison de l'occurrence du nombre de paires de gènes conservées dans au moins p stress dans la matrice d'origine avec la moyenne de cette occurrence dans les matrices aléatoires générées par le test de permutation.

Nombre de paires / Nombre de stress	Matrices aléatoires (moyenne)	Matrice D'origine	Taux d'erreur
2+	68 338	295 105	0.231
3+	733	57 833	0.012
4+	5	19 087	0.00026
5+	0.02	7 997	0
6+	0	3 366	0

III. Visualisation et Analyse du réseau

1. Visualisation et description globale des réseaux obtenus

Pour visualiser ces résultats, j'ai mis en place un graphe où les nœuds sont les gènes et je mets une arête valuée entre deux gènes si cette paire de gènes a été vue dans les mêmes clusters de coexpression dans au moins p catégories de stress. La valeur de l'arête est le nombre de catégories de stress dans lesquelles les 2 gènes sont coexprimés et reflète la force de la corégulation entre les deux gènes. Nous avons choisi de créer des réseaux incluant toutes les arêtes d'une valeur supérieure ou égale à p. Le tableau 9 résume la taille des réseaux obtenus en termes de nombre d'arêtes et nombre de nœuds. Le nombre de facteurs de transcription, de gènes mal caractérisés autrement dit non annotés pour une ontologie et le nombre d'orphelins identifiés au sein de ces réseaux sont également indiqués. Les réseaux obtenus pour chaque seuil de corégulation sont visualisés grâce à l'outil Cytoscape (figure 18). Ainsi plus le seuil augmente, plus la taille du réseau est petite et plus le réseau est épars (ou « sparse » en anglais).

Concernant la caractérisation fonctionnelle des gènes dans les réseaux obtenus, le plus petit réseau de corégulation correspondant au seuil 14 est composé de 3 nœuds (AT1G05420, AT1G69240 et AT1G16510) reliés par 2 arêtes. Ces 3 gènes ne sont pas connus dans la littérature pour être corégulés. Pourtant nous trouvons plusieurs indications montrant leur coopération en plus de leurs réponses coordonnées dans 14 catégories de stress parmi 18 étudiées. En effet, les promoteurs des 3 gènes présentent le motif GATT connu sous le nom d'EIN3 et qui joue un rôle important dans le contrôle transcriptionnel des composantes de signalisation immunitaire. Les gènes AT1G05420 et AT1G69240 reliés par une arête présentent au sein de leurs promoteurs un même motif cis-régulateur « ABRE » connu pour être impliqué dans la défense de la plante et notamment dans la régulation de la réponse à la déshydratation et la salinité chez Arabidopsis et le riz (Yamaguchi-Shinozaki et Shinozaki 2006). Le premier gène est connu pour être impliqué dans la défense de la plante alors que le deuxième gène est annoté comme un régulateur négatif de la transcription (BP). Le gène AT1G05420 est également lié au gène AT1G16510 et ils partagent un motif « MYB » classiquement retrouvé dans les promoteurs des gènes de réponse à la déshydratation. Notons que parmi les catégories de stress dans lesquelles ces deux gènes sont coordonnés, nous retrouvons la sécheresse. Le gène AT1G16510 est annoté en tant que régulateur de la croissance (naturellement

affectée par les perturbations environnementales). Il est également impliqué dans la réponse et le transport de l'auxine.

Le réseau de corégulation au seuil 13 est composé de 8 gènes dont 3 facteurs de transcription. Rappelons que le gène AT1G05420 est impliqué dans la régulation négative de la transcription. Il permet ici de relier tous les gènes du réseau y compris les 3 facteurs de transcription. Ce réseau présente un enrichissement significatif en gènes reliés à l'activité de facteur de transcription en ontologie MF et en localisation « noyau » pour l'ontologie CC. Le réseau est enrichi pour le motif EIN3, présent au sein de 6 promoteurs. Parmi les 8 gènes du réseau, 5 gènes partagent également le motif cis-régulateur ABRE et 3 gènes partagent le motif MYB.

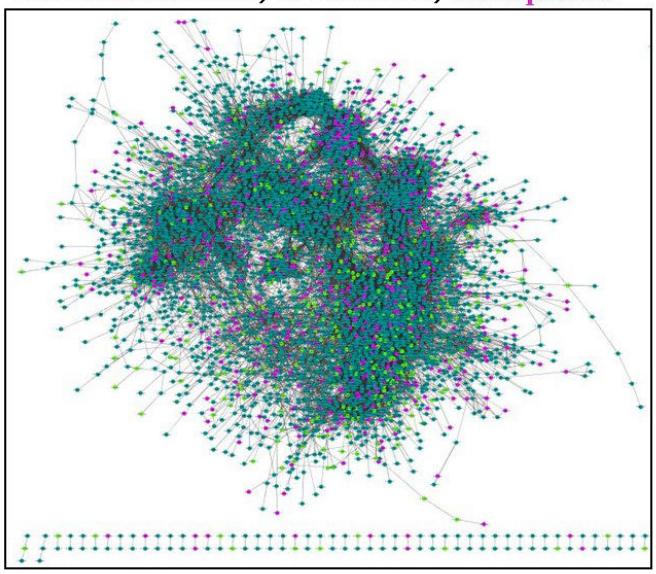
A l'opposé, le réseau de corégulation au seuil 3 stress ou plus est composé de 57 833 arêtes et 5 626 gènes. Ce réseau présente de nombreux enrichissements notamment en facteurs de transcription et en hormones. De nombreux termes GO sont également sur-représentés au sein de ce réseau par rapport à tous les gènes d'*A. thaliana* y compris la réponse au stress, le transport et la transduction du signal. Il est susceptible d'éclairer la fonction de 711 gènes orphelins et d'apporter des éléments de réponse autour de la fonction BP de 1 660 gènes mal caractérisés. Parmi les 5 626 gènes de ce réseau, 479 gènes sont des facteurs de transcription. La figure 19 représente la distribution des degrés des nœuds de ce réseau qui suit une loi de puissance typique des réseaux biologiques. Cette caractéristique se traduit par la présence d'un grand nombre de nœuds à faible degré et d'un petit nombre au degré élevé, appelés des « hubs ». La médiane tracée en rouge sur l'histogramme des distributions des degrés des nœuds indique que 50% des gènes ont un degré supérieur à 9. La distribution du réseau en loi de puissance nous rassure quant à la qualité des données obtenues puisque cette caractéristique est considérée comme un critère pour mesurer la qualité et la pertinence des réseaux inférés (Gillis et Pavlidis, 2012). Certaines études considèrent que les données qui sont en conflit avec cette distribution sont de faible qualité (Gomez *et al*. 2001; Zhang et Horvath, 2005).

Tableau 9 : Informations sur les réseaux de gènes par seuil de catégories de stress.

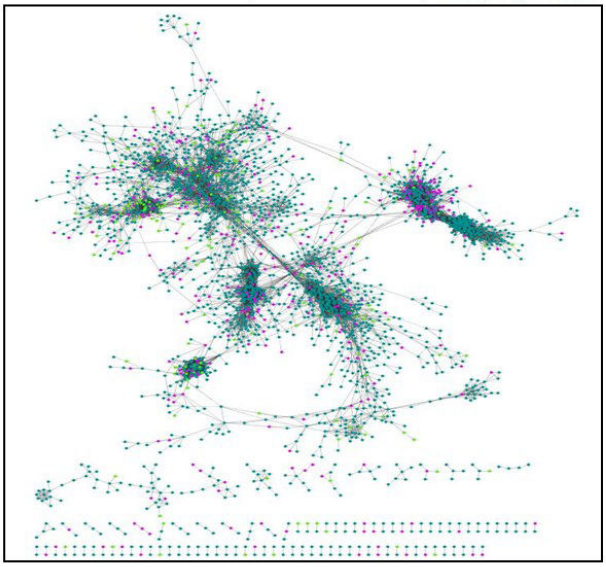
Tableau résumant la taille des réseaux de gènes corégulés dans au moins n catégories de stress (n variant de 3 a 14) énumération du nombre d'arêtes, de nœuds, de facteurs de transcription, de gènes sans annotation pour l'ontologie BP, MF et CC ainsi que le nombre de gènes orphelins identifiés dans chaque réseau.

Stress	Arêtes	Nœuds	TF	Sans_BP	Sans_MF	Sans_CC	Orphelins
3+	57833	5626	479	1660	1848	455	711
4+	19087	2815	195	757	920	224	352
5+	7997	1401	91	343	432	117	155
6+	3782	738	47	176	223	63	82
7+	1908	415	31	91	116	31	40
8+	903	233	22	54	61	16	20
9+	395	120	12	23	33	6	8
10+	190	64	10	13	18	4	6
11+	83	32	6	7	8	3	3
12+	36	20	5	7	8	3	3
13+	9	8	3	3	3	1	0
14+	2	3	0	1	2	0	0

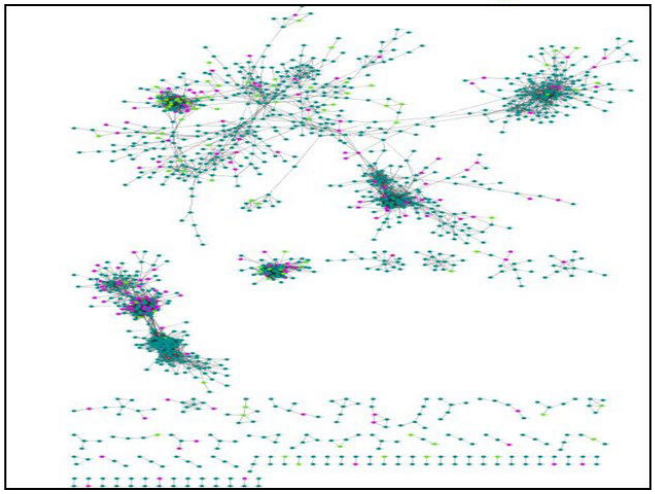
Seuil 3: 5626 nœuds, 57 833 arêtes, 713 orphelins



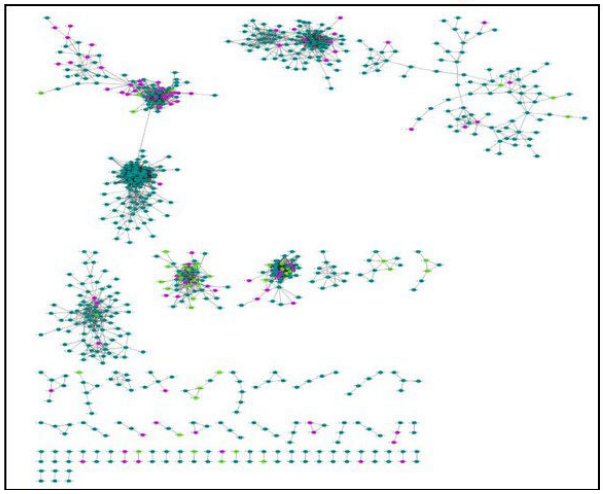
Seuil 4: 2815 nœuds, 19 087 arêtes, 352 orphelins



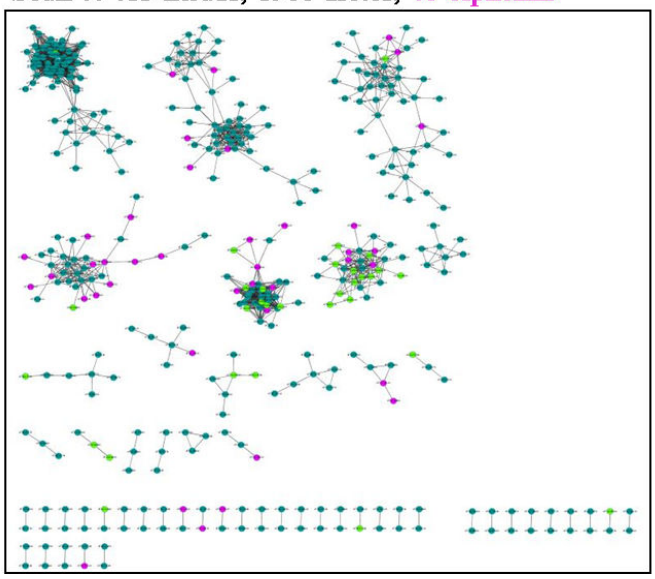
Seuil 5: 1401 nœuds, 7997 arêtes, 155 orphelins



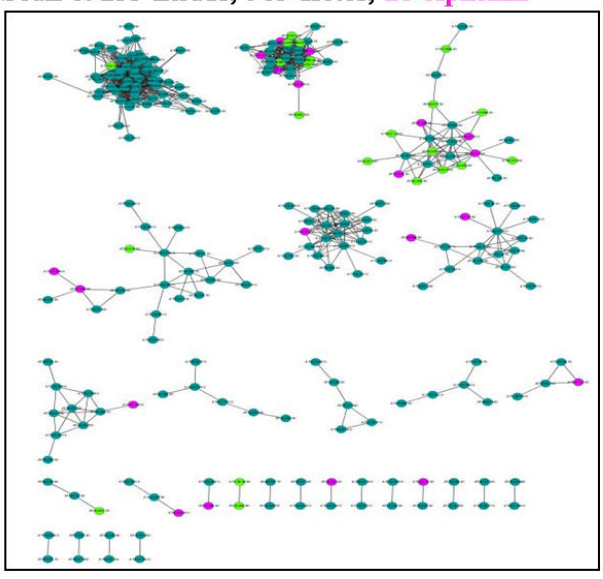
Seuil 6: 738 nœuds, 3782 arêtes, 82 orphelins



Seuil 7: 415 nœuds, 1908 arêtes, 40 orphelins



Seuil 8: 233 nœuds, 903 arêtes, 20 orphelins



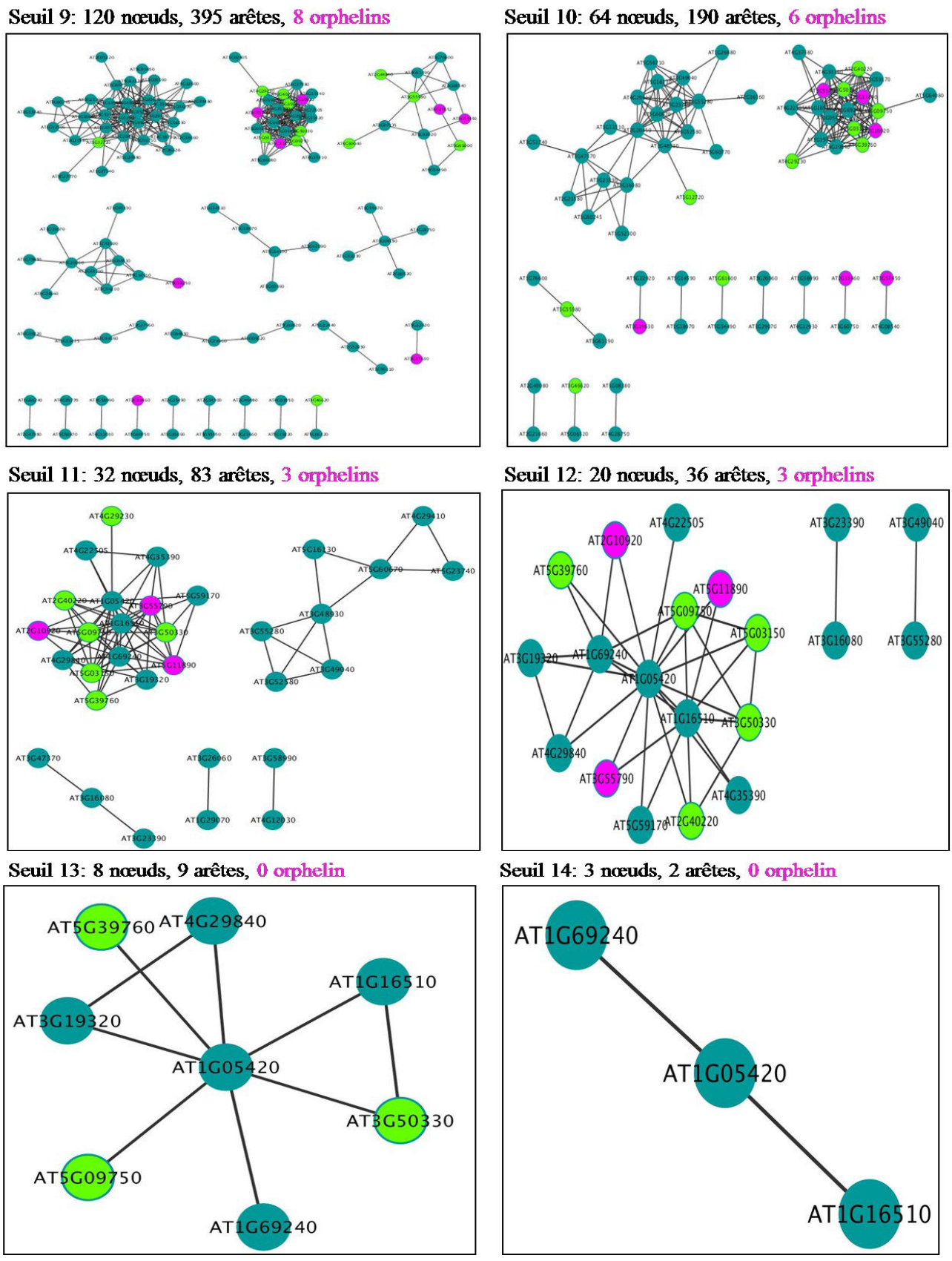


Figure 18 : Représentation graphique de tous les réseaux obtenus pour chaque seuil de corégulation.

Les nœuds roses correspondent aux gènes orphelins, les verts pistache correspondent aux TF

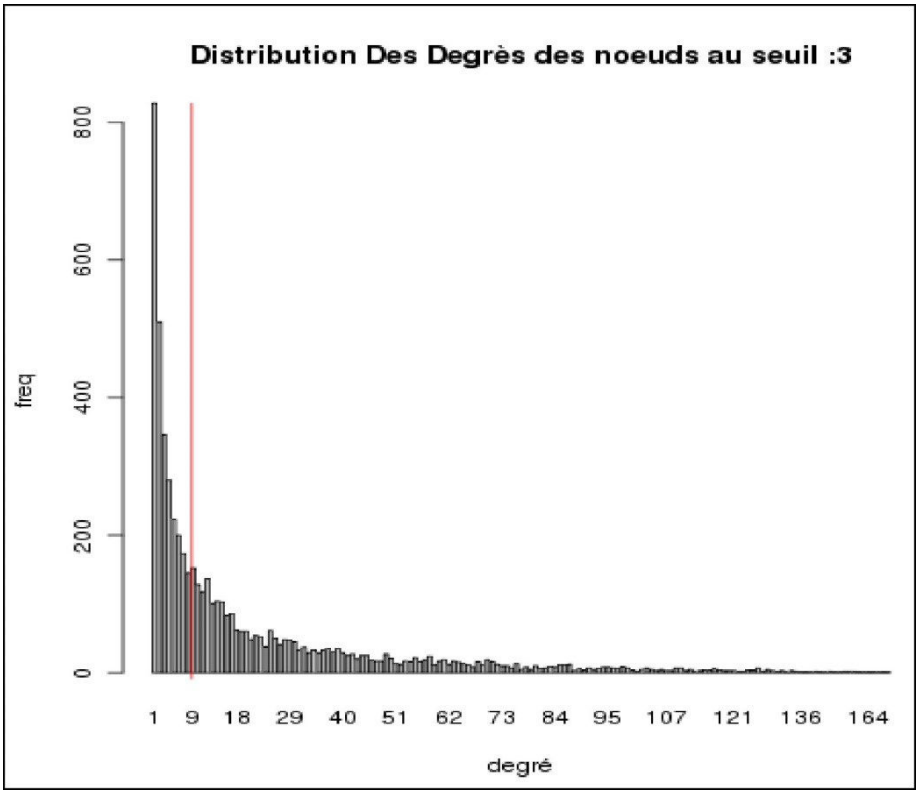


Figure 19 : Distribution des degrés des nœuds du réseau biologique obtenu avec les paires de gènes corégulés dans au moins 3 types de stress.
La droite rouge représente la médiane des degrés des nœuds.

2. Propriétés topologiques du réseau au seuil 7

Pour évaluer le potentiel de nos réseaux de corégulation et illustrer les avantages de notre approche en termes d'inférence de fonction, j'ai décidé d'utiliser le réseau de corégulation au seuil de 7. A ce seuil, les paires de gènes correspondent à des gènes hautement corégulés et le réseau a une taille raisonnable pour une illustration claire. Le réseau se compose de 1 908 arêtes impliquant 415 gènes (figure 20). La distribution des degrés des nœuds de ce réseau est également gouvernée par une loi de puissance avec une médiane égale à 4 (figure 21). La topologie du réseau à ce seuil de coexpression est intéressante et révèle une structure modulaire composée de sous-structures hautement connectées : les gènes sont organisés en 19 composantes connexes suggérant que les membres de ces composantes pourraient être impliqués dans la même fonction biologique ou voie indépendamment du reste des gènes du réseau. Mon objectif à ce stade d'analyse est d'annoter les groupes de gènes obtenus (composantes) et d'étudier leurs biais fonctionnels afin de vérifier si les

membres de ces composantes partagent des propriétés similaires et sont impliqués dans un même processus biologique.

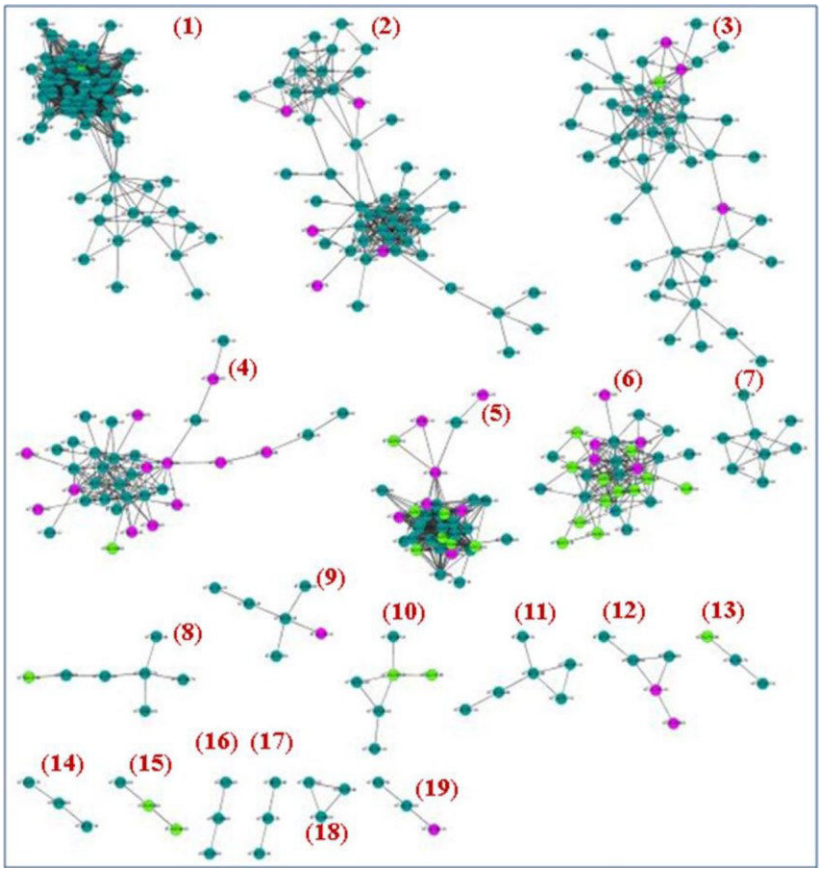


Figure 20 : Réseau de corégulation au seuil 7+, visualisé avec Cytoscape.

Les nœuds roses correspondent aux gènes orphelins, les verts pistache correspondent aux TF. Les numéros indiqués en rouge entre parenthèses correspondent aux numéros attribués aux composantes pour faciliter leur identification.

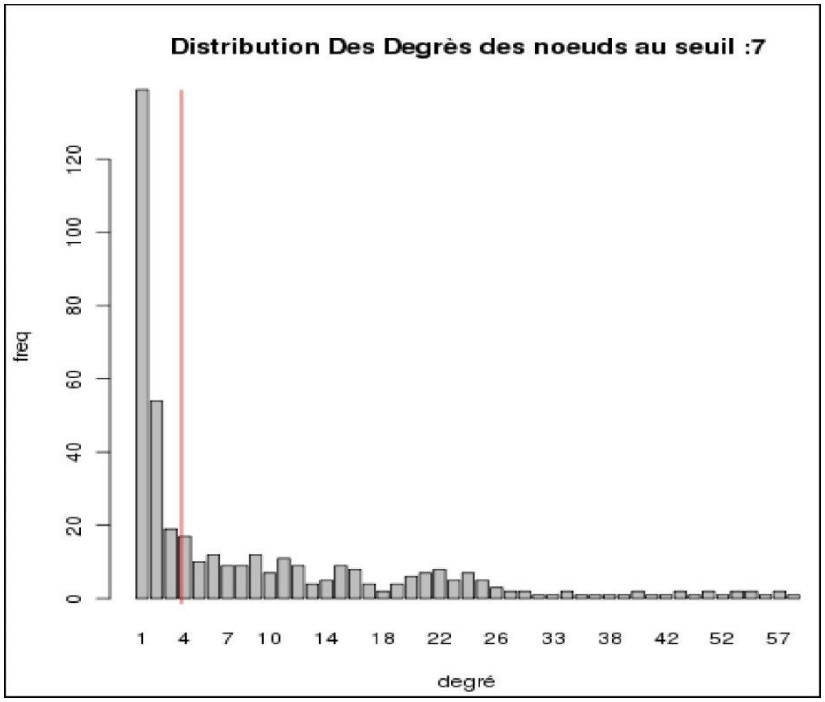


Figure 21 : Distribution des degrés des nœuds du réseau de corégulation au seuil 7.

3. Caractérisation fonctionnelle des modules

Afin d'évaluer la pertinence biologique des composantes connexes obtenues, les pipelines d'annotation fonctionnelle présentés dans le chapitre 1 ont été appliqués. J'ai effectué ces analyses pour chacune des 19 composantes d'une part, et pour la totalité des gènes dans les 19 composantes d'autre part (composante totale) (figures 22, 23 et 24). Les enrichissements des termes GO ont révélé que 68,42 % des composantes ont au moins un terme sur-représenté par rapport au génome pour l'ontologie BP et 42% sont enrichis en termes de l'ontologie MF et CC. De plus, la majorité des composantes enrichies a seulement quelques termes sur-représentés voir un seul et a une plus large proportion de gènes partageant ces termes par rapport à l'ensemble des gènes du réseau. Ces résultats montrent que les composantes sont homogènes et spécifiques. Cette uniformité des annotations GO est particulièrement informative pour l'inférence de fonction puisqu'elle permet d'appuyer fortement la propagation des termes sur-représentés au sein de ces groupes aux gènes orphelins ou mal caractérisés. La différence des proportions des enrichissements entre les ontologies MF et BP est attendue comme démontré dans l'introduction à savoir que les réseaux d'interactions moléculaires sont plus performants pour la prédiction des fonctions de l'ontologie BP que de l'ontologie MF. Cependant pour les composantes ayant un enrichissement en termes MF, cet enrichissement est

spécifique et pertinent où seulement quelques termes sont enrichis dans une composante à la fois et couvre une large proportion de gènes dans le cluster. Comme le montre la figure 23, certains termes MF tels que « transcription factor activity » ou « molecule binding » ne sont pas révélés sur-représentés dans la composante totale, mais sont sur-représentés dans des composantes connexes soulignant ainsi des biais spécifiques dans ces composantes.

Si nous prenons l'exemple de la composante « 1 » indiquée en rouge sur la figure 20, nous trouvons qu'elle présente des enrichissements intéressants, puisque 66 gènes parmi les 81 gènes de la composante sont annotés par le terme « structural molecule activity » de l'ontologie MF. Si nous regardons l'annotation de ces gènes à un niveau moins élevé de l'architecture GO que la GO Slim, alors nous trouvons que ces 66 gènes sont annotés en tant que constituants structuraux de ribosome. Cet enrichissement est largement en accord avec de précédentes études de corégulation chez A. *thaliana* (Haberer *et al*. 2006; Wei *et al*. 2006; Horan *et al*. 2008 ; Sormani *et al*. 2011). En effet les ribosomes sont connus pour jouer un rôle crucial dans la régulation de la croissance et leur expression est très impactée par les facteurs environnementaux. De plus selon l'étude de Horan (2008) l'activité ribosomale nécessite la coordination de nombreuses protéines pour la formation de complexes. Il n'est donc pas surprenant que les gènes correspondants soient étroitement corégulés. L'identification de cette composante ribosomale nous rassure donc quant à la qualité des composantes obtenues et leur pertinence biologique.

J'ai testé également l'enrichissement de ces composantes en éléments cis-régulateurs ou sites de fixation des facteurs de transcription (TFBS). Si ces groupes de gènes sont corégulés alors ils doivent être sous le contrôle des mêmes facteurs de transcription et ainsi être enrichis en motifs cis-régulateurs expliquant cette corégulation. Pour l'ensemble des gènes du réseau de corégulation au seuil 7, j'ai identifié 24 motifs sur-représentés par rapport à tout le génome. Parmi ces motifs, 6 sont des motifs de régulation des gènes induits par la lumière tels que l'I-box, et 3 motifs sont impliqués dans la régulation et la répression du sucre. L'analyse de l'enrichissement des 19 composantes connexes a révélé également que 10 composantes sont enrichies en motifs cis-régulateurs (voir tableau 10). Parmi les 6 motifs reliés à la régulation de lumière sur-représentés dans le réseau au seuil 7, 5 motifs se retrouvent regroupés au sein de la composante numéro 2 et représentent 50% des motifs sur-représentés dans cette composante. La composante 1 est enrichie avec 17 motifs cis régulateurs impliqués dans différents processus dont 2 motifs (AAACAAA et AGCAGC) trouvés dans les gènes aérobies et impliqués dans les voies de fermentation. Ces deux motifs sont présents dans 37,5 % et 12,5 % des promoteurs des gènes de cette composante contre 24,97% et 2,95%

respectivement dans tout le génome. Cette composante est enrichie également en deux motifs reliés à la réponse au stress température : le motif CCAATbox responsable de la réponse aux chocs de chaleur et le motif CCGAAA (ou LTRE-1) responsable de la réponse aux basses températures. La composante numéro 8 est enrichie en 2 motifs qui sont le EIN3 (régulation de signalisation immunitaire) et un motif de fixation des facteurs DOF impliqués dans le développement et la croissance de la plante (Yanagisawa 2004). Ces 2 motifs sont présents dans 100% des promoteurs de la composante analysée. La composante 10 est également enrichie en 2 motifs dont le EIN3 qui est présent dans tous les promoteurs de cette composante et le motif GATA nécessaire pour l'expression spécifique du phloème. Les composantes 16 et 19 présentent également chacune un motif sur-représenté et présent dans tous les promoteurs de la composante à chaque fois.

Ces analyses d'enrichissement en TFBS valident notre approche en montrant que les gènes sont sous le contrôle des mêmes régulateurs confirmant ainsi leur corégulation.

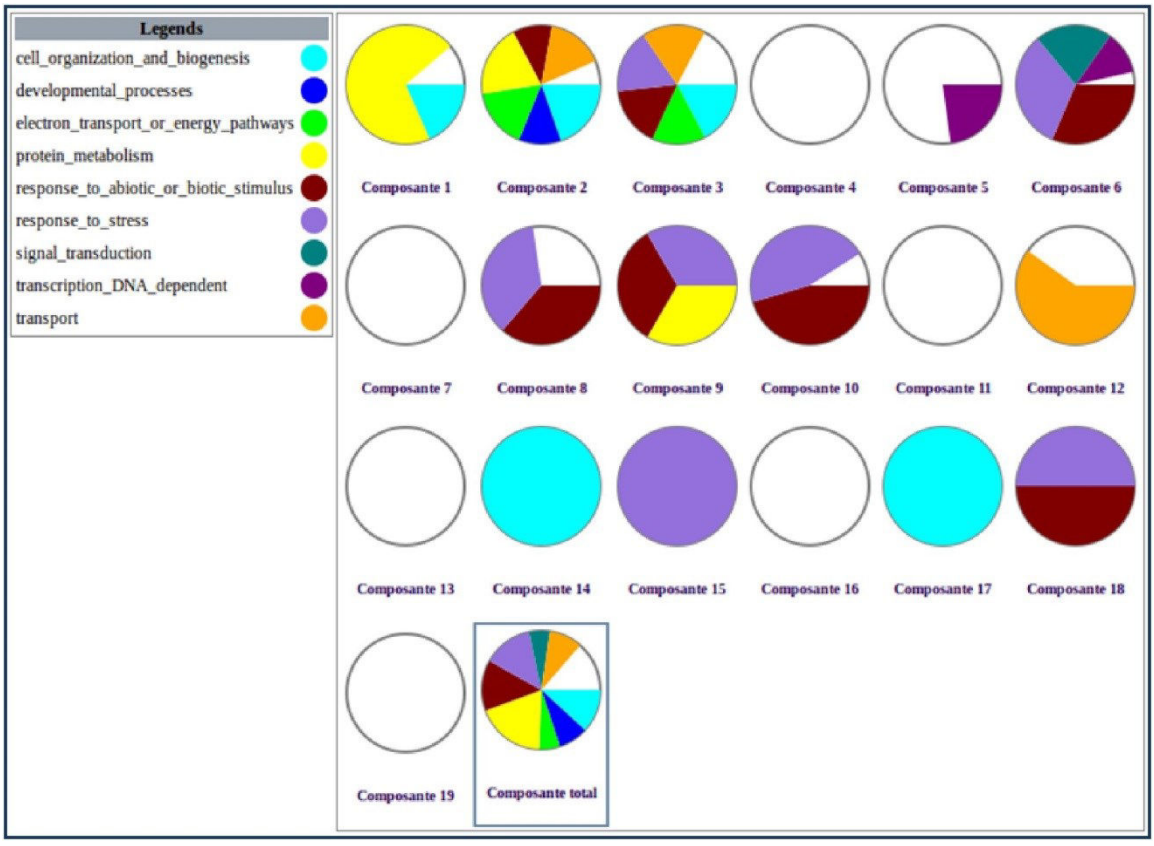


Figure 22 : Représentation graphique de l'enrichissement des composantes connexes en termes GO de l'ontologie BP.

La composante totale correspond aux gènes constituant le réseau de corégulation au seuil 7.

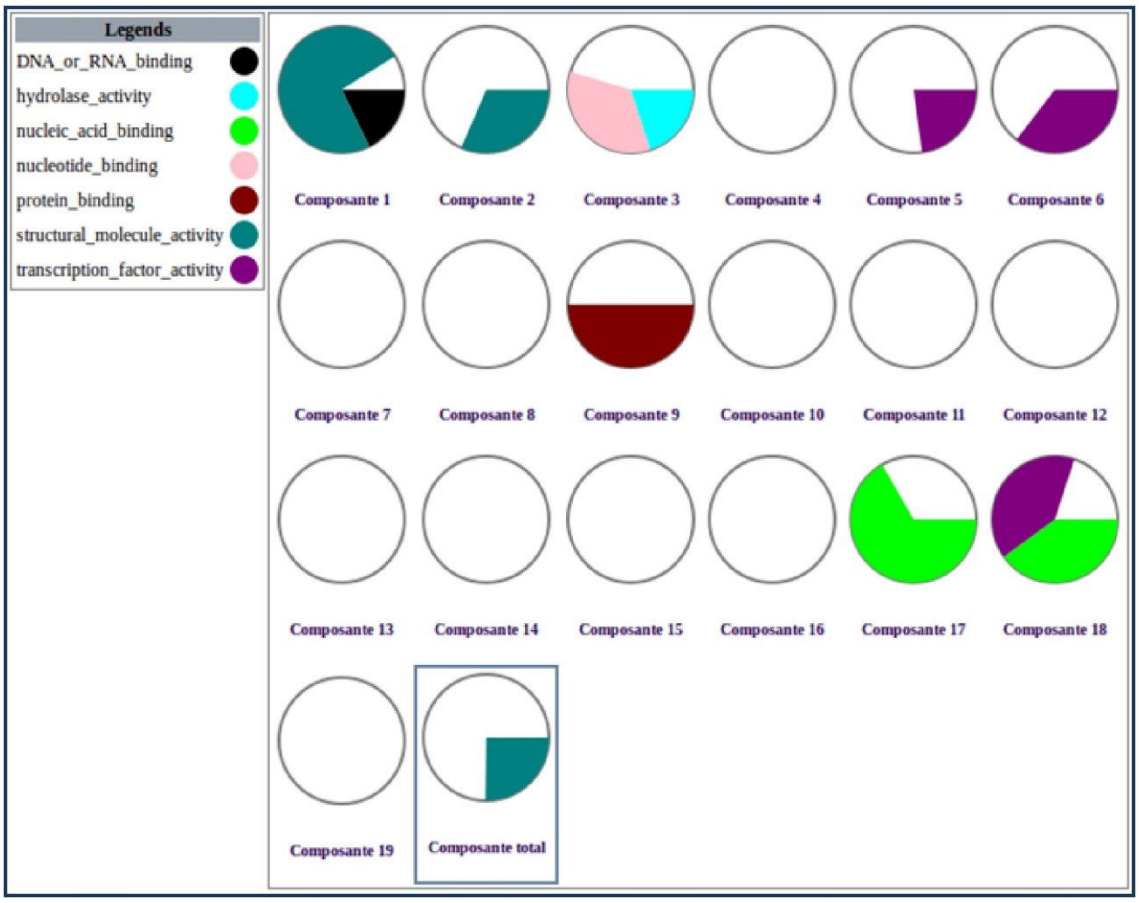


Figure 23 : Représentation graphique de l'enrichissement des composantes connexes en termes GO de l'ontologie MF.

La composante totale correspond aux gènes constituant le réseau de corégulation au seuil 7.

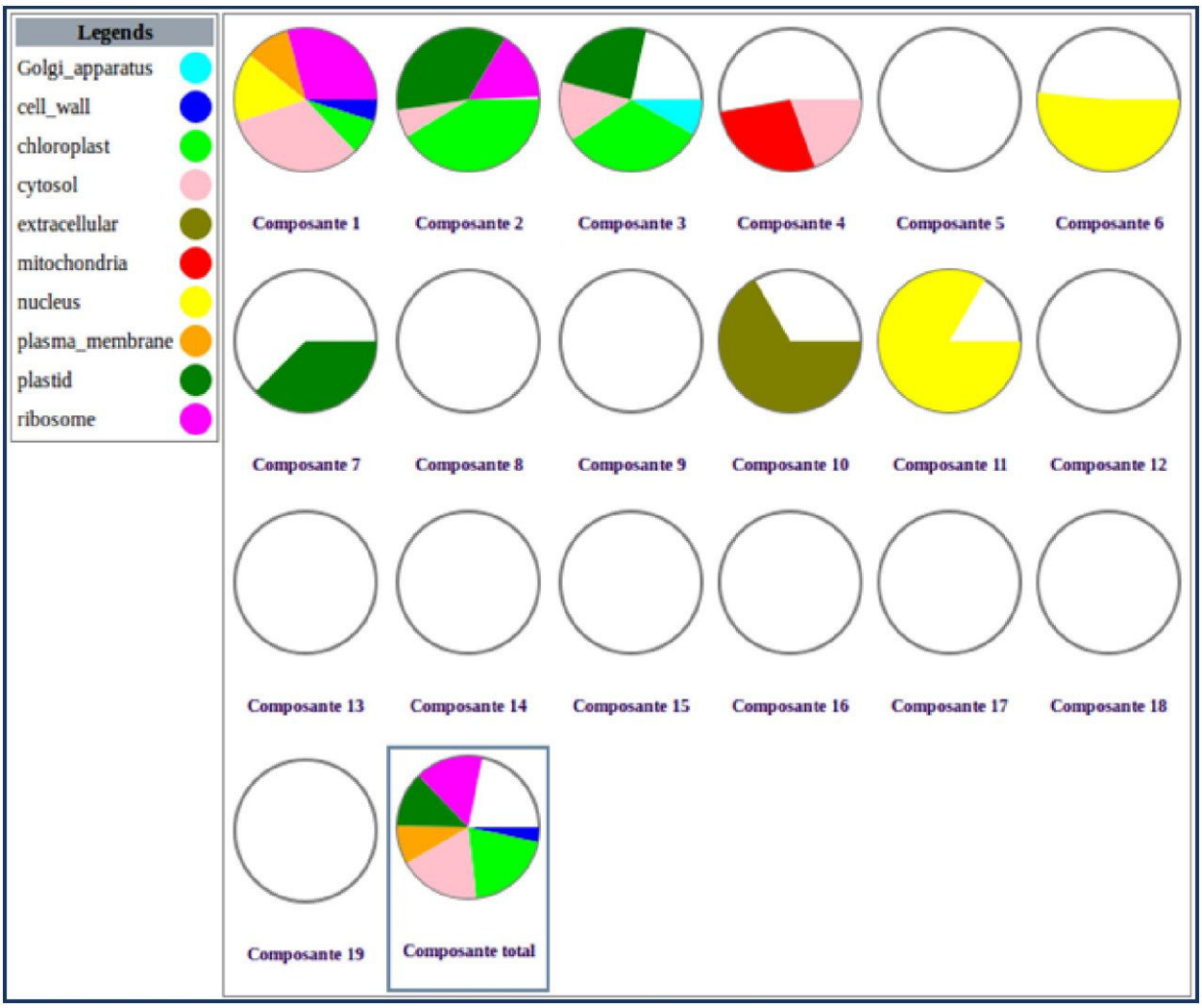


Figure 24 : Représentation graphique de l'enrichissement des composantes connexes en termes GO de l'ontologie CC.

La composante totale correspond aux gènes constituant le réseau de corégulation au seuil 7.

Tableau 10 : PLMs détectés et déterminés comme sur-représentés par rapport au génome au sein de chaque composante connexe du réseau de corégulation au seuil 7.

Composante	Taille cluster	Nbr promoteurs	Motif	Fenêtre fonctionnelle		Promoteurs ayant le motif			P-value
						cluster		génome	
				début	fin	nbr	%	%	
Compos_1	81	80	TATATAA	-67	38	22	27.5	11.97	4.35E-05
			TATAWA	-49	18	45	56.25	32.39	3.33E-06
			TTTATATA	-91	86	16	20	8.57	0.00038212
			CCAAT	-91	-33	27	33.75	16.78	6.29E-05
			AACCCA	-152	-26	12	15	7.71	0.00802972
			AAACAAA	-257	-87	30	37.5	24.97	0.00449554

			GCGT	-32	111	30	37.5	23.01	0.00115621
			ACGC	-25	104	28	35	21.51	0.0018113
			AGCAGC	-8	140	10	12.5	2.95	2.36E-05
			TATTAG	-100	92	20	25	10.13	3.45E-05
			CCGAAA	-195	-33	11	13.75	6.51	0.00551562
			GGATA	-256	-	11	13.75	5.88	0.00243918
			TAAAG	-27	21	22	27.5	15.63	0.00212525
			AACCAA	-261	-	13	16.25	7.89	0.00377623
			AGAAA	-39	-25	16	20	10.96	0.00547208
			TATTCT	-1	44	7	8.75	3.58	0.00792572
			CAAT	-67	-58	18	22.5	10.8	0.00073027
Compos_2	54	54	GATAA	-37	59	35	64.81	23.18	1.37E-11
			GGATA	-60	89	27	50	14.93	2.48E-10
			TATCCA	-78	117	20	37.04	7.93	2.99E-10
			GATA	-23	49	42	77.78	40	3.21E-09
			TGTCA	-110	-28	16	29.63	14.93	0.00165736
			WGATAR	-45	44	26	48.15	22.94	1.29E-05
			GCCAC	-133	53	15	27.78	14.96	0.00450507
			CANNTG	-72	-36	15	27.78	13.8	0.00197113
			RYCGAC	-255	-	7	12.96	4.28	0.0020117
			AACCAA	-276	-	11	20.37	10.36	0.00824094
Compos_3	49	47	GATA	-148	-	18	38.3	20.37	0.00134153
			CAAT	-130	-	15	31.91	15.16	0.00104366
			TAAAATAT	-204	-88	7	14.89	4.88	0.00183169
			GAGAC	55	111	12	25.53	9.18	0.00023003
			AACCAA	52	100	6	12.77	4.31	0.00381176
			RTTTTTR	-129	-71	16	34.04	18.01	0.00241136
Compos_4	36	34	CCAAT	-108	-24	16	47.06	22.22	0.00034542
			AGAAA	-37	-10	15	44.12	19.8	0.00031002
			CANNTG	-74	-42	9	26.47	13.04	0.00946298
Compos_5	35	26	TATAWA	-178	49	21	80.77	55.31	0.0016397
			TATAAAT	-122	62	10	38.46	15.97	0.00126919
			AAAG	-121	-91	17	65.38	42.42	0.00526614
			GRWAAW	-147	-103	16	61.54	37.94	0.00419079
			GTGA	19	62	12	46.15	26.42	0.00859175
			GATT	-265	-244	14	53.85	26.25	0.00068617
		31	TGACY	-236	-140	15	48.39	25.33	0.00151264
			AATAAA	-238	-156	18	58.06	28.19	0.00012235
			AAAG	5	27	19	61.29	37.42	0.00201286
			TATAWA	-76	14	19	61.29	36.86	0.00162683
Compos_7	8	Pas de motif significativement sur-représenté							

Compos_8	7	7	AAAG	-113	18	7	100	97.3	0
			GATT	-328	-261	7	100	59.43	0
Compos_9	6	Pas de motif significativement sur-représenté							
Compos_10	6	6	GATA	-25	63	5	83.33	45.05	0.00835928
			GATT	-157	-	6	100	66.13	0
Compos_11	6	Pas de motif significativement sur-représenté							
Compos_12	5	Pas de motif significativement sur-représenté							
Compos_13	3	Pas de motif significativement sur-représenté							
Compos_14	3	Pas de motif significativement sur-représenté							
Compos_15	3	Pas de motif significativement sur-représenté							
Compos_16	3	3	AAAG	55	28	3	100	60.29	0
Compos_17	3	Pas de motif significativement sur-représenté							
Compos_18	3	3	GATT	-6	14	3	100	73.92	0
Compos_19	3	Pas de motif significativement sur-représenté							

4. Exemple de caractérisation d'un gène mal annoté

Pour illustrer le potentiel des réseaux de corégulation pour l'inférence de fonction, j'ai considéré la composante connexe numéro 1 du réseau au seuil 7 (indiqué en rouge sur la figure 20) comme un exemple pour la caractérisation d'un gène mal annoté au sein de ce module. Cette composante est constituée de 945 interactions impliquant 81 gènes. Rappelons que selon les analyses d'enrichissement, cette composante correspond à un complexe de protéines ribosomales. Elle est enrichie en termes de métabolisme de protéine, d'organisation et de biogenèse de la cellule pour l'ontologie BP et en termes liés à l'activité moléculaire de la structure et de la liaison à l'ADN ou l'ARN pour l'ontologie MF. Je me suis focalisée sur l'analyse du gène AT3G49040 non annoté pour l'ontologie BP et MF au sein de ce module. La seule information disponible autour de la fonction de ce gène actuellement dans TAIR concerne les domaines InterPro identifiés dans ce gène notamment le domaine Fbox et le domaine Leucine-Rich Repeat (LRR). Le domaine Fbox est un domaine conservé présent dans plusieurs protéines présentant une structure bipartite (telles que les protéines ribosomales). Les protéines contenant le domaine LRR sont impliquées dans différents processus biologiques y compris la transcription, la transduction du signal, la résistance et la réponse immunitaire. Ces informations confirment d'une part l'implication de ce gène inconnu dans la réponse au stress et ainsi sa présence dans le réseau de corégulation des gènes impactés dans la réponse à 18 catégories de stress, d'autre part cela suggère une structure proche de celle des

protéines ribosomales constituant le complexe dans lequel il est impliqué. L'étude de la topologie de cette composante permet d'identifier 55 voisins directs de ce gène. Parmi ceux là, 48 sont annotés en tant que constituants structuraux de ribosome et 14 voisins sont annotés avec le terme liaison à l'ARN pour l'ontologie MF. Ces résultats suggèrent ainsi que le gène AT3G49040 code probablement pour une protéine ribosomale ou une protéine impliquée dans la liaison à l'ARN. La comparaison de la séquence nucléique et de la séquence protéique de ce gène avec les séquences correspondantes des gènes voisins n'a pas révélé de similarité de séquence significative. Etant donné qu'il est connu que les séquences des protéines ribosomales sont très conservées, ces résultats nous dirigent vers deux hypothèses. Soit ce gène est associé à la liaison à l'ARN et joue ainsi un rôle dans la régulation de ce complexe protéique, soit il s'agit d'une protéine ribosomale de structure non classique qui, même en n'ayant pas une séquence proche des autres protéines ribosomales du voisinage, présente toutefois un domaine Fbox des protéines bipartites.

Pour l'annotation BP, 53 voisins du gène AT3G49040 sont annotés avec le terme de métabolisme de protéine et suggèrent ainsi que ce gène est annoté probablement avec ce terme. Les résultats de cette analyse permettent de mieux cibler la fonction de ce gène inconnu et donnent des indices forts autour du lien de ce gène avec les protéines ribosomales ou leurs régulations. Ces résultats représentent ainsi une base solide pour un projet biologique guidé pour la confirmation de la fonction de ce gène.

IV. Conclusion

L'intégration des données contextuelles à partir des clusters de coexpression a permis d'identifier des liens entre des paires de gènes ayant la même dynamique de réponse dans de nombreuses conditions de stress allant jusqu'à 14 catégories de stress pour certains couples. Ces interactions ont servi pour la construction de réseaux de corégulation. Les arêtes de ces réseaux sont valuées en fonction du nombre de catégories de stress dans lesquelles les deux partenaires sont coexprimés et informent ainsi sur l'importance de la corégulation entre les gènes. Afin d'évaluer la pertinence des liens identifiés, un test de permutation a été effectué et a permis de montrer la significativité de ces liens à partir d'un seuil à 3 catégories de stress. Ainsi le réseau de corégulation au seuil 3 peut être considéré avec beaucoup de confiance puisque seulement 1,26% des paires de gènes dans ce réseau peuvent être corégulées par hasard.

Les enrichissements des réseaux obtenus et notamment ceux des composantes connexes du réseau au seuil 7 stress ou plus, montrent la force et l'homogénéité des groupes de gènes corégulés identifiés. Ces résultats montrent également la force des liens fonctionnels qui relient les gènes dont la qualité dépasse celle des liens de coexpression des gènes dans une seule catégorie de stress.

La transformation des données contextuelles en données absolues a permis ainsi de passer des données de coexpression à des données de corégulation reliant les gènes avec des liens fonctionnels plus forts. Cette approche a permis également de passer d'une analyse des gènes indépendante par catégorie de stress à une analyse transversale intégrant les informations autour des gènes impactés par les 18 catégories de stress simultanément. De plus, elle a permis la construction de réseaux de gènes qui pourront faciliter l'étape d'inférence de fonction via la propagation des fonctions des gènes connus aux gènes inconnus. L'exemple de caractérisation fonctionnelle du gène mal caractérisé dans le paragraphe précédent montre le potentiel de cette approche et du réseau de corégulation pour l'inférence fonctionnelle notamment pour l'ontologie BP. Cependant pour l'inférence de fonction à haut débit, une méthode automatique est nécessaire afin de caractériser l'ensemble des gènes orphelins et mal caractérisés au sein de ces réseaux.

Chapitre 3 : Annotation fonctionnelle à haut-débit utilisant le réseau de corégulation

I. Contexte et objectifs

L'objectif principal de ce chapitre est l'exploitation du réseau de corégulation pour la prédiction des fonctions des gènes orphelins n'ayant aucune annotation ou mal caractérisés qui sont les gènes n'ayant pas d'annotation pour une branche de l'ontologie GO. Comme décrit en introduction, plusieurs méthodes d'inférence exploitant les réseaux d'interactions moléculaires ont déjà été proposées dans la littérature. Elles sont généralement centrées sur une caractérisation des gènes par des termes. Pourtant, il a été montré que la performance de prédiction varie en fonction de l'ontologie et même des termes analysés (Radivojac *et al*. 2013 ; Ryngajllo *et al*. 2011). Partant de ce constat, j'ai opté pour un autre angle de vue qui est de développer une méthode binaire d'apprentissage supervisée centrée sur les termes par ontologie et non centrée sur les gènes. Cette méthode permet pour chaque terme de prédire si un gène doit être annoté ou pas avec ce terme.

Le principe de la méthode est de considérer une collection de classifieurs qui calculent pour chaque gène un score représentatif de la présence du terme analysé dans son voisinage au sein du réseau de corégulation. Chaque classifieur appliqué à un ensemble de gènes permet d'obtenir une liste décroissante de scores. Une collection de règles de décision est ensuite créée en définissant un score seuil associé à chacune de ces listes qui contrôle la proportion des faux positifs parmi les gènes prédits positifs. Afin de mesurer la performance de la méthode, le jeu de travail est découpé en un jeu d'apprentissage pour définir les listes des scores et les scores seuils associés et un jeu test pour mesurer la performance de ces règles. De plus, j'ai utilisé une procédure de cross-validation (CV) afin de considérer plusieurs jeux d'apprentissage et de test à partir d'un même jeu de travail. La meilleure règle par terme est alors déterminée par l'évaluation de plusieurs métriques et est ensuite appliquée aux gènes orphelins ou mal caractérisés. Grâce à la procédure de cross-validation, j'ai pu aussi proposer un indice de confiance de la prédiction. Le schéma général de la méthode d'inférence est représenté dans la figure 25. Ce chapitre est organisé de la manière suivante : la première section est consacrée à la description du jeu de travail, de la méthode d'apprentissage et des métriques d'évaluation utilisées dans cette analyse. La deuxième section décrit la collection des classifieurs et les paramètres qui les définissent ainsi qu'une analyse de sensibilité de ces paramètres. La section

suivante décrit la collection de règles de décisions ainsi que leur évaluation. La quatrième section présente les résultats de l'application des règles de décision sélectionnées aux gènes orphelins ou mal caractérisés au sein du réseau. Pour finir, la dernière section du chapitre permet de conclure et de discuter les résultats obtenus et les problématiques rencontrées liées à la complexité de l'apprentissage et de l'évaluation dans le cadre de l'annotation fonctionnelle.

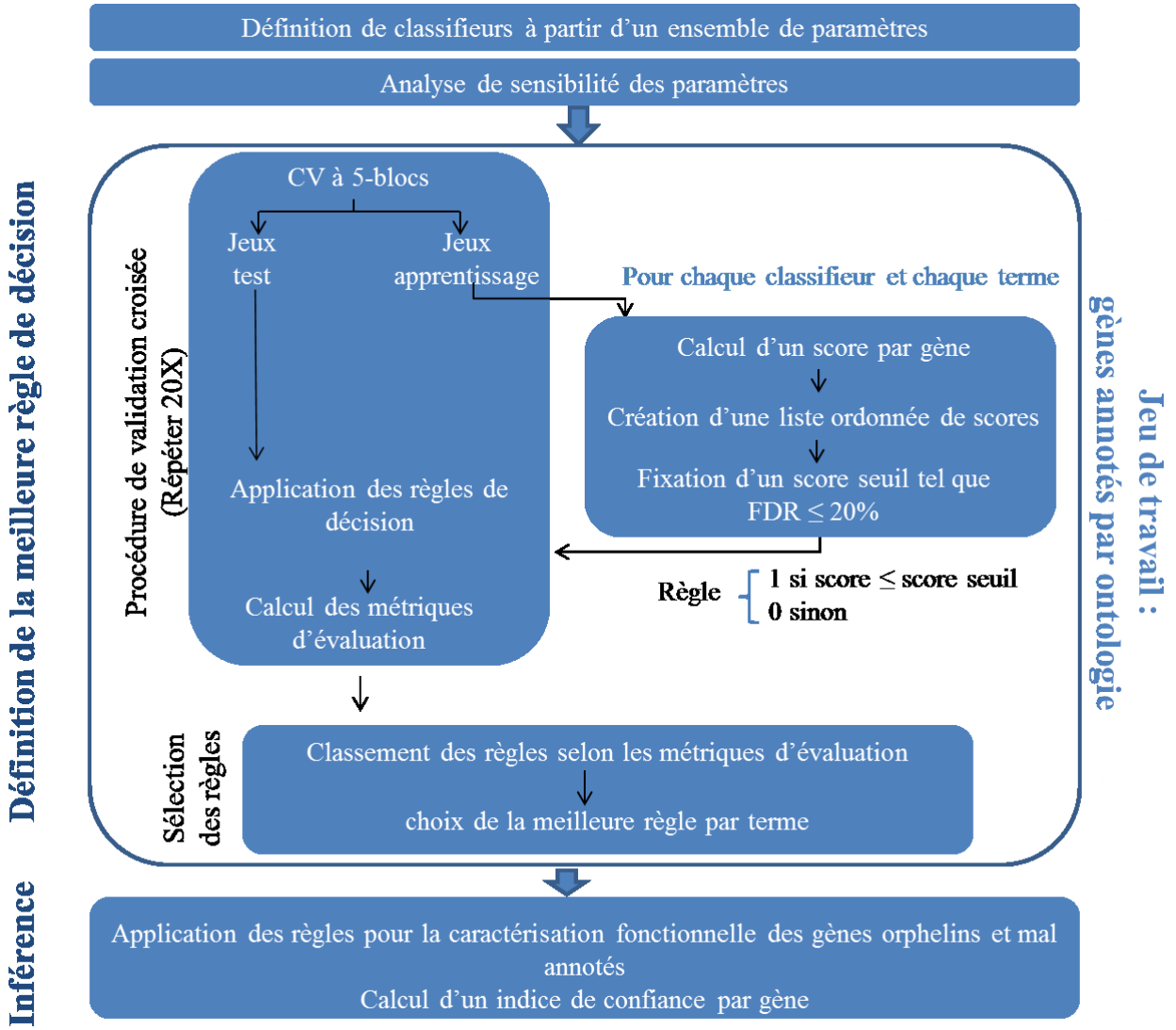


Figure 25 : Schéma général de la procédure d'inférence

II. Description des données, méthodes et métriques d'évaluation

1. Définition du jeu de travail

Un jeu de travail est défini pour chacune des trois ontologies BP, CC et MF en fonction des gènes présents dans le réseau de corégulation à un seuil donné. Ainsi pour tous les termes d'une ontologie, l'ensemble des gènes du jeu de travail est le même, par contre la composition en exemples positifs (gènes annotés par le terme) et exemples négatifs (les gènes n'ayant pas dans leur annotation le terme) dépend du terme.

Les termes étudiés sont les termes du plus haut niveau de la hiérarchie GO, qui sont les termes GO Slim. Ils permettent d'avoir une vision globale des annotations fonctionnelles des gènes ainsi que de limiter le déséquilibre entre les exemples positifs et négatifs. Malgré cela, certains termes sont associés à très peu de gènes ce qui complique leur analyse et dans la littérature ces termes peu représentés, généralement moins de 10 à 20 fois dans le jeu de travail, sont écartés des analyses (Yu *et al*. 2015 ; Radivojac *et al*. 2013 ; Letovsky and Kasif 2003). J'ai appliqué dans cette analyse la même démarche et j'ai considéré uniquement les termes représentés par au moins 20 gènes. Le tableau 11 indique la taille du jeu de travail par ontologie et par seuil de corégulation selon ces critères. Pour être précis et en accord avec les résultats du chapitre 2, nous devrions parler de réseau de coexpression lorsque le seuil est égal à 1 ou 2 puis de réseau de corégulation dès que le seuil est supérieur ou égal à 3. Ici par simplicité de langage, nous oublierons cette subtilité et parlerons toujours de réseau de corégulation même si le seuil est égal à 1 ou 2. D'après le tableau 11, nous observons que pour les réseaux où les gènes sont coexprimés dans au moins 12, 13 et 14 catégories de stress, aucun terme n'est retenu. Par conséquent nous ferons varier ce seuil entre 1 et 11. Nous pouvons constater aussi que pour les trois premiers seuils tous les termes des annotations GO Slim de chaque ontologie sont représentés.

Tableau 11 : Caractérisation du jeu de travail par seuil de corégulation et par ontologie.
Le seuil 1 correspond au réseau de couples de gènes coexprimés dans au moins une catégorie de stress. Pour chaque seuil le nombre de termes GO analysés et le nombre de gènes annotés par ontologie sont indiqués.

Seuil	Ontologie	nbr termes	gènes annotés	Ontologie	nbr termes	gènes annotés	Ontologie	nbr termes	gènes annotés
1	BP	13	8 036	MF	14	7 932	CC	15	11 120
2	BP	13	6 480	MF	14	6 318	CC	15	8 725
3	BP	13	3 966	MF	14	3 778	CC	15	5 171

4	BP	13	2 058	MF	13	1 895	CC	15	2 591
5	BP	13	1 058	MF	13	969	CC	15	1 284
6	BP	12	562	MF	12	515	CC	13	675
7	BP	12	324	MF	10	299	CC	13	384
8	BP	12	179	MF	6	172	CC	11	217
9	BP	8	97	MF	2	87	CC	8	114
10	BP	4	51	MF	1	46	CC	5	60
11	BP	2	25	MF	0	-	CC	0	-

2. Principe de la validation croisée

Généralement un jeu de travail est divisé en deux sous-échantillons distincts. Le premier, appelé jeu d'apprentissage, est utilisé pour l'estimation des paramètres de la règle de décision du modèle. Le deuxième jeu de données, appelé jeu test, est utilisé pour estimer l'erreur de prédiction. Pratiquement la règle de décision est apprise sur le jeu d'apprentissage et est ensuite appliquée sur le jeu test. Ces prédictions sont ensuite comparées avec la vérité. L'intérêt de ce découpage est de tester la règle de décision sur des exemples indépendants afin d'éviter le sur-apprentissage (Boulesteix 2009) et d'estimer la capacité de généralisation du modèle à de nouveaux exemples. Afin que les résultats ne soient pas dépendants du découpage considéré, il est possible d'utiliser la validation croisée, connue sous le terme anglais « cross-validation ».

Elle consiste à itérer plusieurs fois le découpage aléatoire du jeu de travail afin de faire varier le jeu d'apprentissage et le jeu test. Dans cette étude, j'ai utilisé précisément la validation croisée à 5-blocs qui consiste à diviser de manière aléatoire le jeu de travail en cinq sous-échantillons disjoints : un sous-échantillon est sélectionné pour définir le jeu test et les 4 autres forment le jeu d'apprentissage. Cette sélection est répétée jusqu'à ce que chaque sous-échantillon ait été utilisé comme jeu test.

3. Définition des métriques d'évaluation

Dans un système de prédiction binaire, il existe quatre résultats possibles : les vrais positifs (VP), les vrais négatifs (VN), les faux positifs (FP) et les faux négatifs (FN). Les VP et VN sont les prédictions correctes alors que les FP et les FN sont les deux sortes de mauvaises classifications (Tableau 12).

Tableau 12 : Matrice de confusion.

		Réalité	
		Positif	**Négatif**
Prédiction	Positif	VP (Vrais Positifs)	FP (Faux Positifs)
	Négatif	FN (Faux Négatifs)	VN (Vrais Négatifs)

i. Métriques

Ces quatre résultats sont utilisés pour générer des métriques d'évaluation faciles à interpréter et à comparer. La sensibilité et la précision sont des mesures de performance très utilisées dans la littérature.

La sensibilité appelée aussi TPR (True Positive Rate) ou encore le rappel, mesure la proportion des exemples positifs prédits correctement.

$$Sensibilité = \frac{VP}{VP + FN}$$

La précision appelée aussi PPV (Positive Predictive Value) mesure la proportion des exemples positifs parmi les prédictions positives.

$$Précision = \frac{VP}{VP + FP}$$

J'ai également calculé le FDR « False Discovery Rate », qui est le complément de la précision. le FDR mesure le taux de faux positifs parmi les prédictions positives :

$$FDR = \frac{FP}{VP + FP} = 1 - Précision$$

Le FDR est couramment utilisé pour le contrôle des faux positifs dans les tests d'hypothèses lors de la conduite de comparaisons multiples, comme par exemple pour l'identification des gènes différentiellement exprimés à partir de données transcriptomiques. Cette métrique est également appropriée pour les problèmes de classification, pourtant, à notre connaissance, elle n'est jamais utilisée comme un critère de performance en annotation fonctionnelle. L'objectif de l'utilisation de cette mesure dans notre étude est de contrôler le taux de faux positifs parmi les gènes prédits positifs pour chaque terme afin de garantir que la validation expérimentale soit concluante.

Ces critères sont parfois critiqués sur le fait qu'ils sont réducteurs car pour avoir une idée globale de la performance du modèle il faut les étudier simultanément. Ainsi, d'autres critères ont été proposés afin de résumer l'information. C'est le cas du critère Fmeas qui combine la précision et la sensibilité. Cette métrique prend des valeurs comprises entre 0 et 1, et est calculée de la manière suivante :

$$Fmeas = \frac{2(précision.sensibilité)}{(précision + sensibilité)}$$

ii. Courbes ROC et calcul de l'aire sous la courbe

Les courbes ROC « Receiving Operating Characteristics » représentent l'évolution de la proportion de vrais positifs TPR (True Positive Rate ou sensibilité) en fonction de la proportion de faux positifs FPR (False Positive Rate).

Elles permettent ainsi de mesurer et de comparer la capacité des classifieurs à discriminer les exemples positifs des exemples négatifs, sans pour autant assigner une classe à chaque exemple. Cependant, les courbes ROC sont des outils graphiques difficiles à utiliser dès que le nombre de méthodes à comparer est grand. C'est pourquoi, nous calculons l'AUC (Area Under the Curve) qui est un indicateur associé à la courbe ROC qui est plus synthétique, plus facile à interpréter et à comparer. L'AUC représente l'aire sous la courbe ROC et indique la probabilité que le modèle place un exemple positif avant un exemple négatif. Un classifieur parfait aura une valeur d'AUC égale à 1. Par contre un classifieur qui ne prédit aucun exemple positif correctement aura une valeur de 0. Un modèle aléatoire aura une AUC de 0.5.

III. Description des classifieurs et des paramètres considérés

1. Définition des classifieurs et des paramètres

Un classifieur est une fonction définie à un ensemble de paramètres près et qui calcule un score pour chaque gène du jeu d'apprentissage. Ce score mesure la présence du terme considéré dans son voisinage dans le réseau de corégulation, et son calcul dépend des paramètres de la fonction. Un classifieur appliqué à l'ensemble d'un jeu d'apprentissage, permet d'ordonner les gènes grâce à un tri décroissant des scores.

Dans cette étude, nous avons considéré un ensemble de paramètres pouvant potentiellement avoir une influence sur la valeur du score et le classement des gènes du jeu d'apprentissage. Nous les avons organisés en trois catégories.

1) La première catégorie regroupe les paramètres ayant un impact sur la définition du voisinage d'un gène. Elle inclut le seuil de corégulation considéré ainsi que le type de voisinage.
2) La seconde catégorie regroupe les paramètres qui sont directement reliés au calcul du score. Cette catégorie inclut le type de décompte et le type de graphe.
3) La troisième catégorie concerne le type de classement utilisé pour comparer les scores des gènes du jeu d'apprentissage. Ces paramètres sont représentés schématiquement dans la figure 26 et détaillés dans les paragraphes suivants de cette section.

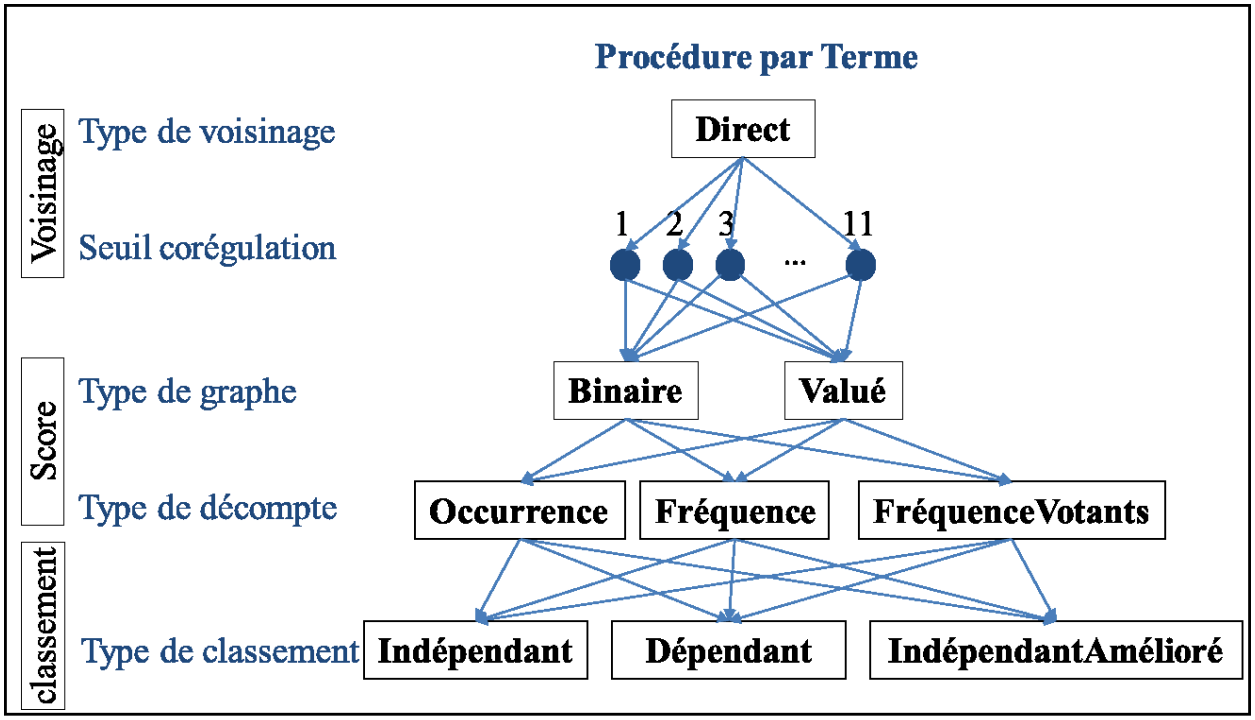


Figure 26 : Représentation schématique des paramètres considérés.

a. *Paramètres de voisinage*

i. Seuil de corégulation

Comme vu dans le deuxième chapitre, il est important de tenir compte de l'occurrence de coexpression des paires de gènes. Cette occurrence définit la force du lien fonctionnel entre les paires de gènes. J'ai également montré dans le chapitre précédent qu'en considérant toutes les occurrences supérieures à un certain seuil, appelé seuil de corégulation, cela impactait la taille du

réseau (voir tableaux 7 et 9) et donc influence le voisinage de chaque gène. C'est pourquoi ce seuil de corégulation est considéré comme un paramètre dans l'approche de prédiction.

ii. *Type de voisinage*

Le voisinage d'un gène dans le réseau de corégulation peut être défini soit par les gènes auxquels il est relié par une arête dans le graphe représentant le réseau, soit par les gènes situés dans un même module identifié grâce à une analyse topologique du graphe. D'après une étude comparative de Sharan *et al.* (2007), la spécificité d'une méthode de vote majoritaire dans le voisinage direct est supérieure à celle d'une méthode qui passe par l'identification de modules. Cette même étude a également mis en évidence la corrélation entre la distance dans le réseau et la similarité fonctionnelle: plus les deux protéines sont proches dans le réseau, plus leurs annotations fonctionnelles sont proches.

Dans mon étude, la structure modulaire identifiée par les composantes connexes du réseau de corégulation nécessite de considérer un seuil supérieur ou égal à 7 ce qui réduit considérablement la taille du jeu de travail. L'exploitation du voisinage direct ayant prouvé son efficacité lors de la caractérisation du gène mal annoté lors du chapitre 2, j'ai choisi de travailler avec les voisins directs d'un gène dans le réseau de corégulation.

b. *Paramètres de score*

i. *Type de graphe*

Le réseau de corégulation est un graphe valué où les arêtes indiquent le nombre de catégories de stress dans lesquelles la paire de gènes est coexprimée. La question est de savoir si la valeur de l'arête apporte plus d'information que son existence pour l'annotation. Ainsi les scores de gènes dans le réseau peuvent être calculés de deux manières. La première consiste à considérer le poids de l'arête dans ce graphe valué. La deuxième manière consiste à considérer que le graphe est non valué, c'est alors un graphe binaire où toutes les arêtes ont une valeur égale à 1.

ii. *Type de décompte*

Pour un terme analysé, le score d'un gène reflète la présence du terme dans son voisinage. Pour le calculer j'ai considéré trois types de décompte:

- ***Occurrence*** : Pour un graphe binaire, le score est égal au nombre d'apparition du terme dans le voisinage. Pour un graphe valué, le score est la somme des valeurs des arêtes des voisins annotés par le terme.
- ***Fréquence :*** le score est égal à la fréquence du terme dans le voisinage. Cela revient à diviser l'occurrence par la taille du voisinage pour un graphe binaire. Pour un graphe valué, cela revient à diviser par la somme des valeurs des arêtes de tous les voisins.
- ***Fréquence Votants :*** parmi le voisinage, certains gènes peuvent ne pas être annotés, donc non associés au terme considéré. Pour un graphe binaire, le score est l'occurrence divisée par le nombre de gènes annotés. Pour un graphe valué, c'est l'occurrence divisée par la somme des valeurs des arêtes de ces voisins annotés.

c. *Paramètre de classement des scores*

i. Type de classement des scores

Pour un terme analysé, un classifieur associe à chaque gène un score qui reflète la présence du terme dans son voisinage. Plus le score est élevé, plus le terme est présent dans le voisinage du gène. Il est donc naturel de classer les gènes selon leurs scores. Pour tenir compte du fait qu'il y ait de nombreux gènes avec le même score et que les termes d'une même ontologie sont dépendants, j'ai défini trois tris décris ci-dessous et j'ai illustré avec l'exemple d'un graphe binaire constitué de 5 gènes pouvant être annotés par 4 termes (tableau 13).

Tableau 13 : Illustration d'une matrice de comptage.
Selon cet exemple, le terme 1 apparait une fois dans le voisinage du gène A alors que le terme 2 apparait 5 fois dans ce même voisinage.

	Terme1	Terme2	Terme3	Terme4
GèneA	1	5	2	6
GèneB	1	1	1	2
GèneC	8	4	0	0
GèneD	0	2	3	1
GèneE	0	7	1	0

- ***Classement indépendant des autres termes***

Pour un terme donné, ce classement ordonne les scores de manière décroissante. Les gènes ayant des scores ex-æquo sont ordonnés de manière aléatoire.

- ***Classement dépendant des autres termes***

Contrairement au type de classement précédent, celui-ci considère les scores calculés pour les autres termes d'une même ontologie. Cela permet ainsi de tenir compte de la dépendance et de la corrélation des termes dans le voisinage. Pour chaque gène, en comparant ses scores obtenus pour tous les termes d'une l'ontologie, les rangs de ces termes dans son voisinage sont alors déduits. Ainsi les rangs des termes sont déterminés par gène. La matrice de comptage de scores du tableau 13 donnera la matrice des rangs illustrée par le tableau 14. Comme notre approche est centrée sur une discrimination des gènes par terme, alors pour chaque terme analysé, un classement croissant de ses rangs dans le voisinage de tous les gènes est effectué. Les gènes ex-æquo c'est-à-dire les gènes ayant le terme analysé au même rang dans leur voisinage, sont départagés grâce à leurs scores en les classant des plus forts scores aux plus faibles scores.

Reprenons l'exemple du tableau 13. Selon ces règles, le classement des gènes potentiellement annotés avec le terme1 dans cet exemple, sera alors le suivant : le gène C, le gène B, le gène E, le gène A et le gène D**.** Dans cet exemple, le terme 1 apparait au même rang pour les deux gènes A et D. Pour les classer j'ai utilisé leurs scores pour le terme 1.

Tableau 14 : Illustration d'une matrice des rangs générée à partir de la matrice de comptage du tableau 13.

	Terme1	Terme2	Terme3	Terme4
GèneA	Rang 4	Rang 2	Rang 3	Rang 1
GèneB	Rang 2	Rang 2	Rang 2	Rang 1
GèneC	Rang 1	Rang 2	Rang 3	Rang 3
GèneD	Rang 4	Rang 2	Rang 1	Rang 3
GèneE	Rang 3	Rang 1	Rang 2	Rang 3

- **Classement indépendant mais amélioré par le rang des autres termes**

J'ai également proposé un troisième type de classement à partir des deux types précédents. Pour un terme analysé, un classement décroissant des scores des gènes est effectué. En cas de scores ex-æquo, la valeur du rang du terme permet de les départager.

Avec ce classement, la liste de gènes classés pour le terme 1 sera la suivante : le gène C, le gène B, le gène A le gène E et le gène D. Ainsi pour les gènes A et B d'une part et les gènes D et E d'autre part, pour lesquels le terme 1 apparait avec la même fréquence dans leur voisinage, j'ai utilisé le rang de ce terme dans le voisinage de ces gènes ex-æquo afin de pouvoir les classer.

2. Analyse de sensibilité des paramètres

J'ai considéré quatre paramètres pouvant influencer la valeur des scores (figure 26). Les différentes combinaisons de ces paramètres définissent une large collection de classifieurs. Afin d'évaluer l'impact de ces différents paramètres sur la classification des gènes et de considérer une collection de classifieurs pertinents, j'ai réalisé une analyse de sensibilité. J'ai tracé la courbe ROC pour chacun des classifieurs appliqués par terme sur l'ensemble des gènes du jeu de travail, et j'ai calculé leurs AUC. J'ai réalisé par la suite une analyse de variance ANOVA sur les AUC pour déterminer les paramètres ayant un impact important sur les classifieurs.

a. *Analyse ANOVA*

Les valeurs d'AUC obtenues tous termes et toutes ontologies confondus, sont extrêmement variables. J'ai réalisé une ANOVA pour déterminer quels sont les facteurs qui expliquent le mieux cette variabilité. Les facteurs pris en compte dans cette ANOVA sont ici les termes, le seuil de corégulation, le type de classement, le type de décompte et le type de graphe sachant que je n'ai pas considéré les interactions entre facteurs. L'analyse a montré que tous les facteurs sont significatifs (pvalues inférieures à $2e^{-16}$) sauf le type de graphe. Ce dernier facteur n'étant pas significatif, j'ai alors décidé de ne considérer dans la suite de ce chapitre que les classifieurs construits à partir des graphes binaires.

J'ai ensuite procédé à une autre ANOVA afin de déterminer s'il existe une influence significative de l'interaction des différents paramètres sur la performance des classifieurs. Dans cette deuxième analyse de variance, j'ai considéré comme facteurs explicatifs : les termes, le seuil de corégulation, le type de classement, le type de décompte et j'ai pris le modèle complet jusqu'aux interactions

d'ordre 4. Cette analyse montre que l'impact de toutes les interactions entre les facteurs est significatif. D'après la somme des carrés moyens, les deux interactions les plus importantes correspondent aux interactions entre termes et seuil de corégulation et entre termes et type de classement.

Tableau 15 : Résultats de l'analyse de la variance des AUC obtenues avec les différents classifieurs par une ANOVA avec interaction entre variables explicatives

Facteurs	Degrés de liberté	Somme des carrés moyens	Pvalue
Terme	41	1.097	$<2e^{-16}$
Seuil de coexpression	10	0.216	$<2e^{-16}$
Type de classement	2	0.104	$<2e^{-16}$
Type de décompte	2	0.147	$<2e^{-16}$
Terme: Seuil de coexpression	287	0.025	$<2e^{-16}$
Terme: Type de classement	82	0.036	$<2e^{-16}$
Seuil de coexpression: Type de classement	20	0.030	$<2e^{-16}$
Terme: Type de décompte	82	0.006	$<2e^{-16}$
Seuil de coexpression: Type de décompte	20	0.024	$<2e^{-16}$
Type de classement: Type de décompte	4	0.061	$<2e^{-16}$
Terme: Seuil: Type de classement	574	0.003	$<2e^{-16}$
Terme: Seuil: Type de décompte	574	0.001	$<2e^{-16}$
Terme: Type de classement: Type de décompte	164	0.001	$<2e^{-16}$
Seuil: Type de classement: Type de décompte	40	0.002	$<2e^{-16}$
Terme: Seuil: Type de classement: Type de décompte	1 148	0.0004	$<2e^{-16}$
Résidus	3 051	0.0000	

b. *Illustration de l'impact des paramètres sur la classification des gènes par terme*

La figure 27 représente une analyse de la distribution des valeurs d'AUC obtenues par terme. Cette figure illustre les résultats obtenus par l'ANOVA en particulier la variabilité des valeurs d'AUC entre les termes indiquant que la performance globale des différents classifieurs varie d'un terme à l'autre. La performance médiane dépassant la valeur de 0,5 pour tous les termes indique ainsi une meilleure performance qu'une classification aléatoire.

Parmi les 15 termes de l'ontologie CC, 11 ont une médiane supérieure ou égale à 0,6 dont deux dépassent 0,8 (« ribosome » avec une valeur AUC médiane à 0,892 et « plastid » à 0,813). Nous remarquons également que tous les termes de cette ontologie ont une valeur d'AUC maximale qui dépasse 0,6. Parmi ces termes, 6 ont la valeur maximale dépassant 0,8.

Pour les 13 termes étudiés de l'ontologie BP, 8 termes ont une médiane supérieure à 0,6 mais aucune médiane ne dépasse 0,8. Le terme pour lequel la valeur de la médiane est la maximale correspond au terme « electron transport or energy pathways » (0,774). Toutefois, tous les termes de cette ontologie ont des valeurs maximales dépassant 0,6 y compris 4 termes dont la valeur maximale dépasse 0,8 correspondant aux termes « response to stress », « electron transport or energy pathways », « protein metabolism » et « other metabolic processes ».

Concernant l'ontologie MF, seulement 3 termes ont une médiane supérieure à 0,6 qui sont « nucleotide binding », « transcription factor activity » et « structural molecule activity ». Les deux derniers termes ont des valeurs maximales supérieures à 0,8 et le terme « structural molecule activity » est le seul à avoir une médiane supérieure à 0,8. Au total 11 termes de cette ontologie ont des valeurs maximales qui dépassent 0,6.

Ces résultats montrent d'une part une différence de performance de classification en fonction de l'ontologie et des termes analysés, et d'autre part globalement une meilleure capacité des données de coexpression à classifier les termes BP et CC que les termes MF.

La variabilité des AUC de certains termes montre l'impact des combinaisons des paramètres utilisés sur la performance de classification des gènes. Pour certains termes, seuls quelques classifieurs atteignent une haute valeur d'AUC indiquant ainsi qu'ils sont les seuls à être capables de discriminer les exemples positifs et négatifs pour ces termes. Par exemple, pour le terme « transcription factor activity », la valeur minimale et la médiane sont respectivement à 0,559 et 0,683 alors que la valeur maximale d'AUC est égale à 0,884. Au contraire, pour d'autres termes, la variabilité est plus faible, ce qui indique que tous les classifieurs discriminent de la même manière les exemples positifs des exemples négatifs. Cela indique ainsi un faible impact des paramètres sur la performance de leur classification. Parmi ces termes, certains sont à une valeur élevée d'AUC tel que le terme ribosome, indiquant ainsi que les données de coexpression sont pertinentes et suffisantes pour leur prédiction. D'autres termes ont au contraire de faibles valeurs d'AUC tel que le terme « receptor binding activity ».

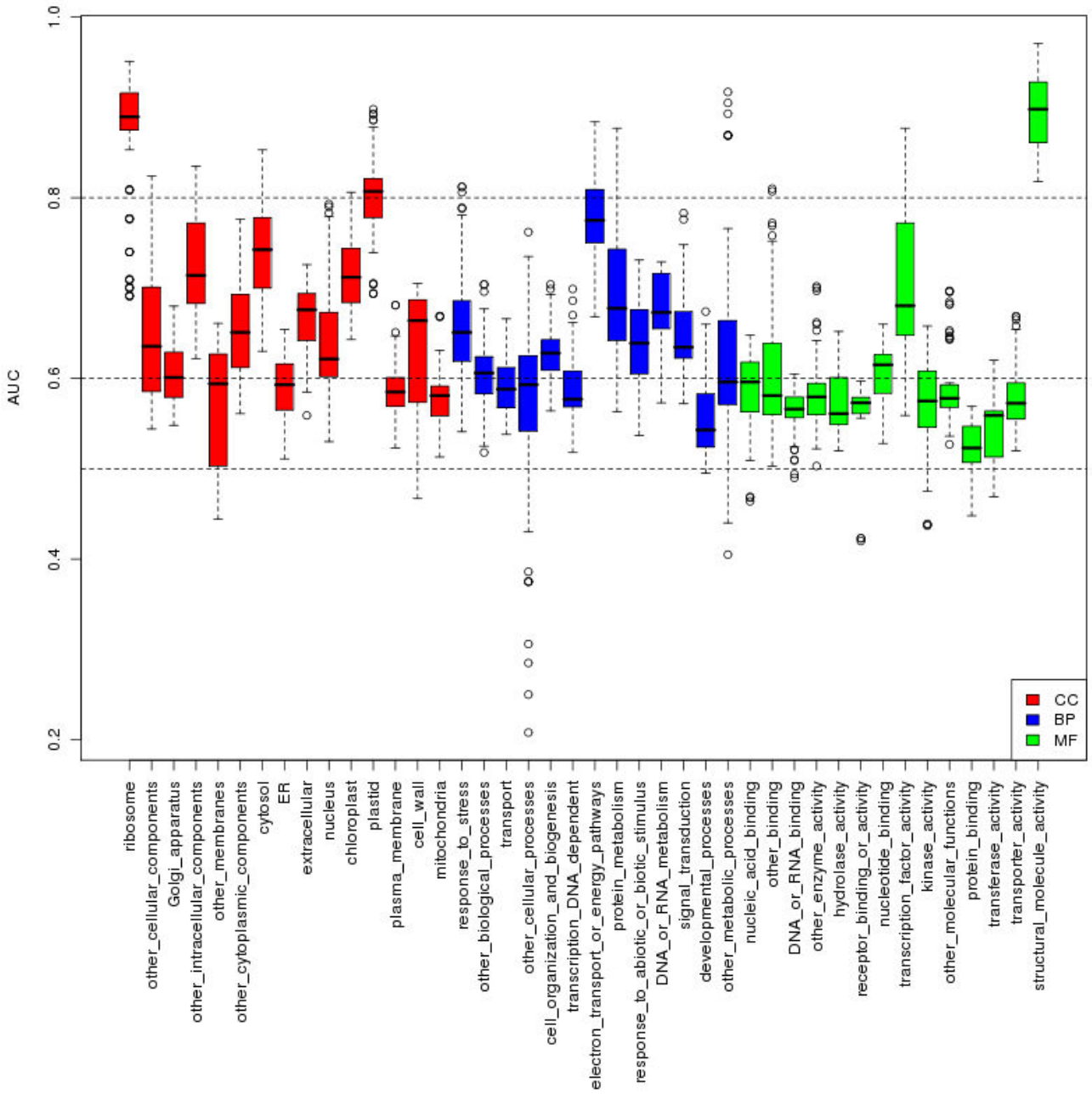


Figure 27 : Analyse des valeurs d'AUC obtenues avec les différents classifieurs par terme.

L'ensemble des valeurs AUC des différents classifieurs par terme sont représentées par une seule boite à moustaches. Le bas et le haut de la boite sont le 1er et le 3ème quartile et la ligne à l'intérieur correspond à la médiane. Les valeurs extrêmes sont représentées par un point. Les boites à moustaches rouges correspondent aux termes de l'ontologie CC, celles en bleu correspondent à l'ontologie BP et en vert représentent les termes de l'ontologie MF.

c. *Conclusion de cette analyse*

L'objectif de cette analyse était d'évaluer l'impact des différents paramètres considérés sur la discrimination des gènes par terme. L'analyse de la variance des AUC a montré qu'il était important de travailler au niveau des termes et que parmi les paramètres considérés, l'information de la valeur

de l'arête dans notre réseau de corégulation n'apporte pas plus d'information que son existence pour l'annotation fonctionnelle. J'ai alors décidé pour la suite des analyses de fixer ce paramètre et de ne considérer que les graphes binaires. Ce choix a permis de réduire de moitié la taille de la collection des classifieurs considérés.

Cette analyse a révélé également l'existence des interactions entre les paramètres. Les interactions les plus significatives entre ces variables explicatives correspondent aux interactions entre les termes et le seuil de coexpression ainsi qu'entre les termes et le type de classement.

Les boites à moustaches illustrent les résultats de l'analyse par ANOVA notamment l'importance de la prédiction par ontologie et par terme. Elle a permis de mettre en évidence les différences de classement entre les termes. Pour certains termes, les données de coexpression sont suffisantes pour leur classification à condition d'une utilisation pertinente de ces données qui passe par le choix de la bonne combinaison des paramètres. Pour d'autres termes, les données de coexpression seules semblent être insuffisantes pour discriminer les exemples positifs et négatifs quels que soient les classifieurs appliqués.

IV. Règles de décision

1. Définition des règles

Idéalement, tous les gènes annotés pour le terme étudié sont associés à des scores élevés. Par conséquent, l'idée pour définir une règle de décision à partir d'un classifieur est de déterminer un score seuil : si le score d'un gène est supérieur à ce score seuil, le terme est attribué au gène et si le score est inférieur, le terme ne lui est pas attribué.

Pour la détermination du score seuil associé à chaque classifieur, j'ai décidé de le définir de manière à ce que la liste de gènes ayant un score supérieur à ce score seuil comporte au maximum 20% de faux-positifs. Cette assurance de qualité est importante car elle garantit aux biologistes de limiter leurs efforts de validation expérimentale qui sont souvent coûteux en temps et en ressources. Cette valeur de FDR autorisée à 20% maximum peut paraître élevée, mais en l'état de l'annotation fonctionnelle actuelle, il peut être considéré comme étant un taux très stringent. En effet, Wang *et al.* (2013) montrent que le FDR de plusieurs méthodes d'annotation actuelles varie entre 85% et 46%, en particulier celui de l'approche du vote majoritaire est de 70%.

L'objectif est de déterminer le score seuil de chaque règle et de sélectionner la meilleure règle par terme. Pour cela j'ai évalué la performance de prédiction de chacune de ces règles à l'aide d'une procédure de validation croisée.

2. Procédure d'évaluation par validation croisée

La procédure d'évaluation que j'ai mise en place est une validation croisée à 5-blocs que j'ai répétée 20 fois. La procédure d'évaluation étant la même pour tous les classifieurs, la description est faite pour un classifieur appliqué à un terme d'une ontologie.

Pour chaque jeu d'apprentissage, le classifieur fournit les scores des gènes et la liste de classement. Le score seuil est ensuite déterminé de manière à ce que le FDR dans cette liste soit au maximum égal à 20 %. Ceci permet de définir une règle de décision qu'on évalue dans un second temps sur les gènes du jeu test en comparant leurs prédictions avec les termes connus pour ces gènes, ce qui permet de calculer le nombre de VP, FP, VN et FN ainsi que le Fmeas et le FDR. Cette procédure appliquée à l'ensemble des jeux test, permet de calculer une moyenne de ces métriques. Notons que dans cette analyse, le FDR permet de définir la règle de décision. Le fait qu'il soit fixé dans le jeu d'apprentissage et calculé sur le jeu test nous permet également de mesurer à quel point il est possible de bien le contrôler.

3. Résultats

Le bilan de la procédure de prédiction par terme est analysé en fonction des métriques calculées (FDR et Fmeas) et des règles considérées lors de la validation croisée.

a. *Facilité d'obtention de score seuil par règle et par terme*

Avec certaines règles et pour certains jeux d'apprentissage, il n'a pas été possible de définir un score seuil car le FDR était toujours supérieur à 20%. Dans ce cas, la prédiction et le calcul de la performance sur le jeu test correspondant ne sont pas effectués et la moyenne du Fmeas et du FDR est calculée uniquement à partir des jeux test évalués. La figure 28 représente la variabilité du nombre de ces jeux d'apprentissage non exploitables par ontologie et par terme, toutes règles confondues. Cette figure montre que seulement trois termes ont un nombre de jeux d'apprentissage non exploités égal à 0 indiquant que pour tous les jeux d'apprentissage un score seuil a pu être défini par toutes les règles; trois autres termes ont une médiane égale à 0 indiquant que pour 50% des

règles, la définition du score seuil a été possible pour tous les jeux d'apprentissage. Au total 11 termes ont une médiane comprise entre 0 et 20. Ils correspondent à : « ribosome », « other intracellular components », « other cytoplasmic components », « cytosol » et « chloroplast » de l'ontologie CC, « response to stress », « other cellular processes », « protein metabolism », « response to abiotic or biotic stimulus » et « other metabolic processes » de l'ontologie BP et « structural molecule activity » de l'ontologie MF. Les 31 autres termes ont une médiane supérieure à 20, voire supérieure à 80 pour 10 termes d'entre eux. Cela signifie que pour la majorité des règles il n'a pas été possible de fixer le score seuil pour avoir un FDR inférieur ou égal à 20% dans de nombreux jeux d'apprentissage. En particulier, pour le terme « receptor binding or activity », aucun score seuil n'a pu être défini. Ces résultats indiquent la difficulté de fixer un FDR inférieur ou égal à 20% et que cette difficulté est plus ou moins grande en fonction du terme analysé. A l'issue de cette analyse, parmi la collection de 3 051 classifieurs, 386 règles de décision ont été éliminées puisqu'elles ne permettent de définir aucun score seuil. Le terme « receptor binding or activity » est aussi écarté de la procédure de prédiction puisqu'aucune règle ne permet de lui définir un score seuil dans tous les jeux d'apprentissage créés.

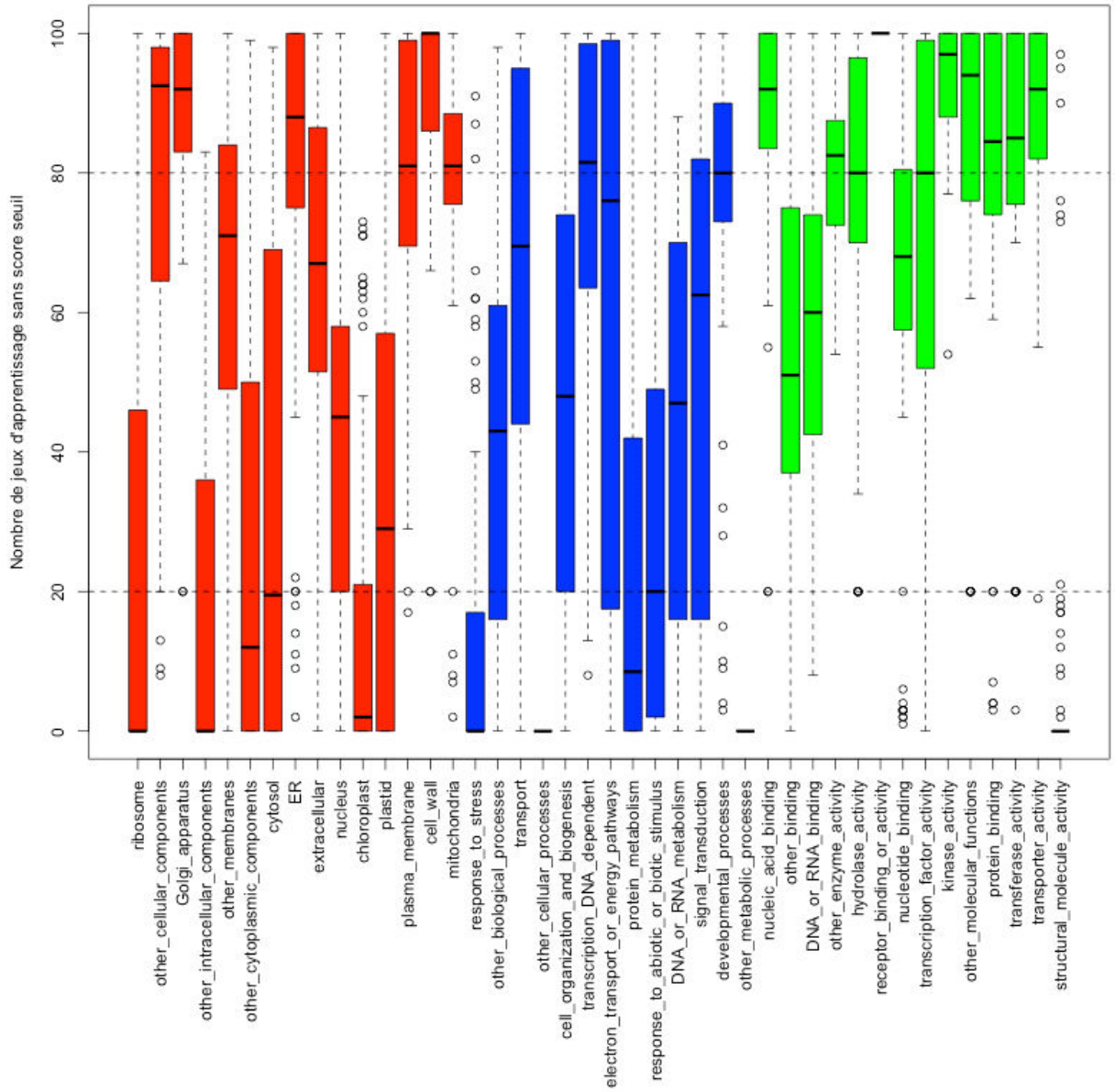


Figure 28 : Représentation du nombre de jeux d'apprentissage non exploitables pour définir un score seuil par terme parmi les 100 jeux créés par la validation croisée (toutes règles confondues).
Chaque boite à moustaches représente les valeurs obtenues avec l'ensemble des règles de décision appliquées à un terme.

b. *Stabilité des valeurs de FDR*

Les scores seuils étant déterminés dans les jeux d'apprentissage pour contrôler le FDR à une certaine valeur inférieure ou égale à 20%, nous nous attendions à ce que les FDR mesurés dans les jeux test soient proches de cette valeur. Afin de vérifier cela, j'ai représenté les valeurs de FDR mesurées dans les jeux test en fonction des valeurs de FDR ayant servi pour définir les scores seuils dans les

jeux d'apprentissage pour les 2 665 règles évaluées (figure 29). Cette figure montre une instabilité des valeurs de FDR pour la majorité des règles puisque seulement quelques règles sont situées sur la première bissectrice. Cela montre que le score seuil fixé sur un jeu d'apprentissage ne garantit pas un contrôle du FDR sur un autre jeu de données. Cette figure montre également des valeurs de FDR obtenues dans les jeux d'apprentissage égales à 0 associées à des valeurs de FDR mesurées dans les jeux test allant de 0 à 1. Cela concerne exactement 888 règles pour lesquelles la définition du score seuil est difficile car le FDR dans cette liste passe de 0 pour les classifications correctes à des valeurs dépassant directement 20 % dès la première fausse classification. La méthode fixe alors le score seuil correspondant à un FDR égal à 0 ce qui génère beaucoup d'instabilité dans les valeurs de FDR mesurées sur les jeux test. En effet, parmi ces 888 règles ayant un FDR à 0 dans le jeu d'apprentissage, 186 génèrent un FDR égal à 1 dans le jeu test, 54 avec un FDR égal à 0 et 329 avec un FDR dans un intervalle de [0,8 : 1[. Contrairement aux 386 règles éliminées dans le paragraphe précédent, ces 888 règles permettent d'identifier un score seuil mais ce dernier est totalement inutilisable pour permettre un contrôle du FDR.

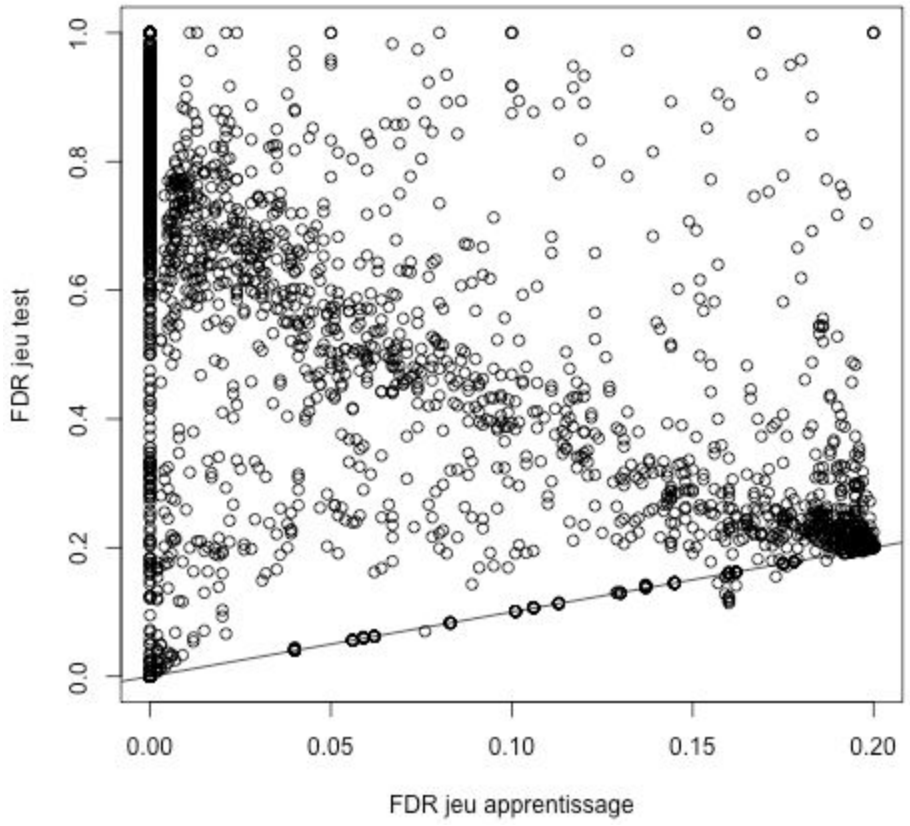


Figure 29 : Représentation des valeurs de FDR mesurées dans les jeux test en fonction des valeurs de FDR associées aux scores seuils dans les jeux d'apprentissage.

Chaque point correspond à la valeur de FDR pour une règle de décision appliquée à un terme. La droite tracée représente la première bissectrice.

c. *Facteurs impactant les valeurs de FDR*

i. Valeurs de FDR en fonction des termes

J'ai représenté les valeurs de FDR mesurées sur les jeux test par terme toutes règles confondues parmi les 2 665 retenues (figure 30). Cette figure montre une variabilité importante des valeurs de FDR obtenues par les différentes règles ainsi qu'une variabilité importante entre les termes. Les valeurs de FDR correspondent à des valeurs élevées dépassant très souvent 0,2 pour la majorité des règles et des termes sauf pour « other cellular processes » et « other metabolic processes ». Cette figure montre également la différence de capacité des règles de décision à contrôler le FDR en fonction de l'ontologie analysée. En effet, 10 termes de l'ontologie CC et 11 termes de l'ontologie BP ont au moins une règle de décision donnant une valeur de FDR inférieure ou égale à 0,2. Cependant, seulement 4 termes de l'ontologie MF ont une valeur minimale inférieure ou égale à 0,2. Globalement, les règles de décision ont plus de facilité à contrôler le FDR pour les termes des ontologies BP et CC que ceux de l'ontologie MF. Notons que le terme avec le FDR le mieux contrôlé de l'ontologie MF, est « structural molecule activity » qui a une médiane de 0,212 et une valeur minimale de 0,190. L'exception de ce terme s'explique par le fait qu'il soit associé aux protéines ribosomales souvent coexprimées et interagissant ensemble pour former des complexes protéiques. Lors du chapitre 2, j'avais déjà montré que ces gènes sont corégulés et constituent une composante connexe très homogène et très spécifique.

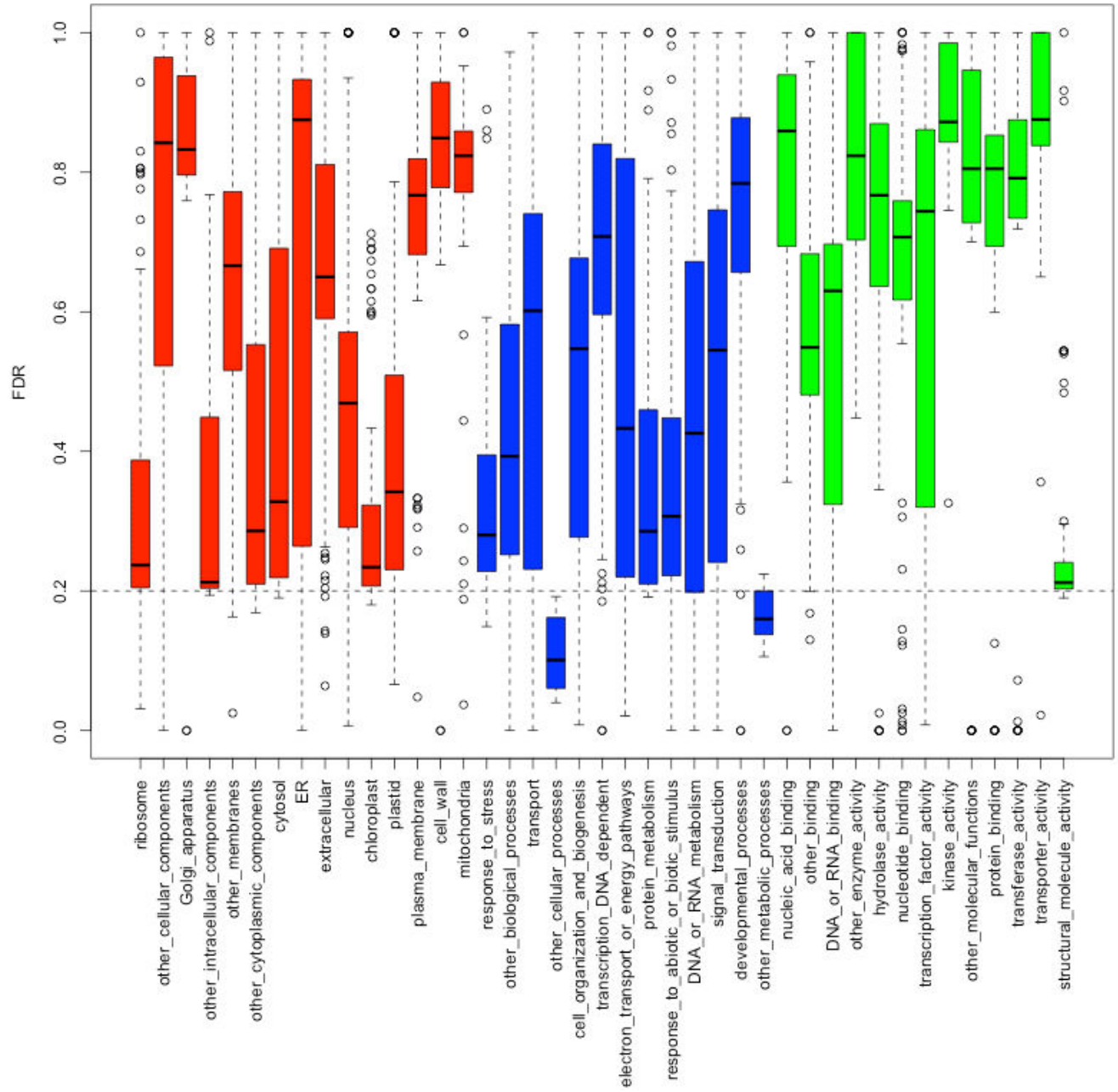


Figure 30 : Analyse des valeurs de FDR mesurées sur les jeux test avec l'ensemble des règles de décision par terme.

ii. Relation entre FDR et nombre de gènes positifs par terme

Le FDR est un rapport des faux positifs parmi tous les gènes prédits positifs. Cette métrique peut donc être naturellement sensible au nombre de gènes annotés présents dans l'échantillon. Pour vérifier cette relation, j'ai représenté les valeurs de FDR mesurées dans les jeux test en fonction du nombre de gènes annotés par le terme étudié pour chacune des règles dans le jeu test (figure 31). Cette figure montre que plus le nombre de gènes annotés par le terme augmente mieux le FDR est

contrôlé. Nous constatons une instabilité des valeurs de FDR pour les termes ayant un nombre de gènes annotés inférieur à 200 et une stabilité des valeurs FDR pour les termes ayant un grand nombre de gènes annotés dans l'échantillon. Il est plus facile ainsi de contrôler le FDR quand plus de 200 gènes sont annotés avec le terme considéré. Pour les termes auxquels très peu de gènes sont associés, la moindre erreur de classification fait augmenter rapidement le FDR. Notons que malgré la condition mise au début de l'analyse, qui consiste à étudier uniquement les termes représentés par au moins 20 gènes dans le jeu de travail, plusieurs jeux d'apprentissage ont un nombre de gènes annotés égal ou proche de 0. Cela est dû à la procédure de validation croisée qui découpe aléatoirement le jeu de travail en jeu d'apprentissage et jeu test sans tenir compte du déséquilibre entre les exemples positifs et négatifs. Dans ces sous-échantillons, le nombre minimal de gènes positifs par terme à 20 n'est plus garanti.

La figure 30 nous a montré que les deux termes dont le FDR est contrôlé à un seuil inférieur à 20% dans les jeux test quels que soient les règles appliquées, correspondent à « other cellular processes » et « other metabolic processes ». A partir de la figure 31, j'ai identifié les règles ayant un FDR stable et associées à un nombre élevé de gènes annotés par le terme. Parmi ces règles, je retrouve toutes celles associées à ces deux termes. En effet, ce sont les termes ayant le plus grand nombre de gènes positifs dans le réseau quels que soient les seuils de corégulation considérés. Pour ces deux termes, le FDR est ainsi contrôlé même pour les règles appliquées à des seuils de corégulation élevés et pour lesquels le nombre de gènes positifs diminue. Cependant, d'autres termes ayant un nombre de gènes positifs élevé n'ont pas un FDR contrôlé à 0,2. Le nombre de gènes positifs de chaque terme a donc un impact sur la stabilité du FDR mais n'explique pas tout pour autant.

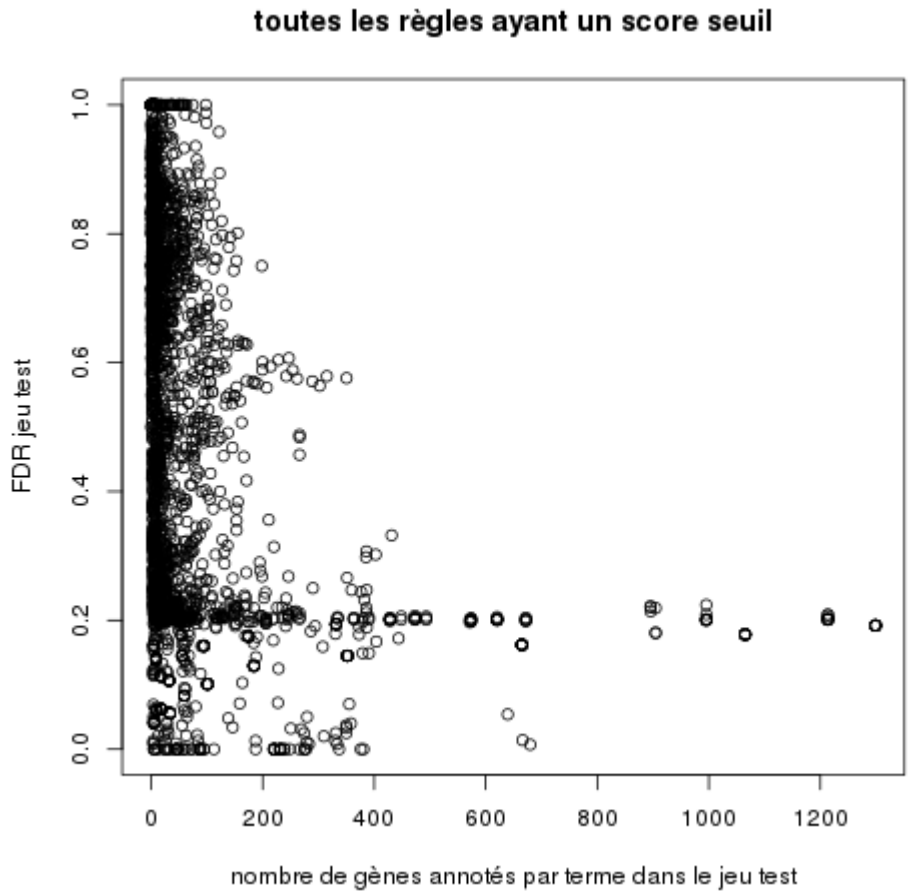


Figure 31 : Représentation des valeurs de FDR mesurées dans les jeux test en fonction du nombre de gènes annotés par le terme étudié par chacune des règles dans les jeux test.
Chaque point correspond à la valeur FDR obtenue par une règle parmi les 2 665 règles évaluées.

iii. Relation entre FDR et représentativité des termes

Afin de mieux comprendre les raisons expliquant l'instabilité du FDR, j'ai représenté ses valeurs en fonction de la représentativité des termes dans les échantillons. Cette représentativité est définie par le nombre de gènes annotés par ce terme dans l'échantillon divisé par la taille de l'échantillon. La figure 32 représente la distribution des valeurs de FDR définies dans les jeux d'apprentissage en fonction de la représentativité des termes dans ces mêmes jeux. La valeur du FDR dans les jeux d'apprentissage est très variable pour les termes représentés moins de 60% dans ces échantillons. Pour les termes représentés moins de 40%, les valeurs de FDR peuvent être égales à 0. Il s'agit des 888 règles identifiées précédemment ayant un FDR du jeu d'apprentissage égal à 0. Ces cas traduisent la difficulté des classifieurs à définir un seuil score dans les listes étudiées et traduisent probablement la stringence du score seuil de FDR à 20% pour les termes représentés moins de 40%. Pour les termes très représentés (>80%) les valeurs de FDR sont plus stables et inversement corrélés

à leur représentativité. En effet, les termes représentés à hauteur de 90% ont un FDR qui ne peut pas dépasser 10%.

La figure 33 représente la distribution des valeurs de FDR mesurées dans les jeux test en fonction de la représentativité des termes dans ces mêmes jeux. La valeur du FDR mesurées dans les jeux test prennent des valeurs très différentes pour les termes peu représentés (<40 %). Pour ces termes, 201 règles ont des valeurs de FDR égales à 1. Ces règles sont associées principalement à un FDR seuil égal à 0 (c'est le cas pour 186 règles). Ce résultat montre que pour ces termes peu représentés, leurs seuils scores identifiés par la majorité des règles sont insuffisants pour la prédiction et génèrent des valeurs de FDR élevées. Cette figure confirme aussi la corrélation entre représentativité des termes et FDR calculés. Ainsi les termes représentés à hauteur de 40 % peuvent avoir des valeurs de FDR allant jusqu'à 60%, et ceux représentés à hauteur de 80% ont des FDR à 20%.

Par ailleurs, ces deux figures montrent que la représentativité des termes dans les jeux test et les jeux d'apprentissage est homogène et que l'impact de cette représentativité sur les valeurs FDR dans le jeu test et le jeu d'apprentissage est globalement le même.

En conclusion, d'après ces analyses, le FDR est sensible à la taille du jeu de données en particulier à la composition de ce jeu en exemples positifs et négatifs. De plus, le FDR est sensible à la représentativité des termes dans l'échantillon ce qui explique pourquoi à un même nombre de gènes positifs les règles contrôlent différemment le FDR des termes.

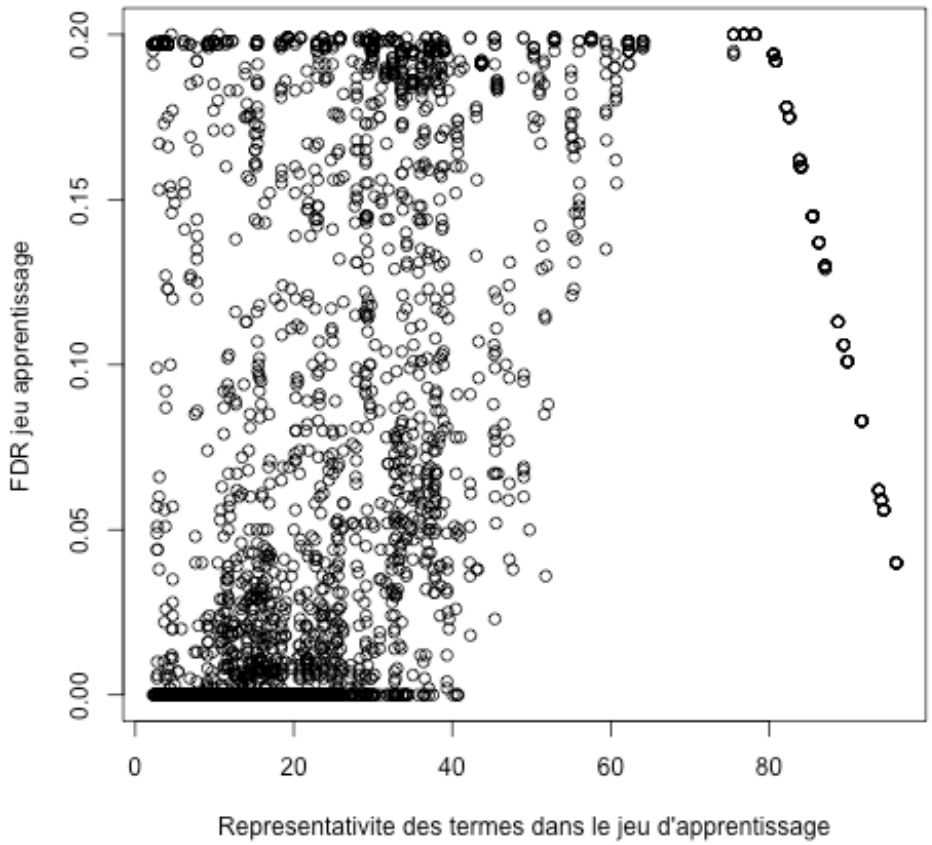


Figure 32 : Représentation des valeurs de FDR déterminées dans les jeux d'apprentissage en fonction de la représentativité des termes dans les jeux d'apprentissage.

Chaque point correspond à une règle parmi les règles ayant défini un score seuil. L'axe des abscisses correspond à la représentativité des termes exprimée en pourcentage.

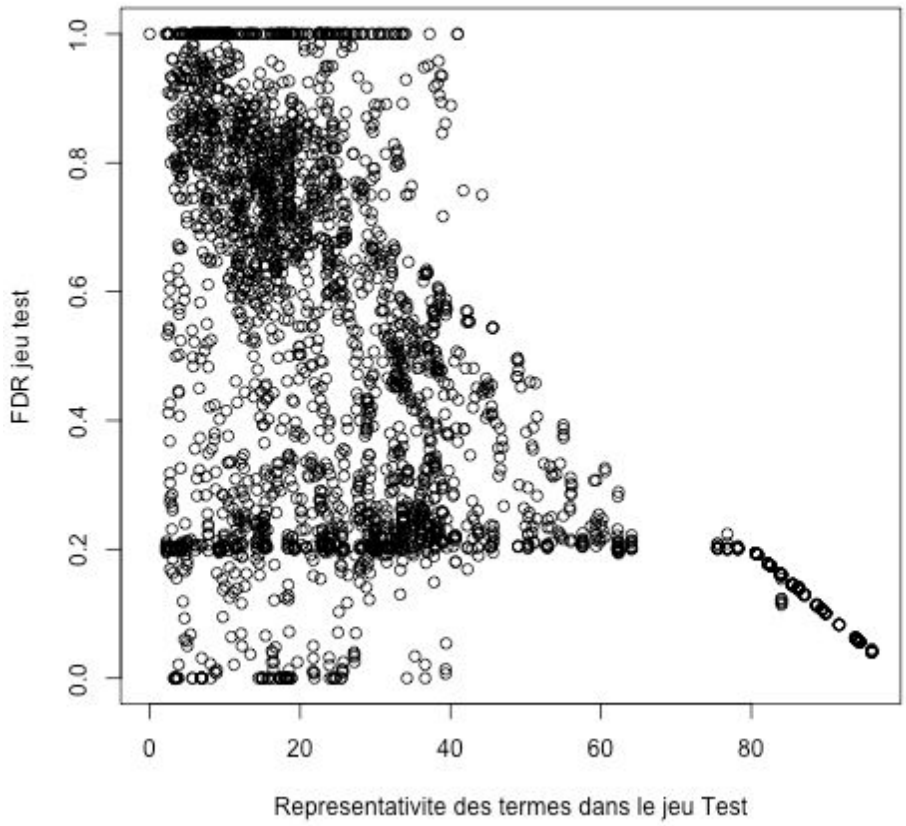


Figure 33 : Représentation des valeurs de FDR mesurées dans les jeux test en fonction de la représentativité des termes dans les jeux test.

Chaque point correspond à une règle parmi les règles ayant défini un score seuil. L'axe des abscisses correspond à la représentativité des termes exprimée en pourcentage.

d. *Sélection des règles contrôlant le FDR*

Afin d'identifier les règles qui permettent de contrôler au mieux le FDR, j'ai par la suite sélectionné uniquement les règles qui vérifient que les valeurs de FDR mesurées dans les jeux test soient égales aux valeurs de FDR obtenues dans les jeux d'apprentissage plus ou moins 0,02. J'ai également mis la contrainte que le FDR du jeu d'apprentissage doit être supérieur à 0,05 afin de ne pas sélectionner les règles qui identifient mal le seuil et s'arrêtent avec les premières erreurs de classification même si elles sont stables entre le jeu d'apprentissage et le jeu test. Au total 430 règles appliquées à 16 termes satisfaisant ces contraintes ont été retenues. Ces termes correspondent principalement à des termes de l'ontologie BP et l'ontologie CC: 8 termes parmi les15 termes de l'ontologie CC, 6 termes parmi les 13 termes de l'ontologie BP et 2 termes parmi les 14 termes de l'ontologie MF. Ce résultat est détaillé dans le tableau 16 avec le nombre de règles retenues par terme.

Cette sélection sur la base du FDR a permis de réduire la taille de la collection des règles de décision considérée mais plusieurs règles par terme sont encore retenues. L'objectif est donc de sélectionner encore parmi ces règles la meilleure règle par terme sur la base d'autres critères.

Tableau 16 : Termes retenus avec la sélection des règles ayant un FDR stable.
Les colonnes #règles représentent le nombre de règles satisfaisant ces critères par terme et par ontologie.

CC	#règles	BP	#règles	MF	#règles
ribosome	33	Response to stress	7	Structural molecule activity	*48*
chloroplast	22	Other biological processes	1	Nucleotide binding	1
cytosol	19	Other metabolic processes	93		
plastid	13	Other cellular processes	90		
Other cytoplasmic components	25	Response to abiotic or biotic stimulus	*6*		
nucleus	2	Protein metabolism	*22*		
ER	1				
Other intracellular components	47				

e. *Performance de prédiction en fonction du Fmeas des règles sélectionnées*

i. Analyse des valeurs de FDR en fonction des Fmeas

J'ai considéré le FDR dans cette analyse comme le moyen pour définir un seuil dans la liste de classement des gènes. Je l'ai également considéré comme le premier filtre pour la sélection des règles, puisque toutes celles qui ne permettent pas d'obtenir un FDR inférieur ou égal à 20% dans les listes d'apprentissage ainsi que toutes celles qui ne permettent pas de contrôler le FDR dans les jeux test dans une fenêtre de décalage de 0,02 ont été éliminées. Cependant, plusieurs règles par terme vérifient encore ces critères. Afin de les départager je m'appuie sur le Fmeas. En annotation fonctionnelle cette métrique prend des valeurs relativement basses: Fmeas de 0,07 avec la méthode de chi deux (Hishigaki *et al*. 2001); Fmeas de 0,33 avec le vote majoritaire de (Schwikowski *et al*. 2000) et Fmeas de 0,44 avec la méthode fondée sur les noyaux de Wang et ses collaborateurs (2013). Dans notre analyse, le Fmeas n'a pas été contrôlé contrairement au FDR. Il est donc intéressant de comparer la performance des règles sélectionnées en fonction de cette métrique. La figure 34

représente les valeurs de FDR mesurées dans les jeux test en fonction de leurs valeurs de Fmeas pour les 430 règles sélectionnées. Cette figure montre que pour les règles ayant un FDR contrôlé autour de 20%, leurs valeurs de Fmeas sont comprises entre 0 et 0.8. La majorité des règles ayant un FDR inférieur à 20% ont des valeurs de Fmeas supérieures à 0,9 mais que certaines règles ont aussi des faibles valeurs de Fmeas (inférieures à 0,02). Cela montre l'importance de sélectionner la règle la plus adaptée pour la prédiction de chaque terme en fonction des deux métriques. Etant donné que lorsque le FDR est contrôlé cela garantit une bonne précision et sachant que le Fmeas est un compromis entre la sensibilité et la précision, ce résultat indique que pour ces règles contrôlant la précision, la sensibilité est de toute façon faible.

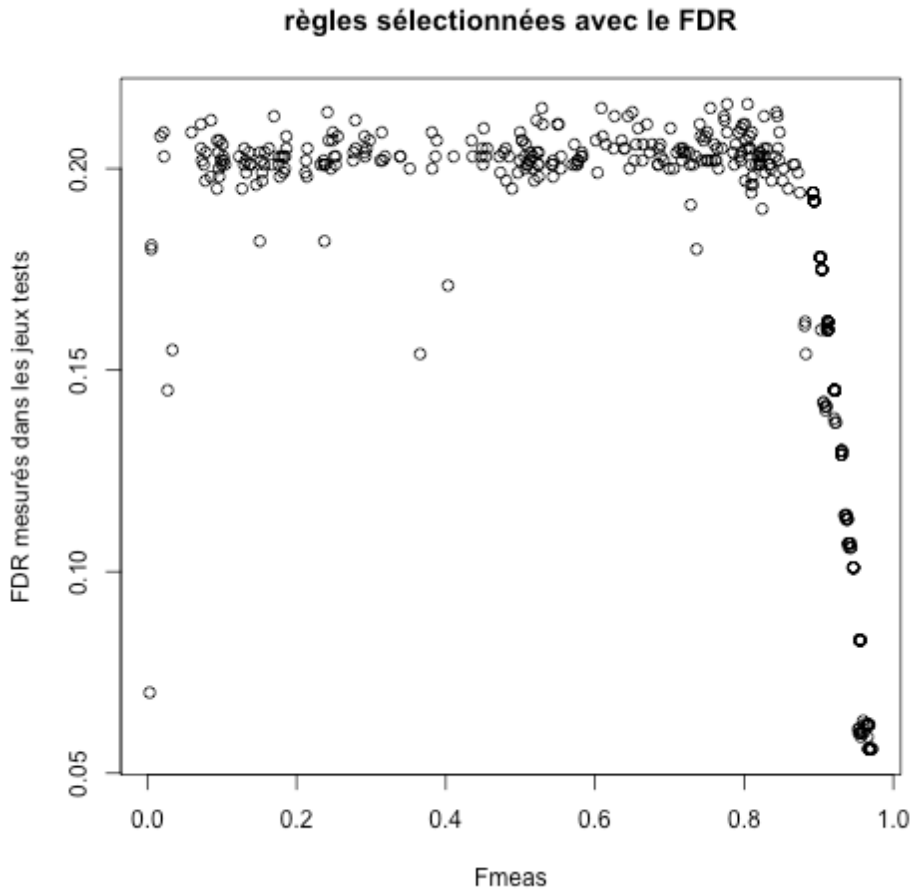


Figure 34 : Représentation des valeurs de FDR mesurées dans les jeux d'apprentissage en fonction de leurs valeurs de Fmeas (pour les 430 règles contrôlant le FDR).

ii. Sélection des règles ayant le meilleur Fmeas par terme

Pour sélectionner la meilleure règle par terme, j'ai considéré celle qui a la plus grande valeur de Fmeas. Dans la figure 35 j'ai représenté les valeurs de FDR de ces 16 règles en fonction de leurs valeurs de Fmeas (partie gauche de la figure 35), ainsi que la boite à moustache des valeurs de

Fmeas (partie droite de la figure 35). La boite à moustache indique que la médiane des valeurs de Fmeas pour ces 16 règles est de 0,750, que la valeur maximale est de 0,971 et la valeur minimale est de 0,003. La partie gauche de la figure montre qu'effectivement 12 règles sélectionnées ont un Fmeas supérieur à 0,2 et 4 règles ont un Fmeas quasiment nul. Ces 4 règles sont associées aux termes « nucleus », « ER », « nucleotide binding » et « other biological processes » qui ne peuvent pas être prédits avec une bonne sensibilité.

La figure 36 représente les valeurs de Fmeas de ces 16 règles en fonction du paramètre seuil de corégulation. Cette figure montre que la performance de prédiction des règles augmente avec le seuil de corégulation. Elle montre en particulier que les 12 règles ayant un Fmeas supérieur à 0,2 sont obtenues à des seuils de corégulation supérieurs à 6 ce qui montre la pertinence du réseau de corégulation et l'importance de considérer ce seuil de corégulation pour l'identification de liens fonctionnels. Cependant, à ces seuils de corégulation, le nombre de gènes orphelins ou mal caractérisés à prédire avec ces règles est très faible: il varie de 6 à 116 gènes au maximum. Il est donc dommage de se limiter dans le cadre d'une annotation fonctionnelle à haut débit aux réseaux de corégulation les plus petits dont le potentiel en terme de nombre de gènes à annoter est limité.

En conclusion, cette analyse nous montre que la sélection des règles sur la base du meilleur Fmeas ne garantit pas une bonne sensibilité pour 4 termes. Pour les 12 autres termes, cette sélection de règles en fonction du meilleur Fmeas limite le potentiel de la méthode à éclairer la fonction d'un plus grand nombre de gènes inconnus.

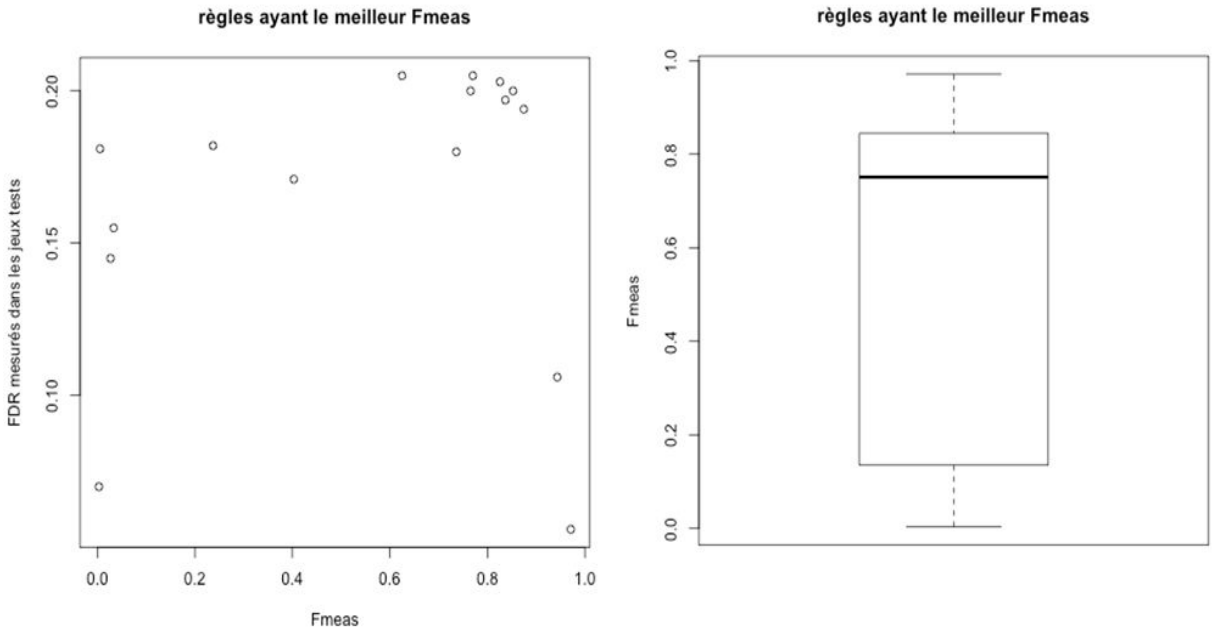


Figure 35 : Analyse des règles ayant le meilleur Fmeas par terme.

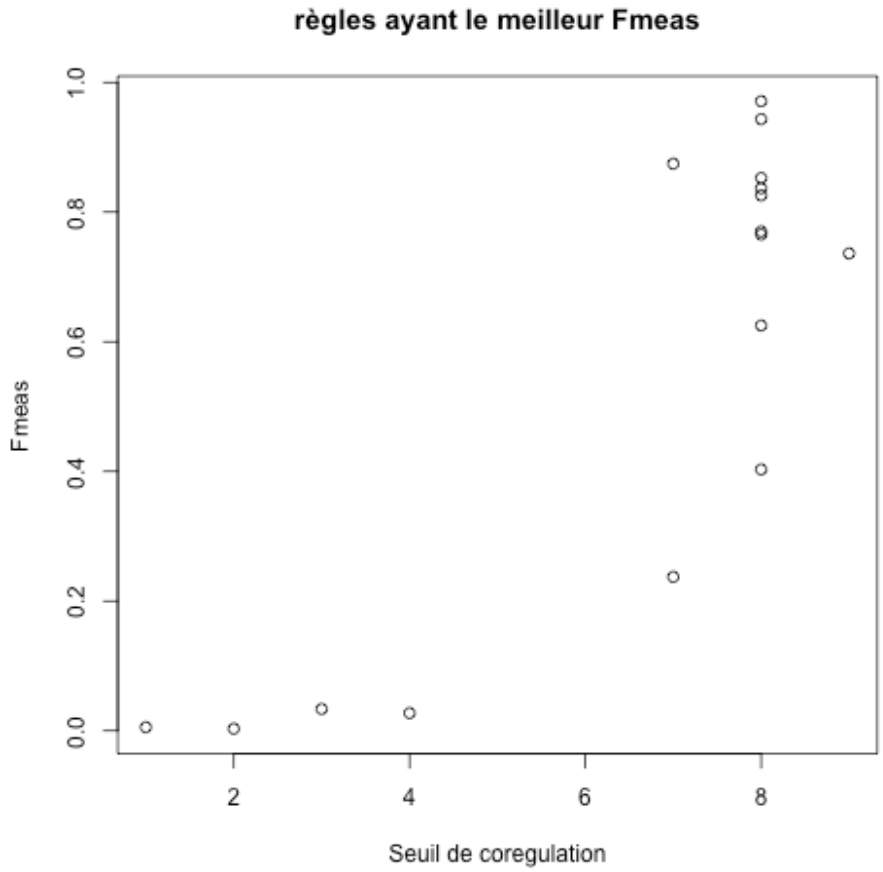


Figure 36 : Représentation des valeurs de Fmeas en fonction des seuils de corégulation correspondants (pour les règles ayant le meilleur Fmeas par terme).

f. *Meilleure règle de décision par terme*

A l'issue des analyses précédentes, j'ai décidé de ne pas prédire les termes ayant un Fmeas inférieur à 0,2 afin de garantir un bon compromis entre sensibilité et précision des termes prédits avec cette méthode. Cela fait passer le nombre de règles contrôlant le FDR de 430 règles à 371 règles appliquées à 12 termes. J'ai également décidé de ne pas choisir la règle ayant le meilleur Fmeas parmi ces règles sélectionnées afin de ne pas limiter le nombre de gènes à annoter avec la méthode. En effet, j'ai opté pour un compromis entre la performance des règles retenues mesurée avec le Fmeas et le nombre de gènes à prédire avec ces règles. Ce compromis est trouvé en maximisant le nombre de gènes annotés prédits positifs. Cela garantit alors de nous placer au seuil de corégulation ayant les meilleures performances et permettant de prédire le maximum de gènes inconnus. Selon ces critères, la meilleure règle par terme est sélectionnée. Les résultats sont résumés dans le tableau suivant.

Tableau 17 : Résultats de la sélection de la meilleure règle par terme et performance correspondante.

Ontologie	Terme	Seuil de coregulation	A predire	Type de classement	Type de décompte	Fmeas	FDR
CC	ribosome	2	891	IndépendantAmélioré	Occurence	0,554	0,2
CC	other_intracellular_components	3	455	Indépendant	Occurence	0,252	0,203
CC	other_cytoplasmic_components	4	224	IndépendantAmélioré	Occurence	0,312	0,202
CC	cytosol	2	891	Indépendant	Occurence	0,213	0,198
CC	chloroplast	2	891	Dépendant	Occurence	0,315	0,202
CC	plastid	3	455	IndépendantAmélioré	Occurence	0,339	0,203
BP	response_to_stress	7	91	IndépendantAmélioré	Occurence	0,351	0,2
BP	other_cellular_processes	1	4337	IndependantAmélioré	Occurence	0,893	0,192
BP	protein_metabolism	4	757	Indépendant	Occurence	0,293	0,205
BP	response_to_abiotic_or_biotic_stimulus	7	91	IndépendantAmélioré	Occurence	0,237	0,182
BP	other_metabolic_processes	2	3135	IndépendantAmélioré	Frequence	0,684	0,201
MF	structural_molecule_activity	3	1848	Dépendant	Occurence	0,614	0,206

Ce tableau montre qu'avec cette sélection, la moyenne des valeurs de Fmeas des 12 règles sélectionnées est de 0,421 et que 7 des règles de décision retenues parmi les 12 correspondent à des seuils de corégulation supérieurs à 3, valeur à partir de laquelle le réseau de corégulation est différent d'un réseau aléatoire. Les cinq termes qui peuvent être analysés à des faibles seuils de corégulation sont « other metabolic processes » et « other cellular processes » qui sont les deux termes les plus représentés dans le réseau pour l'ontologie BP ainsi que les termes « ribosome », « cytosol » et « chloroplast » de l'ontologie CC. Cela indique que ces cinq termes sont facilement prédictibles. Les autres termes sont prédictibles à des seuils de corégulation plus élevés montrant ainsi l'importance de l'analyse de corégulation visant à identifier des liens fonctionnels plus forts entre les gènes. Notons que les termes « response to stress » et « response to abiotic or biotic stimulus » sont prédictibles au seuil 7 auquel des composantes connexes ont été identifiées dans le chapitre2.

Concernant les autres paramètres, la majorité des règles sélectionnées correspondent aux types de classement qui prennent en considération le rang des autres termes dans le voisinage d'un gène analysé pour le calcul de son score: 7 règles avec un classement « IndépendantAmélioré » et 2 règles avec un classement « Dépendant ». Ainsi seulement 3 règles sélectionnées ont un type de classement totalement indépendant des autres termes.

Ce tableau montre également que la performance totale de ces termes prédictibles est de 0,421 en Fmeas et un FDR de 0,199 soit une précision de 0.801. Ces résultats présentent ainsi une amélioration de performance par rapport à celles des méthodes comparées par Wang al (2013).

Par ailleurs, ce résultat montre que notre méthode est capable de faire de la prédiction pour 12 termes dont 8 sont spécifiques et 4 non spécifiques à savoir les termes « other ». Les termes non spécifiques sont probablement moins intéressants pour les biologistes mais donnent toutefois l'information que les gènes qui leur sont associés ne sont vraisemblablement pas annotés par les autres catégories d'annotation plus spécifiques. Parmi les termes spécifiques qui ont été retenus dans cette analyse de prédiction, nous retrouvons ceux reliés au stress. Ce résultat non surprenant est toutefois réconfortant sur la spécificité des processus mis en évidence par les analyses de coexpression et de corégulation.

V. Application des règles sélectionnées pour l'inférence de fonction aux gènes mal caractérisés

1. Procédure

La meilleure règle de décision sélectionnée pour les 12 termes prédictibles est appliquée pour l'annotation des gènes orphelins ou mal caractérisés (effectifs donnés dans le tableau 17). Notre objectif est de prédire si ces gènes peuvent être caractérisés par les 12 termes retenus et d'associer un indice de confiance à ce résultat.

Pour la prédiction, j'utilise les 100 jeux d'apprentissage. Pour chacun et pour chaque gène à annoter, je calcule le score que je compare au score seuil du jeu, Si le score est supérieur au score seuil, alors le gène est prédit positif, sinon il est prédit négatif.

2. Résultats

a. *Gènes prédits positifs par terme*

Je me suis intéressée ici aux gènes qui sont prédits positifs pour un terme dans au moins 80 % des cas (80 fois). Le tableau 18 présente les résultats pour les 12 termes. Le tableau 19 indique le nombre de prédictions positives effectuées et le nombre de gènes annotés pour les 12 termes et pour 8 termes spécifiques n'ayant pas le mot « other ». La liste des gènes prédits positifs pour chacun des termes spécifiques retenus est fournie dans l'annexe F. Le tableau 19 montre que pour les 12 termes, l'approche a déclaré 5 796 prédictions positives dans au moins 80 % des cas et cela permet de caractériser 4 348 gènes. Parmi ces derniers, 4 345 sont prédits positifs 100 fois. Si nous nous

intéressons uniquement aux 8 termes spécifiques, 47 gènes sont prédits positifs dans au moins 80 % des cas et 31 sont tout le temps positifs. Ce tableau montre que le nombre de gènes prédits positifs pour les termes spécifiques est faible par rapport aux prédictions des termes « other ».

Tableau 18 : Nombre de gènes prédits positifs par terme.

Le nombre de gènes est détaillé en fonction du nombre de fois où chaque gène est déclaré positif: dans au moins 80% des cas ou dans 100 % des cas.

Ontologie	Terme	Nbr_Gènes_Positifs dans au moins 80% des cas	Nbr_Gènes_Positifs dans 100% des cas
CC	Ribosome	1	1
	Other intracellular components	10	9
	Other cytoplasmic components	8	8
	cytosol	1	1
	chloroplast	20	15
	Plastid	2	1
BP	Response to stress	8	6
	Other cellular processes	4335	4334
	Protein metabolism	11	10
	Response to abiotic or biotic stimulus	6	6
	Other metabolic processes	1382	1272
MF	Structural molecule activity	12	10

Tableau 19 : Nombre de prédictions et nombre de gènes prédits positifs pour l'ensemble des termes.

Les colonnes « Tous » concernent la totalité des 12 termes retenus. Les colonnes « Spécifiques » concernent uniquement les 8 termes spécifiques (sans les termes « other »). Ce nombre est détaillé en fonction du nombre de fois où les gènes sont déclarés positifs parmi les 100 prédictions (au moins 80% des cas ou 100% des cas).

Termes	Tous		Spécifiques	
Positifs dans au moins x% des cas	80%	100%	80%	100%
Nbr prédictions	5796	5673	61	50
Nbr gènes	4348	4345	47	31

b. *Exemple des résultats de prédiction pour le terme « Structural molecule activity »*

Le terme « Structural molecule activity » de l'ontologie MF est étudié grâce à la règle d'inférence dont les paramètres sont : seuil de corégulation 3, type de classement « Dépendant » et type de décompte « occurrence ». Cette règle de décision a été appliquée à 1 848 gènes mal caractérisés pour

l'ontologie MF, parmi lesquels 712 gènes sont orphelins. Le tableau 20 correspond aux identifiants des gènes déclarés positifs au moins 80 fois pour ce terme parmi les 100 prédictions effectuées. Parmi ces gènes, 10 sont tout le temps déclarés positifs.

Les fonctions des gènes peuvent être régulièrement mises à jour dans les bases de données. J'ai alors vérifié si de nouvelles données concernant la fonction de ces gènes déclarés positifs sont disponibles, pouvant confirmer ou infirmer ces prédictions. Cela m'a permis de trouver quelques informations concernant certains gènes mais pour la majorité aucune nouvelle donnée n'a été récemment publiée. Le gène « AT1G52930 », est associé aux termes « other metabolic processes », « other cellular processes », « developmental processes » et « cell organisation and biogenesis » pour l'ontologie BP mais aucun terme MF ne lui est associé. De nouvelles mises à jour dans les bases de données TAIR et PubMed fin 2015 associent ce gène aux termes « RNA processing » et « ribosomal large subunit assembly ». Ces termes appartiennent aussi à l'ontologie BP mais ils sont cohérents avec la fonction moléculaire prédite à savoir le lien avec l'activité ribosomale. Quant au gène « AT3G59650», je n'ai pas trouvé de nouvelles données le concernant, cependant il appartient à la famille « L51/S25/CI-B8 » qui est une famille connue de protéines ribosomales mitochondriales. Cette information est également réconfortante par rapport à la prédiction effectuée. Quant au gène « AT3G49040 », il correspond à l'exemple analysé lors du chapitre 2 dans la composante connexe au seuil de corégulation 7, enrichie en protéines ribosomales. Ce résultat pouvait être attendu car ce terme GO Slim «Structural molecule activity » englobe le terme « structural constituent of ribosome ». Ainsi les résultats de la méthode d'inférence à « haut débit » sont réconfortants puisqu'ils vont dans le même sens que dans l'analyse précédente.

Tableau 20 : Gènes prédits positifs pour le terme « Structural molecule activity ».

Gene	orphelin	nbr fois Positif
AT3G49040	non	100
AT1G52930	non	100
AT1G73940	oui	100
AT5G19650	non	100
AT2G43780	oui	100
AT2G41650	oui	100
AT2G30990	oui	100
AT1G56110	non	100
AT3G59650	non	100
AT4G30220	non	100
AT5G58250	oui	95
AT3G59840	oui	85

VI. Conclusions de ces analyses

J'ai proposé une nouvelle approche d'annotation fonctionnelle automatique par apprentissage qui permet d'exploiter le réseau de corégulation construit grâce à l'intégration des données de coexpression. Cette approche est différente de l'existant et se distingue par une caractérisation centrée sur les termes et non sur les gènes comme c'est le cas pour la majorité des méthodes d'annotation. De plus, contrairement aux méthodes d'apprentissage telles que les SVM, cette méthode vise à définir des règles de décision explicites et compréhensibles par les biologistes pour l'annotation fonctionnelle. Pour un terme étudié, la méthode est fondée le calcul d'un score pour chaque gène du réseau, représentatif de la présence du terme dans son voisinage et sur un classement des gènes selon leurs scores. Différents choix ont été étudiés pour le calcul du score et le tri. Notre méthode vise ainsi à explorer et à évaluer certains choix qu'un annotateur peut envisager dans le cadre d'une analyse fonctionnelle à partir d'un réseau d'interactions. Elle permet aussi de considérer la particularité du réseau notamment les différents seuils de corégulation et vise à étudier l'impact de ce seuil sur la performance de prédiction des termes. Afin d'examiner ces différentes possibilités, j'ai considéré les voisins directs des gènes comme type de voisinage. J'ai également défini un ensemble de paramètres correspondant au seuil de coexpression, au type de graphe, au type de décompte et au type de classement. Ces paramètres peuvent prendre différentes valeurs et leurs différentes combinaisons ont permis de définir 198 classifieurs pour chacun des 42 termes analysés. L'analyse de sensibilité a permis de montrer la pertinence de l'ensemble de ces paramètres sauf pour le type de graphe indiquant ainsi que la valeur des arêtes dans le réseau de corégulation n'apporte pas plus d'informations que son existence. Dans la suite de l'analyse, j'ai donc décidé de ne considérer que les graphes binaires, réduisant ainsi de moitié la taille de la collection de classifieurs. Cette analyse a révélé également l'existence d'interactions entre les paramètres pertinents. Ces interactions varient d'un terme à l'autre appuyant ainsi l'importance d'identifier une règle par terme.
Pour l'assignation de chaque terme à un gène, j'ai déterminé un score seuil associé à chaque classifieur qui devait contrôler le FDR à 20% au maximum. La procédure de validation croisée pour l'évaluation a révélé l'incapacité de plusieurs règles de décision à définir un score seuil dans les jeux d'apprentissage car le FDR n'est jamais inférieur à 20% dans les listes de classement de ces gènes. J'ai donc éliminé ces règles. J'ai également constaté la difficulté de plusieurs autres règles à contrôler le FDR au seuil demandé. La stabilité du FDR entre le jeu d'apprentissage et le jeu test a été ainsi le deuxième critère déterminant pour la sélection des règles de décision. Grâce à ces deux

critères de sélection la collection de règles est réduite de 85% : 16 termes ont été retenus et les 26 autres termes n'ont pas pu être analysés car aucune règle n'a été retenue.

Etant donné que l'analyse des valeurs AUC a montré, à l'exception de trois termes de l'ontologie MF, qu'il existe pour tous les autres termes au moins un classifieur ayant un AUC élevé, et que j'ai constaté un lien entre le contrôle du FDR et la représentativité des termes dans les jeux de données, cela indique la stringence du FDR à 20% pour les termes peu représentés et au contraire la permissivité de ce taux pour les termes très représentés.

Contrairement à d'autres domaines tel que le diagnostic médical par exemple, où les échantillons d'apprentissage peuvent être constitués de manière équilibrée entre les exemples positifs et négatifs, ici nous sommes limités par la différence de représentativité des termes dans le réseau biologique et du déséquilibre de la composition des jeux. De plus, l'utilisation de la technique de validation croisée avec tirage aléatoire des blocs accentue le déséquilibre entre exemples positifs et exemples négatifs pour l'apprentissage surtout pour les termes les moins représentés dans le réseau. Cela explique probablement en grande partie la difficulté de la méthode à caractériser ces termes et indique ainsi la nécessité d'adapter la valeur maximale autorisée pour le FDR à leur représentativité et donc à chaque terme. Conjointement à cette constatation, notre analyse a montré l'importance d'une caractérisation par terme et la variabilité de l'impact des différents paramètres en fonction du terme analysé ainsi que leurs interactions. Je pense alors que la nécessité de déterminer un seuil associé à un FDR propre à chaque terme en fonction de sa représentativité a un sens et un intérêt biologique.

D'un autre côté, la difficulté des règles à contrôler le FDR pour plusieurs termes reflète probablement aussi l'insuffisance des données transcriptomiques pour la caractérisation de ces termes notamment ceux de l'ontologie MF. En effet, il a été démontré dans la littérature que les méthodes qui reposent sur le principe de culpabilité par association permettent d'améliorer la performance de prédiction des termes de l'ontologie CC et BP mais ils sont moins performants pour la prédiction des termes de l'ontologie MF qui sont mieux prédits par les méthodes fondées sur la similarité de séquence. Cela est dû au fait que les partenaires fonctionnels interagissent ensemble dans l'objectif d'effectuer une même fonction biologique sans pour autant avoir nécessairement la même fonction moléculaire. Cependant certains termes de l'ontologie BP et CC ne peuvent pas être prédits à partir du réseau de corégulation uniquement. L'intégration de ce réseau avec les autres données omiques permettrait probablement d'améliorer la performance de prédiction de ces termes.

Malgré ces remarques, la méthode proposée permet de faire de la prédiction pour 16 termes. Afin de sélectionner la meilleure règle, je me suis appuyée sur la métrique Fmeas communément utilisée dans la littérature pour évaluer et comparer la performance de classifieurs. L'analyse des valeurs de Fmeas a montré que certaines règles ayant un FDR contrôlé peuvent avoir une faible sensibilité, mais dans l'ensemble ces règles ont une meilleure sensibilité.

Cette analyse des valeurs Fmeas a permis également de montrer que la performance de prédiction augmente avec le seuil de corégulation soulignant ainsi la pertinence du réseau de corégulation.

Concernant le paramètre « type de classement », l'analyse des règles sélectionnées montre l'importance de la prise en considération de la dépendance entre les termes qui est bien connue.

Finalement, j'ai sélectionné une règle par terme ayant un Fmeas minimal de 0,2 et permettant de maximiser le nombre de gènes prédits. Au final 12 termes ont été prédits avec une moyenne Fmeas de 0,421 et un FDR de 0,199 soit une précision de 0.80 ce qui correspond à une meilleure performance que celle des méthodes comparées dans *Wang et al*. (2013). La prédiction étant réalisée à l'aide d'une procédure de validation croisée, un indice de confiance défini par le nombre de fois où il est prédit positif parmi 100 prédictions est également disponible. Ces indices permettent ainsi d'augmenter la confiance accordée à la prédiction de ces gènes.

Discussions et Perspectives

Le processus d'annotation fonctionnelle des gènes inconnus reste un défi majeur de la biologie car il est essentiel à la compréhension de la quantité d'informations générées par le séquençage d'un nombre toujours en forte augmentation de nouveaux génomes. Dans la partie état de l'art de cette thèse, j'ai fourni une vue d'ensemble des principales méthodes et des problématiques existantes dans ce cadre. Cette vue globale montre que les méthodes émergentes sont fondées sur la recherche de partenaires notamment à partir des données d'interactions moléculaires. La fonction des gènes inconnus est alors déduite à partir de celles de leurs partenaires en s'appuyant sur l'hypothèse d'association par culpabilité. Ces méthodes ont permis de compléter et de pallier aux limites de l'annotation fonctionnelle par similarité de séquence et elles sont d'autant plus facilitées par l'accumulation d'une quantité importante de données omiques à haut débit notamment les données transcriptomiques. L'intérêt de l'intégration de toutes les données disponibles a été démontré notamment pour la compréhension du fonctionnement complexe et hiérarchisé des systèmes biologiques ainsi que pour pallier aux limites de chaque type de données. Cependant, la majorité des méthodes existantes souffrent du manque de spécificité du contexte biologique et (ou) du manque de contrôle de l'hétérogénéité des données et globalement la performance de prédiction de ces méthodes reste modeste.

Orienté autour de la problématique de l'annotation fonctionnelle, mon projet de thèse s'est articulé autour de deux axes principaux qui sont l'intégration de données pour l'identification de partenaires fonctionnels et l'exploitation de ces données pour l'inférence de fonction aux gènes inconnus. Partant de 681 groupes de gènes d'*Arabidopsis thaliana* ayant la même dynamique de réponse à un stress parmi 18 catégories de stress biotiques et abiotiques, mon objectif était le développement de méthodologies pour l'exploitation et l'enrichissement de ces clusters de coexpression afin d'identifier les processus et les gènes impliqués dans la réponse aux stress. Mon travail de thèse peut se découper principalement en trois étapes importantes: Premièrement, une intégration de données hétérogènes appartenant à différents niveaux moléculaires et organisationnels de la cellule pour l'interprétation biologique des clusters de gènes. Deuxièmement, une intégration des données de coexpression afin de s'affranchir de leurs contraintes contextuelles et environnementales et de les rendre comparables aux autres types de données moléculaires. Finalement, la mise en place d'une

méthode d'apprentissage supervisée pour la prédiction de fonction aux gènes orphelins ou mal caractérisés à partir du réseau d'interactions moléculaires construit.

Concernant la première étape et partant des constats tirés de la littérature, mes travaux se sont situés dans un contexte spécifique qui est la réponse aux stress et se sont recentrés sur le contrôle de l'hétérogénéité des données utilisées en termes de sources et de techniques d'acquisition. Par ailleurs, du fait que l'objectif principal était l'enrichissement des clusters de coexpression et du fait que cette ressource est la plus complète, mon choix était porté sur l'intégration de chacun de ces différents types de données autour des clusters de coexpression. Ce choix a été également guidé par un souci d'interprétabilité des résultats ainsi que par la variabilité de disponibilité des données et la variabilité des objets renseignés par les différents types de données. Selon ces choix, j'ai développé une approche globale intégrant un ensemble de données et de ressources biologiques hétérogènes autour des clusters de coexpression. L'ensemble de cette ressource a été mise à disposition grâce à une interface graphique conviviale qui permet d'observer de manière simple toutes les spécificités des groupes identifiés ainsi que leurs principales différences (Zaag *et al.* 2015). Cette ressource constitue une avancée dans la compréhension des déterminants majeurs de la résistance chez Arabidopsis et de la réponse coordonnée de tous ces acteurs.

Pour les données caractérisant les relations entre les gènes, elles étaient intégrées séparément afin de construire des réseaux d'interactions moléculaires. Du fait de l'abondance des relations de coexpression entre les gènes d'un même cluster par rapport aux interactions de type TF-cibles ou PPI qui les relient ainsi qu'à la variabilité de confiance accordée à chaque type de données, le choix a été porté sur l'intégration des données de coexpression de manière visuelle avec les deux autres types d'interactions. Bien qu'elle ait mis l'accent sur plusieurs hubs qui représentent des candidats potentiels à des régulateurs clés de la réponse aux stress (Frei Dit Frey *et al.* 2014), cette analyse m'a confrontée à la difficulté de l'intégration de ces données du fait de leur hétérogénéité en termes de type et de qualité ainsi que de la variabilité de leur disponibilité. Cette hétérogénéité complique l'interprétation biologique du réseau obtenu et explique probablement le faible chevauchement entre ces données omiques.

Confrontée à ces problèmes, je me suis alors focalisée sur l'intégration des données de coexpression dans l'objectif de contrôler leur hétérogénéité et de renforcer leur pertinence biologique. Cette intégration visait ainsi à affranchir les données de coexpression de leurs contraintes contextuelles liées aux conditions expérimentales que je considère comme la principale divergence avec les données PPI et TF-cibles. En effet, les PPI sont issues de techniques éliminant les variations

environnementales et les TF-cibles utilisées ont été validées par deux techniques indépendantes. C'est pour cela que j'ai considéré que ces deux types de données indépendantes du contexte sont absolus, contrairement à mes données de coexpression et que pour les rendre comparables je les ai transformées. Cette analyse d'intégration de la coexpression correspond ainsi à la deuxième étape de mes travaux de thèse et s'est concrétisée par une analyse transversale intégrant les différentes catégories de stress. Cela a permis de mettre en évidence l'existence de nombreux couples de gènes coexprimés dans plusieurs catégories de stress allant jusqu'à 14 catégories sur les 18 étudiées. Le test de permutation a permis de montrer que la conservation des paires de gènes dans les mêmes clusters à partir de 3 catégories de stress ne peut pas être due au hasard. J'ai conclu alors sur la corégulation de ces gènes à partir du seuil 3, d'autant plus que l'analyse des promoteurs de ces gènes montre la présence de nombreux motifs cis-régulateurs indiquant que ces gènes sont sous le contrôle des mêmes régulateurs. Ces paires de gènes identifiées ont servi à la construction d'un réseau de corégulation où les nœuds correspondent aux gènes et les arêtes correspondent aux relations de corégulation entre chaque couple de gènes et qui prennent la valeur du nombre de catégories de stress dans lesquelles les deux gènes sont coexprimés. La valeur d'une arête prend une note allant de 1 à 14 sur 18 et reflète ainsi la force de la corégulation entre les deux gènes. Il serait probablement intéressant par la suite de préciser pour chaque paire corégulée, le nombre de fois où les deux gènes sont étudiés ensemble parmi les 18 catégories de stress pour affiner la note que prend chaque arête. En effet, il est possible que certains couples de gènes ne soient pas corégulés dans les autres catégories de stress parce qu'ils n'ont tout simplement pas passé les critères pour être inclus dans l'analyse de coexpression. Cela permettra d'augmenter la confiance accordée à certaines paires de gènes qui se retrouvent corégulés et ainsi de mieux spécifier leurs conditions de corégulation : par exemple gènes de réponse aux stress biotiques, abiotiques etc. Effectivement, même avec la construction actuelle, le réseau de corégulation peut servir non seulement pour l'annotation fonctionnelle mais aussi pour la comparaison de la réponse aux stress chez Arabidopsis. L'utilisation de ces réseaux pour la comparaison des stress correspond à un deuxième champ de recherche et nécessite plus d'investigations que je n'ai pas eu le temps de mener pendant ma thèse. Toutefois, j'ai pu effectuer une exploration rapide en spécifiant la catégorie de stress dans laquelle la paire de gènes est impliquée : uniquement biotique, uniquement abiotique ou commune. Cette démarche a permis d'identifier 5 844 arêtes spécifiquement abiotiques, 2 517 arêtes spécifiquement biotiques et 49 472 arêtes communes parmi les 57 833 arêtes du réseau de gènes coexprimés dans au moins 3 catégories de stress. Cela indique ainsi que la majorité des voies de réponse sont communes entre les stress

biotiques et abiotiques. Ces résultats sont en adéquation d'une part avec des constatations précédentes révélant que la majorité des gènes de réponse aux stress ne sont pas spécifiques d'un stress particulier (Rodriguez et Redman, 2005 ; Kilian *et al*. 2007). D'autre part, ils confirment les résultats d'une autre analyse que j'ai menée comparant le nombre de gènes différentiellement exprimés dans les conditions de stress biotiques et celui dans les conditions de stress abiotiques. Cette analyse a révélé un large chevauchement des deux jeux de données qui représente 80% de la taille de leur union. Comme il s'est avéré avec ces premières investigations que la majorité des interactions sont communes, il serait probablement intéressant d'effectuer une analyse topologique du réseau total ou une comparaison topologique des deux réseaux formés par les gènes impliqués spécifiquement dans les stress biotiques et ceux des stress abiotiques afin d'identifier les modules communs et spécifiques de la réponse de la plante à ces deux catégories de stress.

Par ailleurs dans cette analyse transversale, je me suis appuyée sur la recherche de motifs cis-régulateurs au sein des promoteurs des gènes pour valider la corégulation des gènes identifiés puisque ces motifs sont les sites potentiels de la liaison des facteurs de transcription responsables de la régulation de l'expression de ces gènes. Toutefois, la corégulation peut aussi être expliquée par d'autres facteurs tels que les petits ARN ou encore les SMAR (Scaffold Matrix Attachment Region). En effet, les petits ARN jouent un rôle important dans la régulation post-transcriptionnelle des gènes. Quant aux SMAR, ce sont des séquences d'ADN, régulièrement localisées le long du génome et qui pourraient être impliquées dans l'organisation chromatinienne et la modulation de l'expression des gènes. Il serait ainsi intéressant d'une part, d'identifier des cibles connues de petits ARN dans le réseau de corégulation et leur distribution selon les familles auxquelles elles appartiennent, d'autre part nous pourrions intégrer le réseau de corégulation avec les informations disponibles sur les SMARs. Ainsi ce réseau de corégulation pourrait servir d'amorce dans l'orientation de futurs projets de recherches qui pourraient contribuer à leur tour à conforter la validité biologique de la corégulation des gènes identifiés par l'analyse intégrative de la coexpression.

Dans le cadre de ma thèse, mon objectif était d'exploiter le réseau de corégulation pour l'annotation fonctionnelle. Cette exploitation est également un moyen de montrer le potentiel de ce réseau par la prédiction et ainsi appuyer la validité biologique des liens de corégulation et fonctionnels établis. Dans une perspective exploratoire, le choix s'est donc porté sur le développement d'une méthode d'inférence à partir du réseau de corégulation avant son intégration avec d'autres données omiques. A terme, l'objectif serait d'appliquer cette méthode pour l'exploitation du réseau enrichi intégrant toutes les données d'interactions moléculaires collectées.

La comparaison de performance de certaines des méthodes d'inférence dans la littérature montre que les méthodes d'apprentissage automatique telles que celles fondées sur les noyaux ne permettent qu'une légère amélioration de la performance de prédiction par rapport aux méthodes simples telles que le vote majoritaire (Wang *et al.* 2013). De plus, la mise en place et l'utilisation de certaines de ces méthodes pour l'annotation fonctionnelle est difficile et leurs paramètres demandent une certaine maîtrise. Les règles de décisions issues de certaines analyses telles que les SVN représentent des boites noires qui ne sont ni compréhensibles ni interprétables biologiquement. Finalement, j'ai remarqué que dans certaines études la performance de prédiction est non seulement variable en fonction des ontologies analysées mais également en fonction des termes (Radivojac *et al.* 2013 ; Ryngajllo *et al.* 2011). A partir de ces constats, j'ai opté pour le développement d'une méthode binaire d'apprentissage supervisée centrée sur les termes. Habituellement les méthodes d'apprentissage supervisées visent à optimiser les paramètres d'une fonction par minimisation d'un estimateur d'erreur. Ici, l'approche que j'ai proposée est différente et consiste à considérer un ensemble de classifieurs dépendants de plusieurs paramètres dont je calcule les performances. Le meilleur classifieur par terme a été alors choisi en fonction des métriques FDR et Fmeas. Ces classifieurs calculent un score pour chaque gène représentatif de la présence du terme analysé dans son voisinage, stratégie qui peut être vue comme étant proche de celle du vote majoritaire (Schwikowski *et al.* 2000). Cette approche est par contre une méthode de classification multi-classes centrée sur les gènes puisqu'elle propose d'assigner à chaque gène étudié les trois termes les plus fréquents dans son entourage. Notre méthode centrée sur les termes permet de considérer l'ensemble des scores calculés pour tous les gènes du réseau afin d'identifier un score seuil permettant la prédiction des gènes positifs et négatifs pour chaque terme étudié. Je considère alors que c'est une approche globale qui permet d'explorer l'ensemble de l'information biologique codée dans le réseau d'interactions. Notre méthode se démarque également par le fait qu'elle contrôle le taux de faux positifs parmi la proportion de gènes déclarés positifs (FDR), métrique non utilisée à ma connaissance dans le domaine de l'annotation fonctionnelle pour l'évaluation de la performance des classifieurs. Or confrontée au problème d'évaluation des classifieurs, je me suis rendue compte de l'importance de contrôler ce taux afin de renforcer la confiance accordée à nos prédictions par les biologistes qui se soucient de limiter au maximum les coûts de validations expérimentales inutiles. Actuellement, les FDR générés par les méthodes d'annotation restent très élevés. En effet, j'ai pu

constater qu'il est de 70% pour le vote majoritaire et peut même atteindre 85% pour d'autres méthodes.

Dans notre approche, le FDR est utilisé pour fixer un score seuil pour la prédiction des gènes. Il est également utilisé pour l'évaluation de la capacité des classifieurs à le contrôler en comparant sa valeur entre les jeux d'apprentissage et les jeux test. Dans cette analyse, j'ai pu constater que le FDR était sensible à la représentativité des termes et à la composition en exemples positifs et exemples négatifs des jeux d'apprentissage. Ma conclusion est que le FDR maximal autorisé dans notre analyse semble être très stringent pour les termes peu représentés et au contraire trop permissif pour les termes très représentés. Cette métrique favorise donc la prédiction des termes majoritaires dans le réseau. Afin d'adapter cette métrique au contexte d'échantillons déséquilibrés, une solution serait de définir une valeur maximale de FDR par terme; une démarche cohérente avec les résultats démontrés par la méthode à savoir l'inférence par terme et la sélection d'une meilleure règle par terme. Pour cela, une piste serait de faire varier le FDR autorisé jusqu'à une certaine valeur maximale (par exemple 40%) dans les jeux d'apprentissage et étudier les valeurs obtenues dans les jeux test. Le choix de cette valeur maximale dépendrait des besoins et des priorités à savoir : privilégier la confiance aux prédictions ou privilégier le résultat en termes de nombre de prédictions. Certains considèrent que de proposer un FDR par exemple à 60% pour des termes représentés moins de 20% est une amélioration du pouvoir prédictif puisqu'un classifieur aléatoire a 80% de chances de se tromper pour la classification des gènes annotés par ce terme. Pour d'autres, proposer une liste avec 60% de faux positifs n'a pas beaucoup d'intérêt. Pourtant, malgré les avantages du FDR, la limite du FDR dans ce contexte particulier, me pousse à me poser la question de savoir si le FDR était le bon critère pour déterminer le seuil score dans les listes de classements et quels seraient les résultats obtenus avec d'autres critères telles que le Fmeas par exemple qui représente un compromis entre la précision et la sensibilité. D'autres critères ont été également proposés dans la littérature pour chercher des points optimaux sur le graphe rappel-précision ou la courbe ROC, notamment le point BEP (break-even Point) proposé par Sebastiani (2002) qui correspond au point d'égalité entre la précision et la sensibilité ou encore la mesure Pragma proposée par Thomas *et al*. (2007) qui permet de spécifier l'importance accordée à chaque classe ainsi que ses préférences en termes de sensibilité et de précision. En effet, dans cette étude ils considèrent que la sensibilité et la précision sont les mesures les plus adaptées concernant la prédiction d'une classe rare et spécifique et proposent de prendre en compte l'aspect non symétrique des modalités d'une classe. Pour comparer le pouvoir

prédictif de la méthode, il serait probablement intéressant de comparer d'un côté la performance de la méthode actuelle avec celles obtenues en s'appuyant sur d'autres métriques pour fixer le score seuil. D'un autre côté, nous pourrions appliquer une autre méthode d'apprentissage par exemple celle du vote majoritaire au réseau de corégulation. Partant des mêmes données cela permettra de mettre en évidence la performance de la méthode par rapport aux autres.

L'objectif de l'exploitation du réseau de corégulation est de prédire une fonction aux gènes orphelins ou mal annotés en s'appuyant sur la fonction des gènes connus au sein de cet espace. Cependant, les gènes peuvent être partiellement étiquetés, en d'autres termes, les fonctions connues des gènes peuvent ne représenter qu'une fraction de la réalité biologique. Il est aussi connu que certaines erreurs d'annotation parmi cet espace de gènes annotés sont probables et peuvent provenir de la propagation de connaissances d'un gène à un autre que ce soit par similarité de séquence ou par d'autres approches de prédictions. L'apprentissage à partir de ces données incomplètes, dont la qualité peut être discutable, peut ainsi soulever des questions quant à la fiabilité des prédictions ainsi qu'à l'évaluation des classifieurs. Ce constat impose donc d'être attentif et critique à l'égard des annotations disponibles ainsi qu'aux prédictions qui en découlent. Les prédictions doivent être considérées avant tout comme des hypothèses mais essentielles pour orienter les démarches de validation expérimentales. Compte tenu de ces constats, les méthodes d'annotation peuvent alors apporter des éléments de réponse non seulement autour des fonctions des gènes inconnus mais également autour de celles des gènes déjà connus afin d'améliorer leurs annotations. Pour les résultats de notre méthode, il serait intéressant de considérer les prédictions positives avec un indice de confiance élevé affectant aux gènes d'apprentissage de nouvelles fonctions qui ne leur étaient pas encore associées.

Les résultats de la méthode de prédiction proposée ont contribué à associer la fonction de nombreux gènes à un terme parmi les 12 termes prédits. Pour les 30 autres termes, il n'a pas été possible de les prédire principalement à cause de la stringence du FDR imposé et de la difficulté à le contrôler mais également à cause de l'insuffisance des données transcriptomiques à elles seules pour la caractérisation de ces termes. En effet, la majorité des termes MF n'ont pas été retenus pour la prédiction avec la méthode. Or il est connu que les termes de cette ontologie sont mieux prédits par les approches qui s'appuient sur les données de séquences génomiques. Il n'est donc pas étonnant que la méthode n'a pas réussi à les caractériser et que cela n'est pas dû uniquement à la stringence des critères d'évaluation mais aussi aux données utilisées. Certains autres termes de l'ontologie BP et CC n'ont pas été retenus également. Nous pouvons espérer qu'en intégrant les données PPI, TF-

cibles et les données de corégulation, nous arriverons à prédire ces termes. D'autres données pourraient également être intégrées comme celles des interactions métaboliques dont la disponibilité est en constante augmentation et qui pourraient contribuer à améliorer notre compréhension des systèmes biologiques. Bien évidemment la question de comment intégrer toutes ces données de manière appropriée reste le défi principal. Probablement la manière la plus intuitive est l'identification des interactions métaboliques impliquant les gènes du réseau de corégulation ce qui permettra de mettre en évidence les processus métaboliques impliqués dans la réponse aux stress. C'est alors une stratégie qui s'appuie encore sur le principe de chevauchement entre données hétérogènes afin d'augmenter la confiance aux résultats obtenus. Cette stratégie est communément utilisée dans la littérature et je l'ai adoptée également dans les deux premières étapes de ma thèse. En effet, la stratégie proposée lors du premier chapitre s'appuie sur le chevauchement des différents types de données par la mise en évidence des relations entre elles grâce aux enrichissements fonctionnels dans les groupes de gènes coexprimés, ce qui a permis d'appuyer leur validité biologique. L'intégration des données d'interactions moléculaires s'est effectuée en considérant l'ensemble des interactions PPI et TF-cibles impliquant les gènes différentiellement exprimés dans une catégorie de stress et les mettre en relation visuelle avec les données de coexpression. Le faible chevauchement entre ces données dans le réseau reflété par un non regroupement des gènes d'un même cluster, nous a conduits à conclure que l'hétérogénéité du contexte des données de coexpression en est la cause. Donc malgré l'effort pour limiter l'hétérogénéité en contrôlant le contexte, la source et la qualité des données, nous étions confrontés à la différence de la nature des données en particulier le fait que la coexpression est liée à un contexte environnemental précis contrairement aux PPI et TF-cibles. L'intégration des données de coexpression effectuée également par une stratégie de chevauchement, avait pour objectif de s'affranchir du contexte afin de les rendre comparables aux autres données. J'ai montré dans le deuxième chapitre par la mise en évidence des enrichissements fonctionnels et des motifs cis-régulateurs que cette analyse a permis de renforcer la pertinence biologique de ces données. Reste à déterminer si cette analyse qui fait passer la coexpression à la corégulation permet de faciliter leurs intégrations avec les autres données omiques et sur quels critères nous nous appuyons pour confirmer cette amélioration : Est-ce toujours le chevauchement des données ?

Concernant le chevauchement du réseau de corégulation avec les données PPI, il s'est avéré qu'il est quasiment nul. A partir de ce résultat, j'en conclue donc que contrairement à l'hypothèse que j'ai émis à la fin de mon premier chapitre, ce n'est pas l'hétérogénéité du contexte qui explique le faible

chevauchement entre les données omiques, puisqu'elle a été contrôlée au même titre que l'hétérogénéité de leurs sources et de leurs qualités. Ce dernier résultat ne remet pas en cause la qualité des données de corégulation dont nous avons montré la pertinence et qui sont maintenant de même nature que les autres données. En revanche cela remet probablement en cause la stratégie qui consiste à toujours chercher l'intersection entre les données. Le faible chevauchement des données omiques a été largement constaté dans la communauté (De Bodt *et al*. 2009). Cela a été généralement expliqué par la qualité des données ou encore par leur complémentarité. D'un autre côté, il est maintenant connu que les processus biologiques reposent sur un fonctionnement hiérarchisé et intégré qui résulte de l'interaction de différentes entités biologiques appartenant à différents niveaux organisationnels. Chacune de ces données apporte un niveau d'information différent sur la régulation et le niveau organisationnel de la cellule. Ainsi plutôt que de chercher l'intersection il semble plus cohérent et pertinent biologiquement de considérer l'union de toutes ces données du moment que nous avons confiance dans leur qualité. La construction d'un réseau constitué par l'ensemble des interactions permettra ainsi d'enrichir nos connaissances en termes de partenaires fonctionnels et probablement d'améliorer la performance de prédiction de la fonction des gènes. Il serait alors intéressant de comparer la performance de prédiction à partir du réseau de corégulation contre celui construit par l'union des données omiques, ce qui permettrait de mettre en évidence la contribution de chaque type de données pour l'annotation.

Finalement, à terme les informations issues de cette étude chez la plante modèle *A. thaliana* pourront servir d'amorce pour l'annotation fonctionnelle et la découverte de nouveaux gènes de résistance chez les espèces cultivées, volet incontournable pour élaborer les stratégies d'amélioration variétales.

Références

Arabidopsis Interactome Mapping Consortium. 2011. Evidence for network evolution in an Arabidopsis interactome map. *Science* 333: 601-7.

Adamcsek, B., G. Palla, I. J. Farkas, I. Derenyi, and T. Vicsek. 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22: 1021-3.

Albert, R., H. Jeong, and A. L. Barabasi. 2000. Error and attack tolerance of complex networks. *Nature* 406: 378-82.

Alon, U. 2007. Network motifs: theory and experimental approaches. *Nat Rev Genet* 8: 450-61.

Altschul, S. F., and W. Gish. 1996. Local alignment statistics. *Methods Enzymol* 266: 460-80.

Andreeva, A., D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32: D226-9.

Armstrong, N. J., and M. A. van de Wiel. 2004. Microarray data analysis: from hypotheses to conclusions using gene expression data. *Cell Oncol* 26: 279-90.

Arnau, V., S. Mars, and I. Marin. 2005. Iterative cluster analysis of protein interaction data. *Bioinformatics* 21: 364-78.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-9.

Atias, O., B. Chor, and D. A. Chamovitz. 2009. Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network. *BMC Syst Biol* 3: 86.

Aubourg, S., M. L. Martin-Magniette, V. Brunaud, L. Taconnat, F. Bitton, S. Balzergue, P. E. Jullien, M. Ingouff, V. Thareau, T. Schiex, A. Lecharny, and J. P. Renou. 2007. Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. *BMC Genomics* 8: 401.

Bader, G. D., and C. W. Hogue. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4: 2.

Balazsi, G., A. L. Barabasi, and Z. N. Oltvai. 2005. Topological units of environmental signal

processing in the transcriptional regulatory network of Escherichia coli. *Proc Natl Acad Sci U S A* 102: 7841-6.

Barabasi, A. L., and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286: 50912.

Barabasi, A. L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5: 101-13.

Barrett, T., D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. 2007. NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* 35: D760-5.

Barriot, R., J. Poix, A. Groppi, A. Barre, N. Goffard, D. Sherman, I. Dutour, and A. de Daruvar. 2004. New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res* 32: 3581-9.

Barutcuoglu, Z., R. E. Schapire, and O. G. Troyanskaya. 2006. Hierarchical multi-label prediction of gene function. *Bioinformatics* 22: 830-6.

Bashton, M., and C. Chothia. 2007. The generation of new protein functions by the combination of domains. *Structure* 15: 85-99.

Bassel, G. W., A. Gaudinier, S. M. Brady, L. Hennig, S. Y. Rhee, and I. De Smet. 2012. Systems analysis of plant functional, transcriptional, physical interaction, and metabolic networks. *Plant Cell* 24: 3859-75.

Bassel, G. W., E. Glaab, J. Marquez, M. J. Holdsworth, and J. Bacardit. 2011b. Functional network construction in Arabidopsis using rule-based machine learning on large-scale data sets. *Plant Cell* 23: 3101-16.

Bassel, G. W., H. Lan, E. Glaab, D. J. Gibbs, T. Gerjets, N. Krasnogor, A. J. Bonner, M. J. Holdsworth, and N. J. Provart. 2011a. Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. *Proc Natl Acad Sci U S A* 108: 9709-14.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer. 2000. The Pfam protein families database. *Nucleic Acids Res* 28: 263-6.

Berg, J., and M. Lassig. 2004. Local graph alignment and motif search in biological networks. *Proc Natl Acad Sci U S A* 101: 14689-94.

Bergmann, S., J. Ihmels, and N. Barkai. 2004. Similarities and differences in genome-wide expression data of six organisms. *PLoSBiol* 2: E9.

Bernard, V., V. Brunaud, and A. Lecharny. 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.

Bernard, V., A. Lecharny, and V. Brunaud. 2006. Improved detection of motifs with preferential location in promoters. *Genome* 53: 739-52.

Bernstein, F. C., T. F. Koetzle, G. J. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112: 535-42.

Blagosklonny, M. V., and A. B. Pardee. 2002. Conceptual biology: unearthing the gems. *Nature* 416: 373.

Boccalettia S, Latora V, Moreno Y, Chavezf M, Hwanga Yook DU. Complex networks: Structure and dynamics. Physics Reports. 2006;424:4-5.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31: 365-70.

Boser, B.E., Guyon, I.M., and Vapnik, V. 1992. A training algorithm for optimal margin classifers. Proc. 5th Annual ACM Workshop on Computational Learning Theory,144-152.

Boutrot, F., C. Segonzac, K. N. Chang, H. Qiao, J. R. Ecker, C. Zipfel, and J. P. Rathjen. 2010. Direct transcriptional control of the Arabidopsis immune receptor FLS2 by the ethylene-dependent transcription factors EIN3 and EIL1. *Proc Natl Acad Sci U S A* 107: 14502-7.

Braun, P., S. Aubourg, J. Van Leene, G. De Jaeger, and C. Lurin. 2013. Plant protein interactomes. *Annu Rev Plant Biol* 64: 161-87.

Brun, C., F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq. 2003. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol* 5: R6.

Burges, C.J.C. 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 121 -167.

Cabusora, L., E. Sutton, A. Fulmer, and C. V. Forst. 2005. Differential network expression during drug and stress response. *Bioinformatics* 21: 2898-905.

Carmona-Saez, P., M. Chagoyen, F. Tirado, J. M. Carazo, and A. Pascual-Montano. 2007. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol* 8: R3.

Castrillo, G., F. Turck, M. Leveugle, A. Lecharny, P. Carbonero, G. Coupland, J. Paz-Ares, and L. Onate-Sanchez. Speeding cis-trans regulation discovery by phylogenomic analyses coupled with screenings of an arrayed library of Arabidopsis transcription factors. *PLoS One* 6: e21524.

Chen, H., L. Xue, S. Chintamanani, H. Germain, H. Lin, H. Cui, R. Cai, J. Zuo, X. Tang, X. Li, H. Guo, and J. M. Zhou. 2009. ETHYLENE INSENSITIVE3 and ETHYLENE INSENSITIVE3 -LIKE1 repress SALICYLIC ACID INDUCTION DEFICIENT2 expression to negatively regulate plant innate immunity in Arabidopsis. Plant Cell 21: 2527-40.

Cheng, L., H. Lin, Y. Hu, J. Wang, and Z. Yang. 2014. Gene function prediction based on the Gene Ontology hierarchical structure. *PLoS One* 9: e107187.

Chua, H. N., W. K. Sung, and L. Wong. 2006. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 22: 1623-30.

Chung SY, Wooley JC. Challenges Faced in the Integration of Biological Information. In: Lacroix Z, Critchlow T, editors. Bioinformatics: Managing Scientific Data. San Francisco: Morgan Kaufmann; 2003. pp. 11-34.

Clark, C., and J. Kalita. 2014. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics* 30: 2351-9.

Collas, P. 2010. The current state of chromatin immunoprecipitation. *Mol Biotechnol* 45: 87-100.

Corpet, F., F. Servant, J. Gouzy, and D. Kahn. 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* 28: 267-9.

Cozzetto, D., D. W. Buchan, K. Bryson, and D. T. Jones. 2013. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinformatics* 14 Suppl 3: S1.

Cristianini, N., and Shawe-Taylor, J. 2000.An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK.

Crowe, M. L., C. Serizet, V. Thareau, S. Aubourg, P. Rouze, P. Hilson, J. Beynon, P. Weisbeek, P. van Hummelen, P. Reymond, J. Paz-Ares, W. Nietfeld, and M. Trick. 2003. CATMA: a complete Arabidopsis GST database. *Nucleic Acids Res* 31: 156-8.

D'Haeseleer, P., S. Liang, and R. Somogyi. 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16: 707-26.

Dandekar, T., B. Snel, M. Huynen, and P. Bork. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23: 324-8.

Davuluri, R. V., H. Sun, S. K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz, and E. Grotewold. 2003. AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4: 25.

De Bodt, S., D. Carvajal, J. Hollunder, J. Van den Cruyce, S. Movahedi, and D. Inze. 2010. CORNET: a user-friendly tool for data mining and integration. *Plant Physiol* 152: 1167-79.

De Bodt, S., J. Hollunder, H. Nelissen, N. Meulemeester, and D. Inze. 2012. CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. *New Phytol* 195: 707-20.

De Bodt, S., S. Proost, K. Vandepoele, P. Rouze, and Y. Van de Peer. 2009. Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics* 10: 288.

de Lichtenberg, U., L. J. Jensen, S. Brunak, and P. Bork. 2005. Dynamic complex formation during the yeast cell cycle. *Science* 307: 724-7.

Demeter, J., C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, G. Sherlock, and C. A. Ball. 2007. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* 35: D766-70.

Deng, M., Z. Tu, F. Sun, and T. Chen. 2004. Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics* 20: 895-902.

Deng, M., K. Zhang, S. Mehta, T. Chen, and F. Sun. 2003. Prediction of protein function using protein-protein interaction data. *J Comput Biol* 10: 947-60.

Denoux, C., R. Galletti, N. Mammarella, S. Gopalan, D. Werck, G. De Lorenzo, S. Ferrari, F. M. Ausubel, and J. Dewdney. 2008. Activation of defense response pathways by OGs and Flg22 elicitors in Arabidopsis seedlings. *Mol Plant* 1: 423-45.

Derozier, S., F. Samson, J. P. Tamby, C. Guichard, V. Brunaud, P. Grevet, S. Gagnot, P. Label, J. C. Leple, A. Lecharny, and S. Aubourg. 2011. Exploration of plant genomes in the FLAGdb++ environment. *Plant Methods* 7: 8.

Dolinski, K., and D. Botstein. 2007. Orthology and functional conservation in eukaryotes. *Annu Rev Genet* 41: 465-507.

Domazet-Loso, T., and D. Tautz. 2003. An evolutionary analysis of orphan genes in Drosophila.

Genome Res 13: 2213-9.

Dozmorov, M. G., C. B. Giles, and J. D. Wren. 2011. Predicting gene ontology from a global meta-analysis of 1-color microarray experiments. *BMC Bioinformatics* 12 Suppl 10: S14.

Dunn, R., F. Dudbridge, and C. M. Sanderson. 2005. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* 6: 39.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95: 14863-8.

Elemento, O., N. Slonim, and S. Tavazoie. 2007. A universal framework for regulatory element discovery across all genomes and data types. Mol Cell 28: 337-50.

Elkan and Noto, 2008. Learning Classifiers from Only Positive and Unlabeled Data. Proceedings of the 14th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), 213-220.

Elnitski, L., V. X. Jin, P. J. Farnham, and S. J. Jones. 2006. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res* 16: 1455-64.

Emanuelsson, O., H. Nielsen, and G. von Heijne. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci* 8: 978-84.

Enright, A. J., I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402: 86-90.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30: 1575-84.

Eulgem, T., P. J. Rushton, S. Robatzek, and I. E. Somssich. 2000. The WRKY superfamily of plant transcription factors. *Trends Plant Sci* 5: 199-206.

Falda, M., S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, A. Facchinetti, E. Cilia, R. Velasco, and P. Fontana. 2012. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC Bioinformatics* 13 Suppl 4: S14.

Farnham, P. J. 2009. Insights from genomic profiling of transcription factors. *Nat Rev Genet* 10: 605-16.

Fox, A. D., B. J. Hescott, A. C. Blumer, and D. K. Slonim. Connectedness of PPI network neighborhoods identifies regulatory hub proteins. *Bioinformatics* 27: 1135-42.

Fraley C, Raftery AE. 2003. Enhanced software for model-based clustering, discriminant analysis, and density estimation: MCLUST. Journal of Classification. 20:263-286.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. Evolutionary rate in the protein interaction network. *Science* 296: 750-2.

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

Fujishima, K., M. Komasa, S. Kitamura, H. Suzuki, M. Tomita, and A. Kanai. 2007. Proteome-wide prediction of novel DNA/RNA-binding proteins using amino acid composition and periodicity in the hyperthermophilic archaeon Pyrococcus furiosus. *DNA Res* 14: 91-102.

Fukuchi, S., and K. Nishikawa. 2004. Estimation of the number of authentic orphan genes in bacterial genomes. *DNA Res* 11: 219-31, 311-313.

Gagnot, S., J. P. Tamby, M. L. Martin-Magniette, F. Bitton, L. Taconnat, S. Balzergue, S. Aubourg, J. P. Renou, A. Lecharny, and V. Brunaud. 2008. CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Res* 36: D986-90.

Galperin, M. Y., D. R. Walker, and E. V. Koonin. 1998. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res* 8: 779-90.

Garcia-Hernandez, M., T. Z. Berardini, G. Chen, D. Crist, A. Doyle, E. Huala, E. Knee, M. Lambrecht, N. Miller, L. A. Mueller, S. Mundodi, L. Reiser, S. Y. Rhee, R. Scholl, J. Tacklind, D. C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang. 2002. TAIR: a resource for integrated Arabidopsis data. *Funct Integr Genomics* 2: 239-53.

Ge, H., Z. Liu, G. M. Church, and M. Vidal. 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* 29: 482-6.

Gillis, J., and P. Pavlidis. 2013. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA). *BMC Bioinformatics* 14 Suppl 3: S15.

Glenisson, P., P. Antal, J. Mathys, Y. Moreau, and B. De Moor. 2003. Evaluation of the vector space representation in text-based gene clustering. *Pac Symp Biocomput:* 391-402.

Gollery, M., J. Harper, J. Cushman, T. Mittler, T. Girke, J. K. Zhu, J. Bailey-Serres, and R. Mittler. 2006. What makes species unique? The contribution of proteins with obscure features.

Genome Biol 7: R57.

Gollery, M., J. Harper, J. Cushman, T. Mittler, and R. Mittler. 2007. POFs: what we don't know can hurt us. *Trends Plant Sci* 12: 492-6.

Gomez, S. M., S. H. Lo, and A. Rzhetsky. 2001. Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics* 159: 1291-8.

Haas, B. J., J. R. Wortman, C. M. Ronning, L. I. Hannick, R. K. Smith, Jr., R. Maiti, A. P. Chan, C. Yu, M. Farzad, D. Wu, O. White, and C. D. Town. 2005. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol* 3: 7.

Hahn, A., J. Rahnenfuhrer, P. Talwar, and T. Lengauer. 2005. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics* 6: 112.

Han, J. D., N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430: 88-93.

Hanisch, D., A. Zien, R. Zimmer, and T. Lengauer. 2002. Co-clustering of biological networks and gene expression data. *Bioinformatics* 18 Suppl 1: S145-54.

Hanson, A. D., A. Pribat, J. C. Waller, and V. de Crecy-Lagard. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list--and how to find it. *Biochem J* 425: 1-11.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray. 1999. From molecular to modular cell biology. *Nature* 402: C47-52.

Heyndrickx, K. S., and K. Vandepoele. 2012. Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiol* 159: 884-901.

Higo, K., Y. Ugawa, M. Iwamoto, and H. Higo. 1998. PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res* 26: 358-9.

Hishigaki, H., K. Nakai, T. Ono, A. Tanigami, and T. Takagi. 2001. Assessment of prediction accuracy of protein function from protein--protein interaction data. *Yeast* 18: 523-31.

Horan, K., C. Jang, J. Bailey-Serres, R. Mittler, C. Shelton, J. F. Harper, J. K. Zhu, J. C. Cushman, M. Gollery, and T. Girke. 2008. Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol* 147: 41-57.

Horton, P., K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res* 35: W585-7.

Hu, Z., J. Mellor, J. Wu, T. Yamada, D. Holloway, and C. Delisi. 2005. VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res* 33: W352-7.

Huang, H., B. M. Jedynak, and J. S. Bader. 2007. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol* 3: e214.

Hulsen, T., M. A. Huynen, J. de Vlieg, and P. M. Groenen. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* 7: R31.

Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. 2009. InterPro: the integrative protein signature database. *Nucleic Acids Res* 37: D211-5.

Ideker, T., O. Ozier, B. Schwikowski, and A. F. Siegel. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 Suppl 1: S233-40.

Ilic, K., E. A. Kellogg, P. Jaiswal, F. Zapata, P. F. Stevens, L. P. Vincent, S. Avraham, L. Reiser, A. Pujar, M. M. Sachs, N. T. Whitman, S. R. McCouch, M. L. Schaeffer, D. H. Ware, L. D. Stein, and S. Y. Rhee. 2007. The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiol* 143: 587-99.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98: 4569-74.

Jansen, R., D. Greenbaum, and M. Gerstein. 2002. Relating whole-genome expression data with protein-protein interactions. *Genome Res* 12: 37-46.

Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature* 411: 41-2.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. 2000. The large-scale organization of metabolic networks. *Nature* 407: 651-4.

Jiang, Y., W. T. Clark, I. Friedberg, and P. Radivojac. 2014. The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. *Bioinformatics* 30: i609-16.

Jiang, Z., X. Liu, Z. Peng, Y. Wan, Y. Ji, W. He, W. Wan, J. Luo, and H. Guo. 2010. AHD2.0: an

update version of Arabidopsis Hormone Database for plant systematic studies. *Nucleic Acids Res* 39: D1123-9.

Kagaya, Y., and T. Hattori. 2009. Arabidopsis transcription factors, RAV1 and RAV2, are regulated by touch-related stimuli in a dose-dependent and biphasic manner. *Genes Genet Syst* 84: 95-9.

Karaoz, U., T. M. Murali, S. Letovsky, Y. Zheng, C. Ding, C. R. Cantor, and S. Kasif. 2004. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A* 101: 2888-93.

Kelley, R., and T. Ideker. 2005. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 23: 561-6.

Kerrien, S., B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 40: D841-6.

Khatri, P., S. Sellamuthu, P. Malhotra, K. Amin, A. Done, and S. Draghici. 2005. Recent additions and improvements to the Onto-Tools. *Nucleic Acids Res* 33: W762-5.

Kilian, J., D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. D'Angelo, E. BornbergBauer, J. Kudla, and K. Harter. 2007. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J* 50: 347-63.

King, A. D., N. Przulj, and I. Jurisica. 2004. Protein complex prediction via cost-based clustering. *Bioinformatics* 20: 3013-20.

Kolker, E., K. S. Makarova, S. Shabalina, A. F. Picone, S. Purvine, T. Holzman, T. Cherny, D. Armbruster, R. S. Munson, Jr., G. Kolesov, D. Frishman, and M. Y. Galperin. 2004. Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae. *Nucleic Acids Res* 32: 2353-61.

Lalonde, S., D. W. Ehrhardt, D. Loque, J. Chen, S. Y. Rhee, and W. B. Frommer. 2008. Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *Plant J* 53: 610-35.

Lan, H., R. Carson, N. J. Provart, and A. J. Bonner. 2007. Combining classifiers to predict gene function in Arabidopsis thaliana using large-scale gene expression measurements. *BMC*

Bioinformatics 8: 358.

Lanckriet, G. R., M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. 2004. Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput:* 300-11.

Lee, I., B. Ambaru, P. Thakkar, E. M. Marcotte, and S. Y. Rhee. 2010. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nat Biotechnol* 28: 149-56.

Lee, I., Y. S. Seo, D. Coltrane, S. Hwang, T. Oh, E. M. Marcotte, and P. C. Ronald. 2011. Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proc Natl Acad Sci U S A* 108: 18548-53.

Leinonen, R., F. G. Diez, D. Binns, W. Fleischmann, R. Lopez, and R. Apweiler. 2004. UniProt archive. *Bioinformatics* 20: 3236-7.

Letovsky, S., and S. Kasif. 2003. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* 19 Suppl 1: i197-204.

Licata, L., L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40: D857-61.

Lin D. 1998. An Information-Theoretic Definition of Similarity. Proc of Int Conf on Machine Learning (ICML). Morgan Kaufmann, Madison, Wisconsin, USA.

Lin, M., X. Shen, and X. Chen. 2011. PAIR: the predicted Arabidopsis interactome resource. *Nucleic Acids Res* 39: D1134-40.

Luscombe, N. M., M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431: 308-12.

Ma, S., S. Bachan, M. Porto, H. J. Bohnert, M. Snyder, and S. P. Dinesh-Kumar. 2012. Discovery of stress responsive DNA regulatory motifs in Arabidopsis. *PLoS One* 7: e43198.

Ma, S., and H. J. Bohnert. 2007. Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression. *Genome Biol* 8: R49.

Maciag, K., S. J. Altschuler, M. D. Slack, N. J. Krogan, A. Emili, J. F. Greenblatt, T. Maniatis, and L. F. Wu. 2006. Systems-level analyses identify extensive coupling among gene expression machines. *Mol Syst Biol* 2: 2006 0003.

Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. 1999a. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751-3.

Marcotte, E. M., M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg. 1999b. A combined algorithm for genome-wide prediction of protein function. *Nature* 402: 83-6.

Maslov, S., and K. Sneppen. 2002. Specificity and stability in topology of protein networks. *Science* 296: 910-3.

Maugis, C., G. Celeux, and M. L. Martin-Magniette. 2009. Variable selection for clustering with Gaussian mixture models. *Biometrics* 65: 701-9.

Médigue, C., S. Bocs, L. Labarre, C. Mathé, D. Vallenet. 2002. L'annotation in silico des séquences génomiques. Med Sci (Paris), Vol. 18, N° 2; p. 237-250 ; DOI :
10.1051/medsci/2002182237

Meyer, P. E., F. Lafitte, and G. Bontempi. 2008. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* 9: 461.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298: 824-7.

Moreno-Risueno, M. A., W. Busch, and P. N. Benfey. Omics meet networks - using systems approaches to infer regulatory networks in plants. *Curr Opin Plant Biol* 13: 126-31.

Mostafavi, S., and Q. Morris. 2010. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* 26: 1759-65.

Mostafavi, S., D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris. 2008. GeneMANIA: a realtime multiple association network integration algorithm for predicting gene function. *Genome Biol* 9 Suppl 1: S4.

Mushegian, A. R., and E. V. Koonin. 1996a. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* 93: 10268-73.

—. 1996b. Sequence analysis of eukaryotic developmental proteins: ancient and novel domains. *Genetics* 144: 817-28.

Mutwil, M., B. Usadel, M. Schutte, A. Loraine, O. Ebenhoh, and S. Persson. 2010. Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiol* 152: 29-43.

Nabieva, E., K. Jim, A. Agarwal, B. Chazelle, and M. Singh. 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21 Suppl 1:

i302-10.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48: 443-53.

Nehrt, N. L., W. T. Clark, P. Radivojac, and M. W. Hahn. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol* 7: e1002073.

Nepusz, T., H. Yu, and A. Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 9: 471-2.

Ning, K., H. K. Ng, S. Srihari, H. W. Leong, and A. I. 2010. Nesvizhskii. Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. *BMC Bioinformatics* 11: 505.

Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 96: 2896-901.

Pandey, G., C. L. Myers, and V. Kumar. 2009. Incorporating functional inter-relationships into protein function prediction algorithms. *BMC Bioinformatics* 10: 142.

Pandey, S. P., and I. E. Somssich. 2009. The role of WRKY transcription factors in plant immunity. *Plant Physiol* 150: 1648-55.

Parkinson, H., M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma. 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35: D747-50.

Pavlidis, P., J. Weston, J. Cai, and W. S. Noble. 2002. Learning gene functional classifications from multiple data types. *J Comput Biol* 9: 401-11.

Pavlovic, V., A. Garg, and S. Kasif. 2002. A Bayesian framework for combining gene predictions. *Bioinformatics* 18: 19-27.

Peng, W., J. Wang, J. Cai, L. Chen, M. Li, and F. X. Wu. 2014. Improving protein function prediction using domain and protein complexes in PPI networks. *BMC Syst Biol* 8: 35.

Pereira-Leal, J. B., A. J. Enright, and C. A. Ouzounis. 2004. Detection of functional modules from protein interaction networks. Proteins 54: 49-57.

Persson, S., H. Wei, J. Milne, G. P. Page, and C. R. Somerville. 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci U S A* 102: 8633-8.

Przulj, N., D. A. Wigle, and I. Jurisica. 2004. Functional topology in a network of protein

interactions. *Bioinformatics* 20: 340-8.

Radivojac, P., W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza,

Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Toronen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kassner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Honigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Bjorne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. Sternberg, N. Skunca, F. Supek, M. Bosnjak, P. Panov, S. Dzeroski, T. Smuc, Y. A. Kourmpetis, A. D. van Dijk, C. J. ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, et al. 2013. A large-scale evaluation of computational protein function prediction. *Nat Methods* 10: 221-7.

Raftery AE and Dean N. 2006. Variable selection for model-based clustering. Journal of the American Statistical Association, 101:168-178.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. 2002. Hierarchical organization of modularity in metabolic networks. *Science* 297: 1551-5.

Rhee, S. Y., and M. Mutwil. 2014. Towards revealing the functions of all genes in plants. *Trends Plant Sci* 19: 212-21.

Rider AK, et al. Proceedings of the 12th International Symposium on Intelligent Data Analysis (IDA 2013) Springer; 2013. Classifier evaluation with missing negative class labels; pp. 380-391.

Rives, A. W., and T. Galitski. 2003. Modular organization of cellular networks. *Proc Natl Acad Sci U S A* 100: 1128-33.

Robinson, M. D., J. Grigull, N. Mohammad, and T. R. Hughes. 2002. FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* 3: 35.

Rodriguez, R. and Redman, R. 2005. Balancing the generation and elimination of reactive oxygen species. Proc. Natl Acad. Sci. USA, 102, 3175–3176.

Rost, B. 2002. Enzyme function less conserved than anticipated. *J Mol Biol* 318: 595-608.

Ruepp, A., A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G.

Mannhaupt, M. Munsterkotter, and H. W. Mewes. 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* 32: 5539-45.

Ryngajllo, M., L. Childs, M. Lohse, F. M. Giorgi, A. Lude, J. Selbig, and B. Usadel. 2011. SLocX: Predicting Subcellular Localization of Arabidopsis Proteins Leveraging Gene Expression Data. *Front Plant Sci* 2: 43.

Sallet, E., J. Gouzy, and T. Schiex. EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics* 30: 2659-61.

Samanta, M. P., and S. Liang. 2003. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci U S A* 100: 12579-83.

Sass, S., F. Buettner, N. S. Mueller, and F. J. Theis. 2013. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Res* 41: 9622-33.

Schachter, V. 2002. Bioinformatics of large-scale protein interaction networks. *Biotechniques* Suppl: 16-8, 20-4, 26-7.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467-70.

Schietgat, L., C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Dzeroski. 2010. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics* 11: 2.

Schoner, D., S. Barkow, S. Bleuler, A. Wille, P. Zimmermann, P. Buhlmann, W. Gruissem, and E. Zitzler. 2007. Network analysis of systems elements. *Exs* 97: 331-51.

Schwikowski, B., P. Uetz, and S. Fields. 2000. A network of protein-protein interactions in yeast. *Nat Biotechnol* 18: 1257-61.

Sebastiani, F. 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47.

Segal, E., H. Wang, and D. Koller. 2003. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* 19 Suppl 1: i264-71.

Schaeffer SE. 207. Graph clustering. Comp Sci Rev; 1:27-64.

Shaik, R., and W. Ramakrishna. 2013. Genes and co-expression modules common to drought and bacterial stress responses in Arabidopsis and rice. *PLoS One* 8: e77261.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498-504.

Sharan, R., S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. 2005. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* 102: 1974-9.

Sharan, R., Ulitsky I and Shamir R. 2007. Network-based prediction of protein function. Mol Syst Biol 3, 88.

Shen-Orr, S. S., R. Milo, S. Mangan, and U. Alon. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* 31: 64-8.

Shoval, O., and U. Alon. 2010. SnapShot: network motifs. *Cell* 143: 326-e1.

Sigrist, C. J., L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, and N. Hulo. 2010. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 38: D161-6.

Simonis, N., J. van Helden, G. N. Cohen, and S. J. Wodak. 2004. Transcriptional regulation of protein complexes in yeast. *Genome Biol* 5: R33.

Small, I., N. Peeters, F. Legeai, and C. Lurin. 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4: 1581-90.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol* 147: 195-7.

Snel, B., G. Lehmann, P. Bork, and M. A. Huynen. 2000. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 28: 3442-4.

Spirin, V., and L. A. Mirny. 2003. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* 100: 12123-8.

Stark, C., B. J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* 39: D698-704.

Stuart, J. M., E. Segal, D. Koller, and S. K. Kim. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302: 249-55.

Suhre, K., and J. M. Claverie. 2004. FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res* 32: D273-6.

Swarbreck, D., C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang, and

Huala. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* 36: D1009-14.

Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 39: D561-8.

Tamames, J., G. Casari, C. Ouzounis, and A. Valencia. 1997. Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol* 44: 66-73.

Thomas, R. Gruber, Towards Principles for the Design of Ontologies Used for Knowledge Sharing in Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers, 1993.

Thomas, J., P.-E. Jouve, et N. Nicoloyannis. 2007. Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de données déséquilibrés. 3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique.

Tian, W., and J. Skolnick. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* 333: 863-82.

Tian, W., X. Dong, Y. Zhou, R. Ren. 2011. Predicting gene function using omics data : from data preparation to data integration. In Protein Function Prediction for Omics Era (Kihara, D., ed), pp. 215-242, Springer.

Titz, B., M. Schlesner, and P. Uetz. 2004. What do we learn from high-throughput protein interaction data? *Expert Rev Proteomics* 1: 111-21.

Tornow, S., and H. W. Mewes. 2003. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res* 31: 6283-9.

Troyanskaya, O. G., K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U S A* 100: 8348-53.

Tsuda, K., H. Shin, and B. Scholkopf. 2005. Fast protein classification with multiple networks. *Bioinformatics* 21 Suppl 2: ii59-65.

Usadel, B., T. Obayashi, M. Mutwil, F. M. Giorgi, G. W. Bassel, M. Tanimoto, A. Chow, D. Steinhauser, S. Persson, and N. J. Provart. 2009. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ* 32: 1633-51.

Valentini, G. 2011. True path rule hierarchical ensembles for genome-wide gene function prediction.

IEEE/ACM Trans Comput Biol Bioinform 8: 832-47.

Vardhanabhuti, S., J. Wang, and S. Hannenhalli. 2007. Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res* 35: 3203-13.

Varemo, L., F. Gatto, and J. Nielsen. 2014. Kiwi: a tool for integration and visualization of network topology and gene-set analysis. *BMC Bioinformatics* 15: 408.

Vazquez, A., A. Flammini, A. Maritan, and A. Vespignani. 2003. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* 21: 697-700.

Velankar, S., Y. Alhroub, C. Best, S. Caboche, M. J. Conroy, J. M. Dana, M. A. Fernandez Montecelo, G. van Ginkel, A. Golovin, S. P. Gore, A. Gutmanas, P. Haslam, P. M. Hendrickx, E. Heuson, M. Hirshberg, M. John, I. Lagerstedt, S. Mir, L. E. Newman, T. J. Oldfield, A. Patwardhan, L. Rinaldi, G. Sahni, E. Sanz-Garcia, S. Sen, R. Slowley, A. Suarez-Uruena, G. J. Swaminathan, M. F. Symmons, W. F. Vranken, M. Wainwright, and G. J. Kleywegt. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res* 40: D445-52.

Vidal, M., M. E. Cusick, and A. L. Barabasi. Interactome networks and human disease. *Cell* 144: 986-98.

Von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31: 25861.

Von Mering, C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417: 399-403.

Wachi, S., K. Yoneda, and R. Wu. 2005. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 21: 42058.

Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* 18: 1283-92.

Wang, H., H. Huang, and C. Ding. 2013. Function-function correlated multi-label protein function prediction over interaction networks. *J Comput Biol* 20: 322-43.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57-63.

Warde-Farley, D., S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris. 2010. The GeneMANIA prediction server: biological network

integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 38: W214-20.

Watson, J. D., and F. H. Crick. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171: 737-8.

Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393: 440-2.

Willis, R. C., and C. W. Hogue. 2006. Searching, viewing, and visualizing data in the Biomolecular Interaction Network Database (BIND). *Curr Protoc Bioinformatics* Chapter 8: Unit 8 9.

Wolfe, C. J., I. S. Kohane, and A. J. Butte. 2005. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics* 6: 227.

Wortman, J. R., B. J. Haas, L. I. Hannick, R. K. Smith, Jr., R. Maiti, C. M. Ronning, A. P. Chan, C. Yu, M. Ayele, C. A. Whitelaw, O. R. White, and C. D. Town. 2003. Annotation of the Arabidopsis genome. *Plant Physiol* 132: 461-8.

Wren, J. D. 2009. A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics* 25: 1694-701.

Wrobel, G., F. Chalmel, and M. Primig. 2005. goCluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics* 21: 3575-7.

Wu, X., and X. Qi. Genes encoding hub and bottleneck enzymes of the Arabidopsis metabolic network preferentially retain homeologs through whole genome duplication. *BMC Evol Biol* 10: 145.

Wuchty, S., A. L. Barabasi, and M. T. Ferdig. 2006. Stable evolutionary signal in a yeast protein interaction network. *BMC Evol Biol* 6: 8.

Wuchty, S., Z. N. Oltvai, and A. L. Barabasi. 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* 35: 176-9.

Xenarios, I., D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. 2000. DIP: the database of interacting proteins. *Nucleic Acids Res* 28: 289-91.

Xia, K., D. Dong, and J. D. Han. 2006. IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics* 7: 508.

Yamaguchi-Shinozaki, K., and K. Shinozaki. 2006. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu Rev Plant Biol* 57: 781-803.

Yang, C., E. Zeng, T. Li, and G. Narasimhan. 2005. Clustering genes using gene expression and text

literature data. *Proc IEEE Comput Syst Bioinform Conf:* 329-40.

Yeung, K. Y., M. Medvedovic, and R. E. Bumgarner. 2004. From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biol* 5: R48.

Yilmaz, A., M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold. 2011. AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Res* 39: D1118-22.

Yokoyama, K. D., U. Ohler, and G. A. Wray. 2009. Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Res* 37: e92.

Yoo, S. D., Y. H. Cho, G. Tena, Y. Xiong, and J. Sheen. 2008. Dual control of nuclear EIN3 by bifurcate MAPK cascades in C2H4 signalling. *Nature* 451: 789-95.

Yu, G., H. Zhu, C. Domeniconi, and M. Guo. 2015. Integrating multiple networks for protein function prediction. *BMC Syst Biol* 9 Suppl 1: S3.

Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa,

Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C.

Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. 2008. High-quality binary protein interaction map of the yeast interactome network. *Science* 322: 104-10.

Zaag, R., J. P. Tamby, C. Guichard, Z. Tariq, G. Rigaill, E. Delannoy, J. P. Renou, S. Balzergue, T. Mary-Huard, S. Aubourg, M. L. Martin-Magniette, and V. Brunaud. GEM2Net: from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response. *Nucleic Acids Res* 43: D1010-7.

Zhang, B., and S. Horvath. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4: Article17.

Zhang, B., S. Kirov, and J. Snoddy. 2005. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* 33: W741-8.

Zhang M-L, Zhou Z-H. 2013. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering.

Zhu, D., and Z. S. Qin. 2005. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics* 6: 8.

Zien, A., R. Kuffner, R. Zimmer, and T. Lengauer. 2000. Analysis of gene expression data with

pathway scores. *Proc Int Conf Intell Syst Mol Biol* 8: 407-17.

Zimmermann, P., L. Hennig, and W. Gruissem. 2005. Gene-expression analysis and network discovery using Genevestigator. *Trends Plant Sci* 10: 407-9

.

Annexes

I. Annexe des articles présentés dans ce travail de thèse

II. Annexe des posters présentés lors de conférences internationales (European Conference on Computational Biology 2014 et Tri National Arabidopsis Meeting 2014)

III. Annexe Inférence

Annexe A

Functional analysis of *Arabidopsis* immune-related MAPKs uncovers a role for MPK3 as negativeregulator of inducible defences

Nicolas Frei dit Frey, Ana Victoria Garcia, Jean Bigeard, Rim Zaag, Eduardo Bueso, Marie Garmier, Stéphanie Pateyron, Marie-Ludivine de Tauzia-Morea, Véronique Brunaud, Sandrine Balzergue, Jean Colcombet, Sébastien Aubourg, Marie-Laure Martin-Magniette and Heribert Hirt.

Genome Biology

RESEARCH Open Access

Functional analysis of *Arabidopsis* immune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defences

Nicolas Frei dit Frey[1,2†], Ana Victoria Garcia[1†], Jean Bigeard[1], Rim Zaag[1], Eduardo Bueso[1], Marie Garmier[3], Stéphanie Pateyron[1,4], Marie-Ludivine de Tauzia-Moreau[1], Véronique Brunaud[1], Sandrine Balzergue[1,4], Jean Colcombet[1], Sébastien Aubourg[1], Marie-Laure Martin-Magniette[1,5,6] and Heribert Hirt[1,7*]

Abstract

Background: Mitogen-activated protein kinases (MAPKs) are key regulators of immune responses in animals and plants. In *Arabidopsis*, perception of microbe-associated molecular patterns (MAMPs) activates the MAPKs MPK3, MPK4 and MPK6. Increasing information depicts the molecular events activated by MAMPs in plants, but the specific and cooperative contributions of the MAPKs in these signalling events are largely unclear.

Results: In this work, we analyse the behaviour of *MPK3*, *MPK4* and *MPK6* mutants in early and late immune responses triggered by the MAMP flg22 from bacterial flagellin. A genome-wide transcriptome analysis reveals that 36% of the flg22-upregulated genes and 68% of the flg22-downregulated genes are affected in at least one *MAPK* mutant. So far MPK4 was considered as a negative regulator of immunity, whereas MPK3 and MPK6 were believed to play partially redundant positive functions in defence. Our work reveals that MPK4 is required for the regulation of approximately 50% of flg22-induced genes and we identify a negative role for MPK3 in regulating defence gene expression, flg22-induced salicylic acid accumulation and disease resistance to *Pseudomonas syringae*. Among the *MAPK*-dependent genes, 27% of flg22-upregulated genes and 76% of flg22-downregulated genes require two or three MAPKs for their regulation. The flg22-induced MAPK activities are differentially regulated in *MPK3* and *MPK6* mutants, both in amplitude and duration, revealing a highly interdependent network.

Conclusions: These data reveal a new set of distinct functions for MPK3, MPK4 and MPK6 and indicate that the plant immune signalling network is choreographed through the interplay of these three interwoven MAPK pathways.

Background

Plants fend off most microbial attacks thanks to a multi-layered immune system, which is activated through the recognition of diverse microbial features. The first layer of induced defences relies on pattern recognition receptors (PRRs) that detect conserved microbe-associated molecular patterns (MAMPs) and initiate a defence program called pattern-triggered immunity (PTI). All known plant PRRs are located at the plasma membrane where they recognise and bind extracellular MAMPs [1]. The best studied example is FLS2 (flagellin-sensing 2), a receptor kinase with an extracellular leucine-rich repeat (LRR) domain that binds the conserved flg22 epitope derived from bacterial flagellin [2]. This recognition event induces immediate FLS2 association to the co-receptor BAK1 (BRI1-associated kinase 1) and their reciprocal kinase activation, which in turn initiates a series of responses important for defence activation [3]. In plants, MAMP perception induces early and late cellular processes, such as calcium fluxes, kinase cascades, production of reactive oxygen species (ROS), transcriptional reprogramming and reinforcement of the cell wall via deposition of callose [4].

* Correspondence: Heribert.Hirt@kaust.edu.sa
†Equal contributors
[1]Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165 - Université d'Evry Val d'Essonne - ERL CNRS 8196 - Saclay Plant Sciences, 2 rue Gaston Crémieux, Evry 91057, France
[7]Center for Desert Agriculture, 4700 King Abdullah University of Sciences and Technology, Thuwal 23955-6900, Saudi Arabia
Full list of author information is available at the end of the article

The importance of PTI was highlighted by the identification of pathogen effector molecules that target PTI components to suppress host defences and allow invasion [5]. Through the use of secretion systems, pathogens deliver a suite of effectors to the plant apoplast and intracellular compartments to modify the host cell to their benefit. As a counterpart, plants evolved intracellular receptors with nucleotide-binding and leucine-rich repeat domains (NB-LRR) that sense effectors and activate effector-triggered immunity (ETI) [5]. ETI is an amplified PTI response that results in disease resistance and is often associated with the accumulation of the hormone salicylic acid (SA) and a localised programmed cell death referred to as hypersensitive response (HR). While this response is efficient against biotrophic pathogens, necrotrophic pathogens that kill host cells are fought through activation of defences mediated by the hormones jasmonic acid (JA) and ethylene (ET) [6].

Following the detection of pathogens, MAPK cascades become activated and are central to the regulation of the immune system in animals and plants [7]. These conserved signalling modules are generally composed of a MAPKKK (MAPK kinase kinase), a MAPKK (MAPK kinase) and a MAPK, and function to translate extracellular stimuli into intracellular responses. In plants, MAPKs play important roles in different developmental processes and stress responses, but the far best studied examples are the roles of the MAPKs MPK3, MPK4 and MPK6 in disease resistance [7]. In Arabidopsis, flg22 recognition activates at least two MAPK signalling pathways. One of these MAPK cascades is defined by the MAPKKs MKK4 and MKK5, which act redundantly to activate the MAPKs MPK3 and MPK6 [8]. The second cascade activated by flg22 is defined by the MAPKKK MEKK1, which activates MKK1 and MKK2 that act redundantly on MPK4 [9,10]. It was recently shown that this cascade negatively regulates the MAPKKK MEKK2 (SUMM1) and the NB-LRR SUMM2, whose activation initiates defence responses [11-13]. As a consequence, the double mutant *mkk1 mkk2* and the single mutants *mekk1* and *mpk4* exhibit similar autoimmune phenotypes, such as dwarfism, cell death lesions, ROS accumulation and constitutive SA-mediated defences [9,10,14-16]. Furthermore, *mkk1 mkk2* and *mpk4* plants show enhanced resistance to the biotrophic pathogens *Hyaloperonospora arabidopsidis* and *Pseudomonas syringae* (*P. syringae*) and susceptibility to the necrotrophic fungi *Botrytis cinerea* (*B. cinerea*) (*mkk1 mkk2*) and *Alternaria brassicicola* (*mpk4*) [9,10,16,17]. These phenotypes are partially suppressed by the expression of the bacterial salicylate hydroxylase *NahG* or by mutations that impair SA accumulation [10,15,16]. Recently, the activity of a fourth MAPK, MPK11, was shown to be induced by flg22 and to play redundant functions with the other stress-induced MAPKs in embryo development, but no major function in disease resistance could be detected [18]. Besides induction of the MAPK activities, MAMP treatment also leads to transcript accumulation of *MPK11* and *MPK3* but not of *MPK4* and *MPK6* [19]. The importance of these protein kinases for immunity was further highlighted by the identification of pathogen effectors that target MAPK cascades. *P. syringae* encodes at least two effectors that reduce MAPK activation: the ADP-ribosyltransferase HopF2 that inactivates MKK5 and the phosphothreonine lyase HopAI1 that presumably dephosphorylates MPK3, MPK4 and MPK6 [13,20,21].

While *mpk4* has severe developmental defects, *mpk3* and *mpk6* single mutants resemble wild type plants and only the combination of both mutations impairs normal development. Indeed, MPK3 and MPK6 redundantly regulate stomatal development and the *mpk3 mpk6* double mutant is embryo lethal [22]. MPK3 and MPK6 are believed to be redundant also during plant immune responses, but increasing evidence points to additional independent functions. MPK3 and MPK6 phosphorylate and stabilise the ET biosynthetic enzymes ACS2 and ACS6 and thereby drive ET production in response to *B. cinerea* [23]. Furthermore, both kinases mediate the *B. cinerea*-induced phosphorylation of the transcription factors ERF6 and WRKY33, which in turn regulate defence gene expression and the accumulation of the antimicrobial compound camalexin, respectively [24-26]. In contrast to these redundant functions, MPK3 and MPK6 play different roles in the defence response to *B. cinerea.* While *mpk3* plants are more susceptible to *B. cinerea*, *mpk6* mutants show wild type susceptibility levels and are compromised in the elicitor-induced fungal resistance [25,27]. Furthermore, *mpk6* but not *mpk3* was shown to suppress exacerbated stress responses, as the enhanced resistance of mutants in the phosphatase MKP1 [28], the constitutive stress responses triggered by a dominant allele of the receptor-like wall associated kinase WAK2 (generated by a WAK2-cTAP fusion) [29], or the deregulated cell death triggered by fumonisin B1 [30,31]. MPK3 and MPK6 have also been proposed to play both redundant and distinct roles in the flg22-induced pathway [8,27,32]. Both *mpk3* and *mpk6* single mutants are defective in flg22-induced stomatal closure, a key defence step against pathogen entry into leaves [32]. In contrast, *mpk3* but not *mpk6*, shows increased responses to flg22 in terms of ROS production and growth inhibition [33]. In response to flg22, MPK3 and MPK6 regulate the transcription factors WRKY22 and WRKY29 [8], whereas the ET-related ERF104 is specifically targeted by MPK6 [34]. In agreement with these partially overlapping roles, the use of random peptide libraries and protein arrays suggested common and specific substrates for these immune related MAPKs,

among which numerous transcription factors are found [35-37].

All these data indicate that MPK3, MPK4 and MPK6 are key regulators of the transcriptional reprogramming in response to many stresses including MAMP perception. Nevertheless, no transcriptome analysis has been reported that would give insight into the genes controlled by these three MAPKs in response to flg22. Whereas *mpk4* adult plants show strong dwarfism, young *mpk4* seedlings display less severe developmental changes and therefore facilitate the phenotypic analyses of the mutant. In this work, we performed a comparative analysis of MPK3, MPK4 and MPK6 mutants for early (flg22-induced transcriptome changes and MAPK activities) and late (SA production, callose deposition and resistance against *P. syringae*) immune responses. By using a clustering approach based on the transcriptome analysis and a gene network construction method we were able to predict specific transcription factors involved in the flg22-induced transcriptional reprogramming modulated by the individual MAPKs. The analysis of the flg22-induced transcriptomes and the differential regulation of the MAPK activities in the *MAPK* mutants revealed extensive cooperative and inhibitory cross-talk between the MAPK signalling pathways. These analyses also identified new functions for MPK3 and MPK4. Although our and other groups have documented a negative role of MPK4 in the regulation of SA-mediated immunity [16,17,38], our present analysis revealed that MPK4 also functions as a positive regulator of early flg22-induced transcriptional reprogramming. Moreover, MPK3 was found to repress the constitutive and flg22-induced expression of defence genes, inhibit flg22-induced SA accumulation and resistance to *P. syringae*.

Results

General overview of the transcriptomes of *mpk3*, *mpk4* and *mpk6* in response to flg22

MPK3, MPK4, MPK6 and recently also MPK11 have been described to be rapidly and transiently activated in response to flg22 and other MAMP treatments [8,9,18]. MAPKs are important regulators of gene transcription in animals [39] and plants [7,36]. To identify genes regulated by the MAPKs in response to flg22, we performed a whole transcriptome analysis of Col-0, *mpk3*, *mpk4* and *mpk6* after mock or 30 min treatment with 1 μM flg22. We took advantage of the root developmental phenotype present in young *mpk4* seedlings to select for homozygous mutant plants from a segregating population (see Material and Methods). The *mpk11* mutant displays only minor and non-reproducible alterations in the flg22-induced transcriptional reprogramming [18] and was therefore not included in the analysis.

mpk3 and mpk4 display major and partially overlapping transcriptional changes under standard growth conditions

In control conditions, we observed 1,235 genes differentially expressed in *mpk4*, 496 genes in *mpk3* and only 61 genes in *mpk6* in comparison to Col-0 (Table 1, Additional file 1: Table S1 and Additional file 2: Table S2). As the variances of the expression differences were of the same order, we concluded that the transcriptome differences were not due to any experimental error and reflected the impact of the mutations. The reprogramming observed in *mpk4* included the specific upregulation of genes related to stress responses, cell death and SA production (Additional file 2: Table S2), in agreement with previous transcriptome data of adult plants [16]. Strikingly, approximately half of the genes that were differentially expressed in *mpk3* (51% of downregulated and 60% of upregulated genes) displayed a similar regulation in *mpk4* (Figure 1A and Additional file 3: Table S3). Within this group of 286 commonly regulated genes in *mpk3* and *mpk4*, genes controlling glucosinolate biosynthesis were upregulated and genes responding to sugar and amino acid metabolism were downregulated. In *mpk6*, 10 out of the 45 upregulated genes are chloroplast genes encoding regulators of photosynthesis and light reactions (Additional file 2: Table S2). As MAPKs are known to regulate MAMP-induced transcriptional responses, we wondered if the absence of one kinase would trigger basal changes in the transcriptome that resemble those triggered by flg22 treatment. Only a small subset of the differentially regulated genes in *mpk3*, *mpk4* and *mpk6* in mock-treated conditions was similarly regulated in Col-0 after a 30 min treatment with flg22 (Additional file 4: Figure S1). As expected, *mpk4* showed the highest overlap with the flg22-induced response but this represented only 24% of the differentially expressed genes (292 of 1,235 genes). This indicates that the basal transcriptome changes observed in these mutants do not mimic the transcriptional reprogramming triggered in Col-0 in response to flg22.

Table 1 Number of differentially expressed genes in the different transcriptome comparisons of the microarray analysis

Comparison	Genes up	Genes down
mpk3 vs Col	305	191
mpk4 vs Col	969	265
mpk6 vs Col	45	16
mpk3 + flg22 vs *mpk3*	1,519	877
mpk4 + flg22 vs *mpk4*	1,442	634
mpk6 + flg22 vs *mpk6*	1,468	690
Col + flg22 vs Col	1,529	862

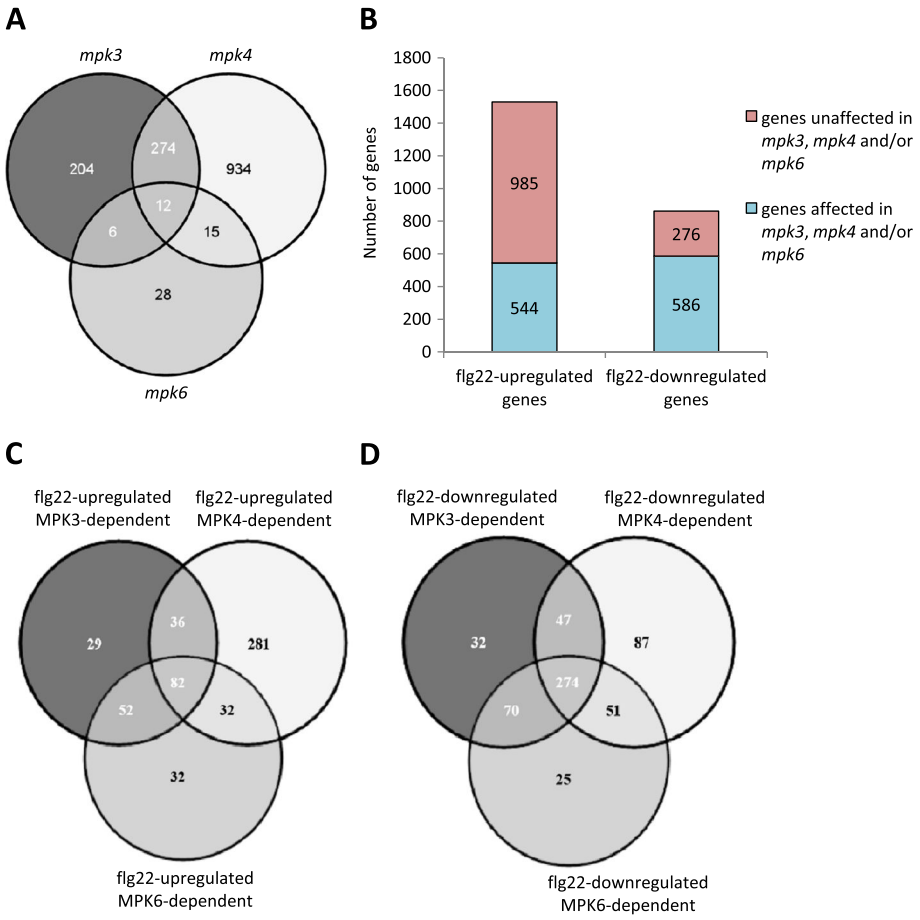


Figure 1 Transcriptome analysis of *mpk3*, *mpk4* and *mpk6*. **(A)** Venn diagram of the overlap between the differentially expressed genes (up- and downregulated) in control conditions in *mpk3*, *mpk4* and *mpk6* compared to Col-0. **(B)** Number of genes losing part of their flg22-regulation (at least 1 log ratio) in at least one *MAPK* mutant and number of genes not affected by *MAPK* mutations. **(C)** Venn diagram of the overlap between genes showing flg22-upregulation in Col-0 and affected in *mpk3*, *mpk4* or *mpk6*. **(D)** Venn diagram of the overlap between genes showing flg22-downregulation in Col-0 and affected in *mpk3*, *mpk4* or *mpk6*.

Identification of MAPK-dependent genes in the flg22-triggered transcriptional reprogramming

We next analysed the transcriptome changes in response to a 30 min treatment with flg22. Col-0 reacted to the flg22 treatment with the upregulation of 1,529 genes, enriched in GO terms involved in signalling, enzymatic functions and with membrane or cell periphery targeting, in agreement with a coordinated response to an extracellular pathogen-derived signal (Additional file 5: Table S4). The downregulated genes (962 genes) showed enrichment in genes involved in hormone metabolism and signalling, RNA metabolism, transcription and response to sugar and were targeted to different subcellular compartments (Additional file 5: Table S4). These observations are in line with previous analyses of the transcriptional responses of wild type plants to flg22 [4,40,41].

We then assessed the proportion of genes that loose totally or partially their flg22-dependent regulation in the *MAPK* mutants by assessing the number of genes with at least 1 log ratio difference in the flg22-induced expression in each *MAPK* mutant as compared with Col-0 (Table 1 and Additional file 6: Table S5). We observed that 36% of the flg22-upregulated genes and 68% of the flg22-downregulated genes were affected in at least one *MAPK* mutant (Figures 1B). This revealed that besides the known role of *MPK3*, *MPK4* and *MPK6* in regulating gene induction, the three kinases have a major role in flg22-induced gene repression. As 82% of the *MAPK*-dependent flg22-induced genes and 93% of the *MAPK*-dependent flg22-repressed genes are not affected in *mpk3*, *mpk4* or *mpk6* in control conditions (Additional file 4: Figure S1) the reduction in the flg22 response cannot be explained by the basal transcriptome changes observed in the mutants in the absence of stress.

MPK4 positively regulates flg22-upregulated genes

We next assessed the contribution of each *MAPK* to the observed flg22-induced gene upregulation (Additional

file 5: Table S4). Within the *MAPK*-dependent flg22-upregulated genes, the majority of genes (63%) showed differential regulation in only one kinase mutant and strikingly, we found that 52% (281/544 genes) are differentially regulated only in *mpk4* (Figure 1C, Additional file 6: Table S5). Only 24% of these *MPK4*-regulated genes can be attributed to a basal upregulation in untreated *mpk4* (Additional file 4: Figure S1A and Additional file 7: Figure S2). This subset of upregulated genes in untreated *mpk4* was enriched in GO terms associated with immune responses, cell death, SA, JA and ROS (Additional file 8: Figure S3D). Interestingly, many of the genes showing reduced flg22 induction in *mpk4* and not affected in mock-treated samples were associated to ET biosynthesis and signalling (Additional file 8: Figure S3B), which points to a positive role of *MPK4* in mediating the flg22-induced transcriptional reprogramming of the ET pathway. This observation fits with previous data suggesting a role for *MPK4* in mediating the induction of the JA- and ET-responsive gene *PDF1.2* in response to *P. syringae* effectors or hormone treatments [16,17,42]. The group of flg22-induced *MPK4*-dependent genes encode important regulators of plant defence such as the cell death inhibitor BAP1 [43], the calcium-dependent protein kinase CPK5 [44], the exocyst complex subunit EXO70B2 [45], the cyclic nucleotide-gated ion channel CNGC11 [46] or the BAK1-like receptor kinase BKK1/SERK4 [47].

MPK3, MPK4 and MPK6 equally contribute to regulate flg22-downregulated genes

We next analysed the participation of each *MAPK* in the regulation of the *MAPK*-dependent flg22-repressed genes. We found that 47% (274/586 genes) of the flg22-repressed genes lose partially or completely their regulation in *mpk3*, *mpk4* and *mpk6* (Figure 1D and Additional file 6: Table S5). Among these flg22-repressed *MAPK*-dependent genes we found enrichment in GO terms related to sugar response and the metabolism of the branched-chain amino acids (BCAA) leucine, isoleucine and valine. This group of genes is highly co-regulated and, in some cases, constitutively repressed in *mpk3* and *mpk4* (Additional file 3: Table S3 and Additional file 6: Table S5). Interestingly, isoleucine and other BCAA-related metabolic products are involved in the homeostasis of defence hormones, as for example isoleucic acid (ILA) that induces the SA pathway and resistance against *P. syringae* [48]. This highlights a role for the three *MAPKs* in repressing primary metabolic pathways in response to flg22 that may impact hormone metabolism. In addition, *mpk4* showed specific deregulation of a subset of genes involved in rRNA metabolism and in regulating transcriptional responses during morphogenesis, hormone responses and circadian rhythm, (Additional file 6: Table S5).

Functional analysis of transcriptome data through clustering of co-regulated genes

The differential analysis performed on the transcriptome data identified genes with statistically significant differential expression but did not reveal whether the genes are regulated by common regulators. To identify genes that behave similarly across the seven comparisons, we performed a co-expression analysis based on model-based clustering. Hereby, genes with similar expression patterns are grouped in clusters that may share a similar regulatory protein or mechanism. In contrast to a clustering method based on a metric distance, like K-means or hierarchical clustering, model-based clustering assumes that the data are generated by a finite mixture of distributions. Hence, the clustering is done with a global point of view and provides a statistically rigorous framework to determine the cluster numbers and the gene assignments [49]. For this analysis, we considered probes that were differentially expressed in at least one of the seven comparisons according to the Bonferroni *P* value adjustment to limit the number of false positives. This corresponded to a total of 4,378 probes representing 4,177 genes. The clustering method found 29 clusters of co-expression and 1,928 probes corresponding to 1,876 genes were assigned into the clusters after a classification based on a threshold Maximum A Posteriori rule (Additional file 9: Figure S4 and Additional file 10: Table S6). For the biological interpretation of the analysis, for each cluster and for each comparison, the percentage of genes differentially expressed according to Bonferroni is indicated on the top of the cluster profiles. We considered for further analysis those clusters where more than 50% of the genes were differentially expressed in at least one comparison. For this reason, clusters 1 and 17 were excluded from the interpretation.

We obtained three major groups of clusters: (1) clusters with genes modulated by flg22 in Col-0 (15 clusters); (2) clusters with genes affected only in *mpk4* under standard growth conditions (6 clusters); and (3) clusters showing differential expression in the *MPK3*, *MPK4* or *MPK6* mutants but not in Col-0 (6 clusters). Four representative clusters with interesting profiles are shown in Figure 2. Among the 15 clusters with flg22-regulated genes in Col-0, 10 group flg22-induced genes (Clusters 2, 5, 10, 11, 15, 18, 20, 22, 28, 29) and five group flg22-downregulated genes (Clusters 3, 19, 21, 25, 26). We then found five (Clusters 4, 6, 7, 12, 27) and one (Cluster 9) clusters that are defined by genes up- and downregulated, respectively, only in *mpk4* in standard growth conditions. We wondered whether those differentially regulated genes in *mpk4* could be regulated by flg22 at other time points not analysed in our study. To assess this, we made use of a microarray analysis performed in similar conditions as ours, which identified genes differentially expressed in Col-0 seedlings at 1 and 3 h after treatment with 1 μM

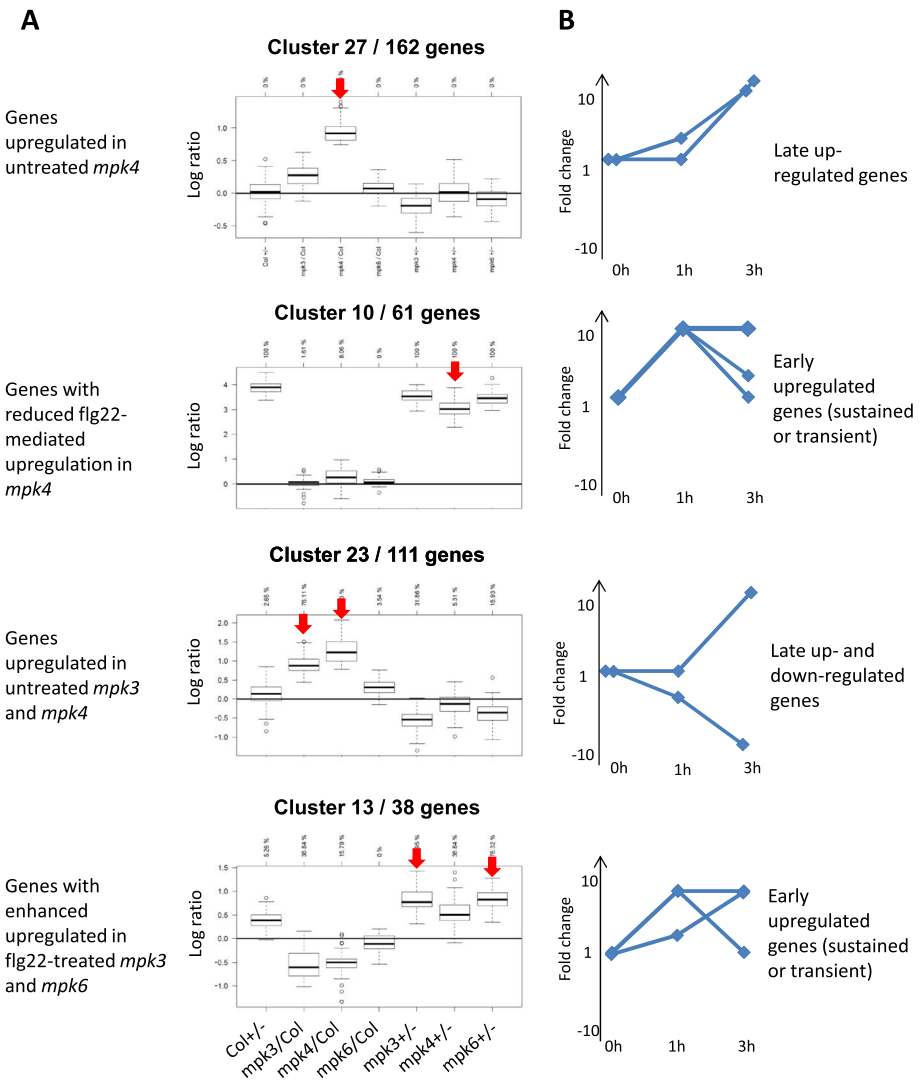


Figure 2 Selection of clusters displaying interesting gene co-regulation. (A) Profiles of the indicated clusters displaying the fold change expressions in the seven comparisons. The sentences on the left and the red arrows indicate the main message of the cluster. Y scale is log ratio. On the X scale, the following comparisons are shown: Col-0 + flg22 *vs.* Col-0, *mpk3 vs.* Col-0, *mpk4 vs.* Col-0, *mpk6 vs.* Col-0, *mpk3* + flg22 *vs. mpk3*, *mpk4* + flg22 *vs. mpk4*, *mpk6* + flg22 *vs. mpk6*. Profiles are represented as boxplots, where the bottom and top of the box are the first and third quartiles and the band inside the box is the median. Data not included between the whiskers are represented by a dot. On top is indicated the percentage of genes differentially regulated (*P* value <0.05) in the different comparisons. '-' indicates mock treatment and '+' indicates flg22 treatment. **(B)** Kinetic behaviour of the cluster as predicted by the comparison with 'late' flg22-regulated genes.

flg22 [40]. Hereafter, we refer to these genes as 'late' flg22-regulated genes, which we grouped into different classes according to their expression kinetics (Figure 2 and Additional file 11: Figure S5). Out of the five clusters showing upregulation only in *mpk4*, three clusters (Clusters 6, 12, 27) showed enrichment in late flg22-induced genes (Figure 2 and Additional file 12: Figure S6). The genes of these three clusters and those of cluster 11 (flg22-upregulated genes induced in untreated *mpk4*) were also induced at 1 and 3 h, suggesting that the genes induced in untreated *mpk4* correspond to late flg22-upregulated genes. In contrast, clusters 10, 18, 20 and 29 group flg22-upregulated genes in Col-0 that loose part of their flg22-regulation in *mpk4* and are not differentially regulated in untreated *mpk4* (Additional file 13: Figure S7). Interestingly, these clusters were strongly enriched for early induced genes (Figures 2 and 3).

Clusters 8, 13, 14, 16, 23 and 24 showed a differential regulation in the MAPK mutants with or without flg22 treatment while they were not regulated by flg22 in Col-0. Specifically, the clustering approach revealed one cluster with genes upregulated in untreated *mpk3* (Cluster 14) and, as observed in the differential analysis,

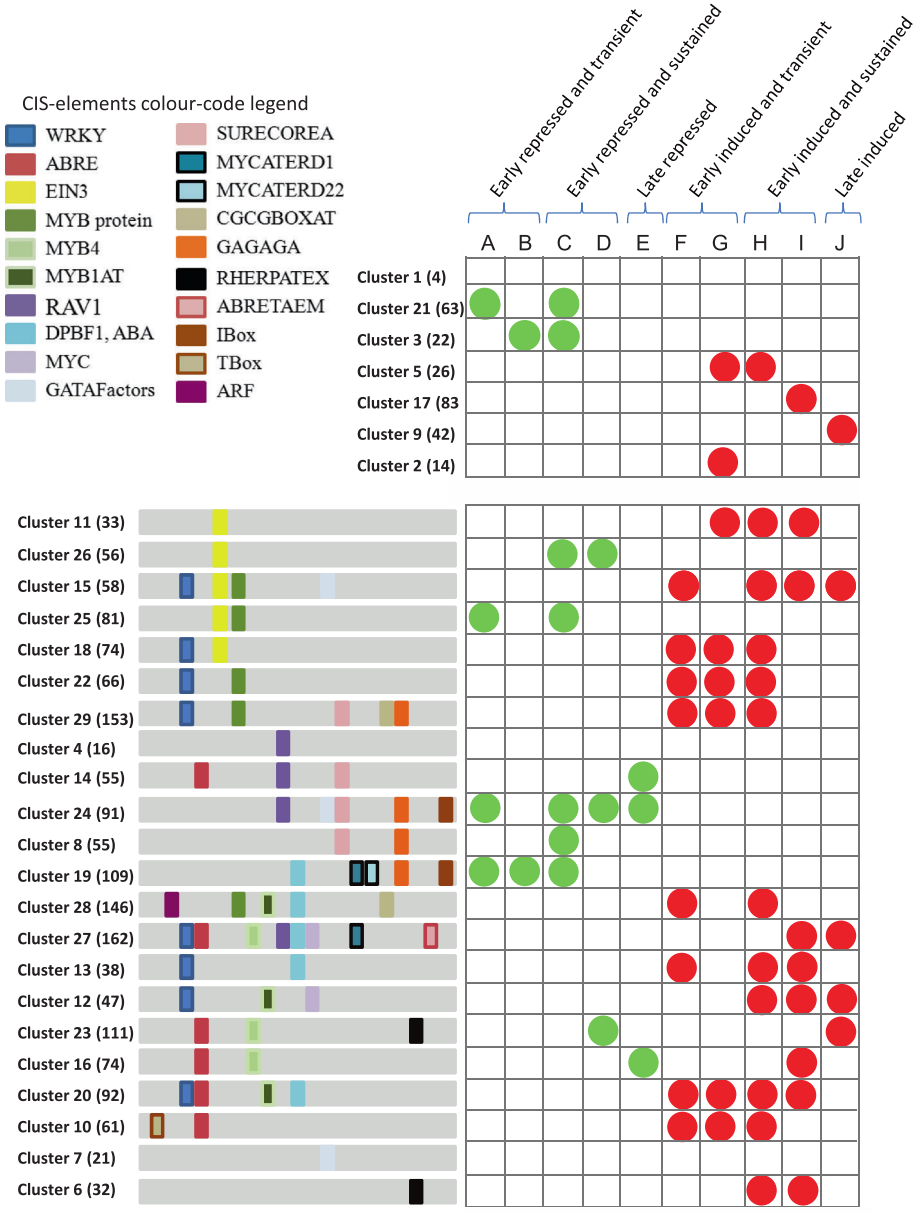


Figure 3 Representation of the cis-element enrichment and the expression kinetics in the different clusters. The left panel shows clusters sharing cis-elements. The cluster number and the number of genes of each cluster (in brackets) are shown on the left. The right panel represents the kinetic behaviour of the clusters as predicted by the comparison with publicly available data on 'late' flg22-regulated genes. Class F genes are induced after 1 h flg22 treatment and return to basal level after 3 h. Class G genes are induced after 1 h flg22 treatment and induced to a lesser extent at 3 h after treatment. Class H genes are induced after 1 h flg22 treatment and to a same extent at 3 h. Class I genes are induced after 1 h flg22 treatment and show further induction at 3 h. Class J genes are not induced at 1 h but show upregulation 3 h after flg22 treatment. Class A, B, C, D, E show the same tendencies but with flg22-induced downregulation. No cis-element enrichment was found for clusters 1, 2, 3, 5, 9, 17, 21. Clusters 1, 4 and 7 do not show significant enrichment in 'late' flg22-regulated genes. For detailed explanation, see Additional file 11: Figure S5.

two clusters with similar regulation in untreated *mpk3* and *mpk4*: cluster 23 displays upregulated genes and cluster 16 downregulated genes. Therefore, the clustering approach confirmed the transcriptional similarities between unchallenged *mpk3* and *mpk4* already observed in the differential analysis. Cluster 23 grouped 86 of the 183 genes commonly upregulated in *mpk3* and *mpk4* and cluster 16 grouped 42 of the 103 genes commonly downregulated in the two *MAPK* mutants (Additional file 3: Table S3). Network building of known co-expressed genes in the cluster 23 revealed genes involved in flavonol metabolism and genes coding for enzymes that control glucosinolate production (Additional file 14: Figure S8 and Additional file 10: Table S6). The genes controlling BCAA metabolism

were found within cluster 16 of downregulated genes in *mpk3* and *mpk4* (Additional file 15: Figure S9 and Additional file 10: Table S6). Interestingly, three clusters contained genes that are not affected in flg22-treated Col-0 samples but are differentially expressed in one of the *MAPK* mutants after flg22 treatment. Cluster 8 shows specific downregulation in flg22-treated *mpk6*, cluster 24 downregulation in flg22-treated *mpk3* (and to a lesser extent *mpk4* and *mpk6*) and cluster 13 flg22-triggered upregulation in both *mpk3* and *mpk6* (and *mpk4* to a lesser extent). Therefore, these clusters contain genes whose expression is maintained unchanged upon flg22 treatment due to the concerted action of these three kinases. Interestingly, the genes contained in these clusters are similarly regulated in Col-0 at the later time points of 1 or 3 h after flg22 treatment (Figure 4 and Additional file 16: Figure S10). This suggests that the MAPKs are not only required to regulate the rapid transcriptional responses to flg22 treatment, but also to prevent premature regulation of flg22-responsive genes.

Analysis of cis-elements

The clustering approach allows the identification of genes showing similar expression patterns, thus putatively controlled by the same upstream regulators. Therefore, we used the promoters of the genes in each cluster to assess the enrichment of known cis-elements [50]. Additional file 17: Table S7 compiles all data concerning the cis-element enrichment analysis, the size of the clusters and the number of genes that presents each given cis-element. A schematic representation of this analysis is presented in Figures 3 and 4. Many clusters containing flg22-induced genes contain promoters with enrichment in W-boxes bound by WRKY transcription factors. Several WRKY transcription factors are known MAPK substrates and play important roles in stress-related transcriptional reprogramming in plants [51]. Our comparative analysis with the kinetic data from Denoux et al. [40] also indicates that six out of the eight clusters enriched for early and transiently flg22-activated genes contain W-boxes, suggesting that these transcription factors function in the early phases of flg22-mediated gene regulation. MYB binding sites were mostly present in clusters containing flg22-induced genes and were associated in three out of five clusters with WRKY binding sites (Clusters 15, 22 and 29). Until now, MYB51 is the only transcription factor of this family that has been reported to have a role in MAMP-triggered immunity [52], while other MYB factors are involved in different plant defence mechanisms [53]. Interestingly, protein microarrays identified several WRKY and MYB factors as putative targets of MPK3, MPK4 and MPK6 [36]. We found binding sites of the ET-related transcription factor EIN3 enriched in five clusters containing genes either up- or downregulated by flg22. This factor is regulated through direct phosphorylation by MPK3 and MPK6 and plays an important role in the transcriptional control of immune signalling components such as *SID2* and *FLS2* [54-56]. The classical ABA-responsive ABRE motifs were poorly associated with flg22 transcriptional regulation, but we found ABRE-related binding sites called DPBF1-binding elements that were associated with many clusters responding to flg22. DPBF1 belongs to the A-group of bZIP transcription factors, which are mostly related to ABA signalling [57], and

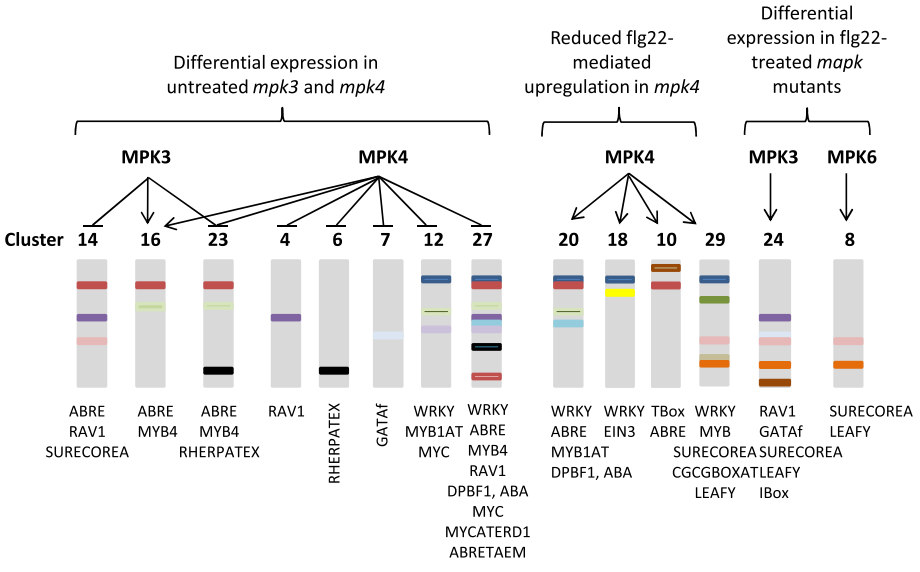


Figure 4 Cooperative and Specific and roles of MPK3, MPK4 and MPK6 as revealed by cluster analysis. Under standard growth conditions or in response to flg22, different MAPKs appear to play specific roles in the control of distinct clusters. This regulation may be under the control of the indicated enriched CIS-elements.

DPBF2 was found in protein microarrays as putative MPK4 and MPK6 substrates [36]. Among the clusters displaying genes differentially regulated in *mpk3* and *mpk4*, we found enrichment in a motif bound by RAV1, an AP2/EREBP transcription factor. Interestingly, these clusters are not regulated by flg22 in Col-0 and they group genes that are upregulated in untreated *mpk3* (Cluster 14) or *mpk4* (Clusters 4 and 27) or that present an altered expression pattern in flg22-treated *mpk3* (Cluster 24). The activity of this transcription factor is induced by mechanical stimuli [58] and may therefore be negatively controlled by *MPK3* and *MPK4*. Overall, the clustering approach coupled with the cis-element analysis allowed us to predict the function of specific transcription factors in the regulation of *MAPK*-dependent genes upon flg22 treatment.

Construction of gene interaction networks and identification of putative regulators of the MAPK-dependent transcriptional reprogramming

We then capitalised on the transcriptome data to build a gene network that is based on publicly available experimental data and displaying validated transcription factor-target and protein-protein interactions (see Material and Methods) (Additional file 18: Figure S11). This approach allows the identification of regulators which are not transcriptionally regulated and are therefore not identified through conventional transcriptome analysis. The absence of transcriptional regulation in response to flg22 is not a detrimental criterion for biological relevance, as rapid transcriptional reprogramming responses are usually controlled by preformed factors that are post-translationally regulated. Among the different hubs we found known regulators of immune responses such as the calmodulin-like protein CML9 [59] (Additional file 18: Figure S11B) and the calcium-dependent kinase CPK11 [60] (Additional file 18: Figure S11C). CML9 belongs to cluster 29 of flg22-upregulated genes and previous protein microarray experiments showed that it interacts with three other proteins of this cluster: two leucine-rich repeat protein kinases (AT3G02880 and AT3G28450) and the cytoplasmic kinase CAST-AWAY/KIN4 (AT4G35600) [61,62]. The protein microarray also showed that these three CML9-interacting proteins share seven common interacting proteins, which are all calmodulin-like proteins [61]. Interestingly, the kinase CAST-AWAY/KIN4 is a putative MPK6 phosphorylation substrate [36], suggesting the existence of a highly interconnected network related to Ca^{2+} signalling that may be regulated by MAPKs during immune responses. The transcription factors HY5, PIF1 and AP2 involved in different aspects of plant development were also revealed as hubs (Additional file 18: Figure S11). Indeed, HY5 and PIF1 are important regulators of the transcriptional reprogramming that occurs during light-regulated processes, which are primarily regulated post-transcriptionally in response to light [63,64]. Analysis of the genes that are connected to HY5 and may constitute transcriptional targets, positioned HY5 upstream of the clusters 10 and 20 of early flg22-regulated genes modulated by *MPK4*. A similar analysis placed PIF1 upstream of cluster 29 together with several other transcription factors.

Compensatory mechanisms at the level of MAPK protein activity occur in *mpk3* and *mpk6*

The complex relationships between the *MAPKs* revealed by the transcriptome analysis suggested that the absence of one MAPK could influence the function of the other MAPKs. We therefore analysed the flg22-induced MAPK activities in Col-0, *mpk3*, *mpk4* and *mpk6* using an anti-pTpY antibody that recognises the dual phosphorylated activation loop of MAPKs (TEY motif). Interestingly, *mpk3* showed higher and longer activation of MPK4 and MPK6 in response to flg22 treatment, whereas *mpk6* displayed higher and longer activation of MPK3 and MPK4 (Figure 5A). In contrast, despite the increased flg22-induced MPK3 and MPK6 activities observed in mutant plants of the upstream kinase MEKK1 [15], the absence of *mpk4* did not have an impact on the flg22-induced MPK3 or MPK6 activities (Figure 5B). To further assess the impact of *mpk4*, we generated and analysed *mpk3 mpk4* and *mpk6 mpk4* double mutants. Interestingly, while the double mutant plants resembled *mpk4* phenotypically (Additional file 19: Figure S12), they showed flg22-induced MAPK activities that resembled the respective *mpk3* and *mpk6* single mutants (Figure 5B). We concluded that MPK4 does not influence the regulation of MPK3 and MPK6 activities. The *mpk3 mpk6* double mutant was not included in this analysis due to its embryo lethal phenotype [22]. In all experiments, mock-treated samples showed no signal or only weak MPK6 activity (data not shown). Importantly, the observed differential regulation of the kinase activities were not due to different MAPK protein levels (Additional file 20: Figure S13). These results indicate that MPK3 regulates MPK4 and MPK6 activities whereas MPK6 regulates MPK3 and MPK4 activities. The observed regulation could be accomplished through a direct MAPK-MAPK phosphorylation event that could negatively regulate the MAPK activities and therefore explain an enhanced kinase activity in the mutant backgrounds. To assess this hypothesis, we tested if wild type and constitutively active versions of MPK3, MPK4 and MPK6 [38] could directly phosphorylate kinase-dead versions of MPK3 or MPK6. Whereas all wild type and constitutively active MAPK proteins showed kinase activity on the MBP substrate (Additional file 21: Figure S14A) and displayed autophosphorylation (Additional file 21: Figure S14B), none

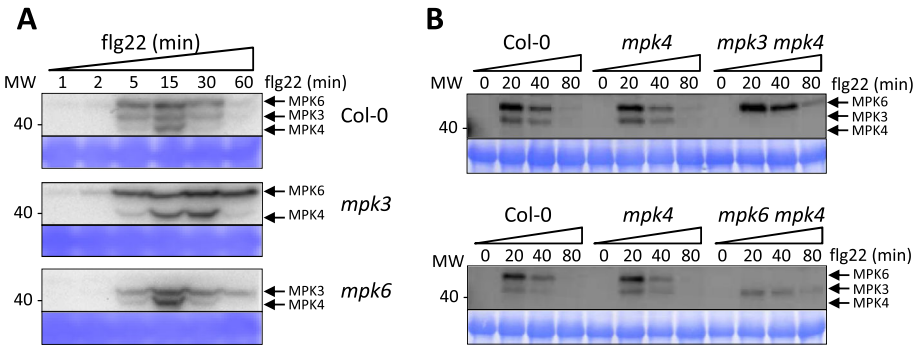


Figure 5 Flg22-induced activation of MPK3, MPK4 and MPK6 in Col-0, *mpk3*, *mpk4* and *mpk6* single and in *mpk3 mpk4* and *mpk6 mpk4* double mutant plants. Western blot analysis of Col-0, *mpk3* and *mpk6* plants **(A)** and Col-0, *mpk4*, *mpk3 mpk4* and *mpk6 mpk4* plants **(B)** at the indicated time-points after flg22 treatment, using an anti-pTpY antibody to detect activated MPK3, MPK4 and MPK6. The arrows indicate the activated forms of MPK3, MPK4 and MPK6. Blots were stained with Coomassie blue and the protein band corresponding to the RuBisCO large subunit shows equal loading.

of the MAPKs phosphorylated the dead versions of MPK3 or MPK6 in *in vitro* kinase assays (Additional file 21: Figure S14B). We concluded that MPK3 and MPK6 regulate MAPK activities through an indirect mechanism that may involve upstream kinases or phosphatases. As we did not detect any phosphorylation of the dead MPK6 or dead MPK3 by the respective wild type or constitutively active kinase versions, these data show that MAPK autophosphorylation is caused by intramolecular and not intermolecular phosphorylation.

Flg22-induced SA production is enhanced in *mpk3*

mpk4 displays constitutive activation of the SA pathway and enhanced resistance to biotrophic pathogens [16]. Since *mpk3* and *mpk4* transcriptomes presented a significant overlap in unchallenged conditions, we tested whether the SA pathway was also activated in *mpk3*. Contrary to *mpk4*, we observed no differential expression in genes involved in SA biosynthesis and signalling in either control or flg22-treated *mpk3* plants (that is, *SID2*, *EDS1*, *PAD4*, *ACD6*, *NDR1*, *ALD1*, *EDS5* and *NPR1* [65]). However, flg22-mediated induction of *SID2* occurs at later time points and is not the rate limiting step in flg22-induced SA accumulation [40,66]. To assess whether *MPK3* could play a role in the regulation of the SA pathway, we quantified SA accumulation in resting and flg22-challenged seedlings in a similar way as for the expression analyses. Flg22 leaf infiltration induces a 10-fold increase in SA levels in adult Arabidopsis plants [66], but the flg22-induced SA accumulation in Arabidopsis seedlings has not been reported yet. In response to 1 μM flg22 treatment, Arabidopsis wild type seedlings displayed a 10- to 20-fold increase in total SA (Figure 6A and 6B). Whereas *mpk4* seedlings displayed enhanced SA accumulation in resting conditions and higher flg22-induced SA levels compared to Col-0 (Figure 6A), *mpk3* showed wild type levels of SA in resting conditions but enhanced flg22-induced SA levels 24 h after treatment (Figure 6B). The increased SA accumulation correlated with the moderately increased flg22-induced transcript levels of the SA biosynthetic gene *SID2* and the SA signalling genes *EDS1* and *PAD4* (Figure 6C). On the contrary, the JA and ET marker gene *PDF1.2* displayed reduced flg22-induced accumulation in *mpk3*. These results suggest that the observed similarities in the transcriptomes of unchallenged *mpk3* and *mpk4* are not due to a similar deregulation of the SA pathway. We concluded that in contrast to the role of MPK4 that regulates constitutive and inducible SA levels, MPK3 has a role in dampening the SA pathway upon perception of MAMPs and possibly pathogens.

Flg22-induced callose deposition is reduced in *mpk3*

We next analysed the *MAPK* mutants for callose deposition, another defence hallmark induced by flg22 and proposed to be regulated by MAPK cascades. Previous studies on transgenic plants expressing the *P. syringae* effector HopAI1 or $MKK5^{DD}$, which, respectively, inactivate and activate both MPK3 and MPK6, suggested that these two MAPKs are necessary for callose accumulation [21]. In addition, *mekk1* plants show constitutive callose deposition [15]. In agreement with this, we found that *mpk4* plants also display constitutive callose accumulation (Figure 7A). The transcriptome analysis showed that *mpk3* and *mpk4* share the upregulation of certain genes involved in the indole glucosinolate-dependent production of callose (Additional file 14: Figure S8). We therefore tested if MPK3 could be involved in the regulation of this inducible defence mechanism. In contrast to *mpk4*, *mpk3* and *mpk6* displayed no constitutive callose accumulation (Figure 7A), but infiltration of flg22 into leaves of adult plants led to a reduced number of callose deposits in *mpk3* whereas *mpk6* had an intermediate behaviour between wild type and *mpk3* (Figure 7B).

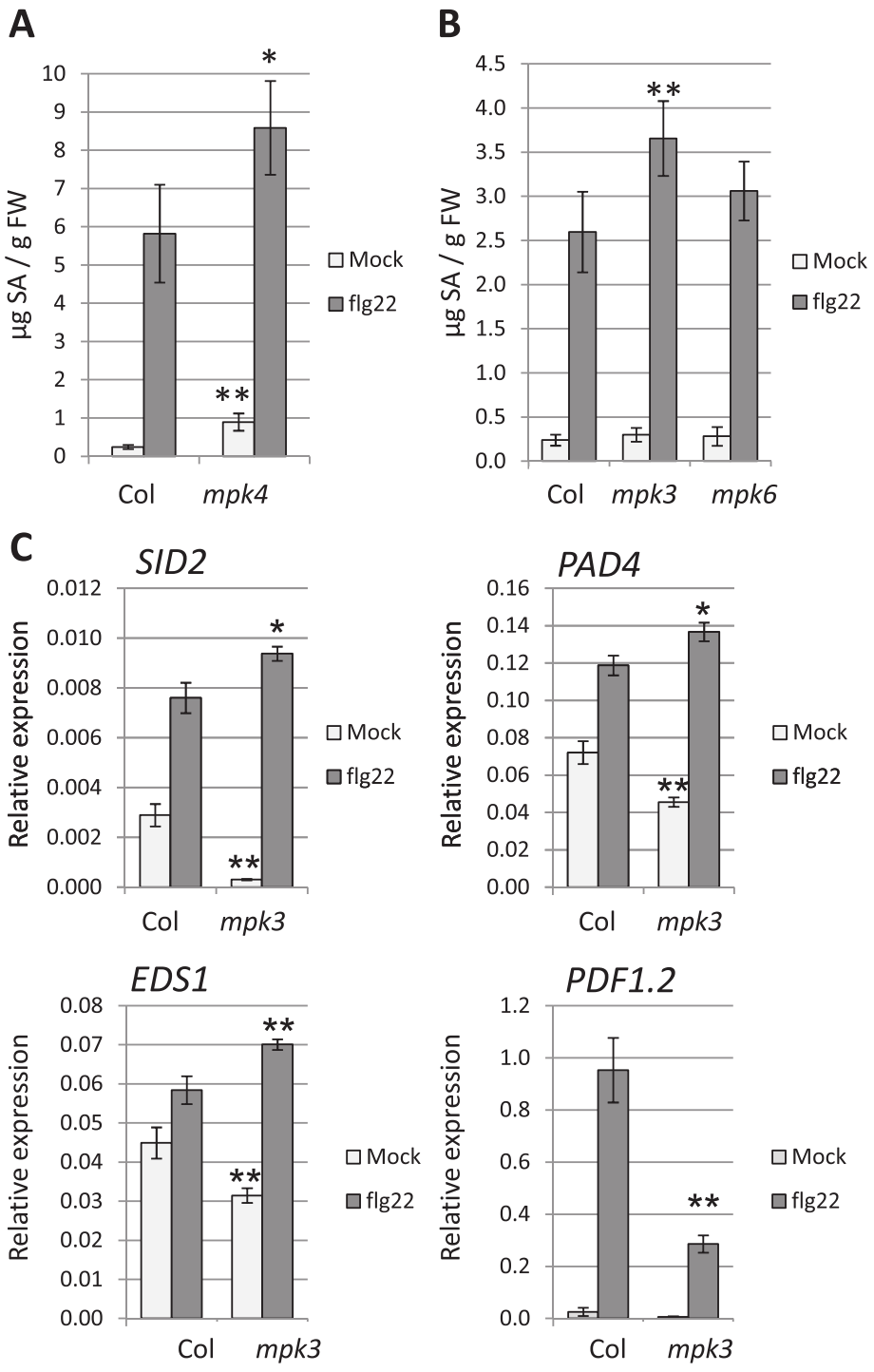


Figure 6 *mpk3* shows enhanced flg22-mediated activation of SA-mediated defences. SA accumulation 24 h after mock-treatment or treatment with 1 µM flg22 in Col-0 and *mpk4* **(A)** and in Col-0, *mpk3* and *mpk6* seedlings **(B)**. Bars are means ± SD. **(C)** qPCR analysis of the expression of SA marker genes *SID2*, *EDS1* and *PAD4* and the ET/JA marker gene *PDF1.2* in Col-0 and *mpk3* seedlings 24 h after treatment with 1 µM flg22. Transcript accumulation is expressed relative to the reference gene *ACTIN2*. Bars represent means ± SE of three independent biological replicates and each replicate is composed of three technical replicates. Stars indicate significant difference with Col-0 under the same conditions based on a two-tailed Student's ttest. ***P* value <0.01 and **P* <0.05.

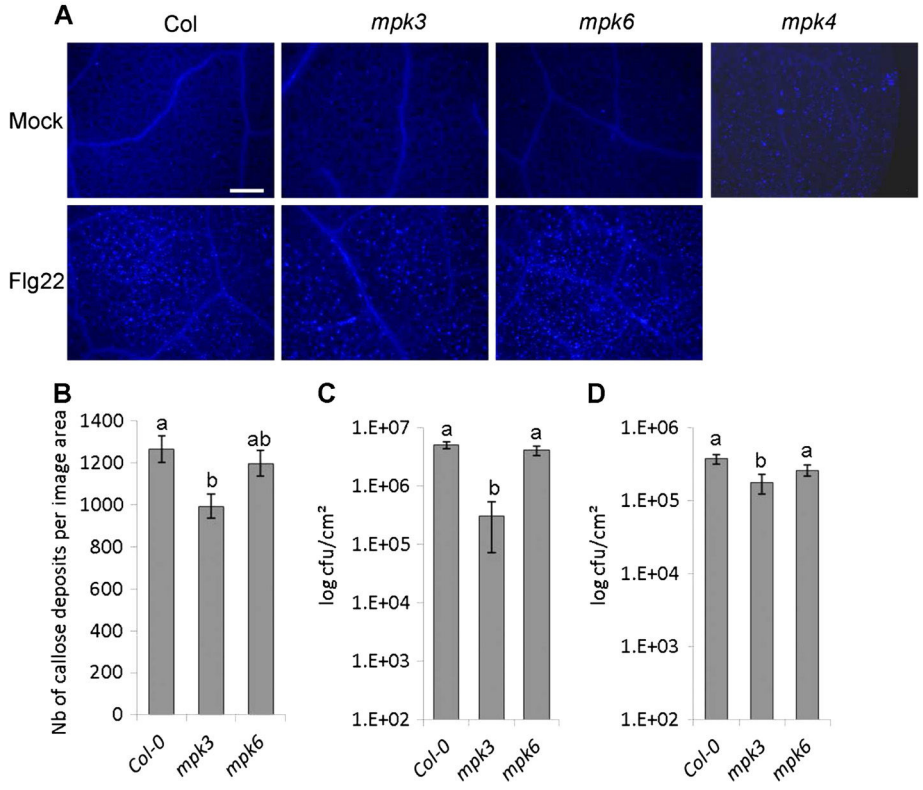


Figure 7 *mpk3* adult plants show reduced flg22-induced callose accumulation and enhanced resistance to virulent *P. syringae*. **(A)** Pictures of aniline-blue stained leaves of the indicated phenotypes. Leaves were either untreated (*mpk4*), infiltrated with mock (H_20) or a 1 μM flg22 solution for 24 h. Bar is 200 μm. **(B)** Relative quantification of callose deposition in leaves of Col-0, *mpk3* and *mpk6* after infiltration of 100 nM flg22. **(C)** *Pst* DC3000 bacterial titres 3 days post spray-inoculation. Bars are means ± SD (n = 5). **(D)** *Pst* DC3000 bacterial titres 3 days post inoculation by syringe-infiltration. Bars are means ± SD (n = 4). Letters indicate significant difference based on a Kruskal & Wallis test (α <0.05).

mpk3 displays reduced susceptibility to virulent *Pseudomonas syringae*

To assess the impact of the altered flg22-induced transcriptional reprogramming, SA production and callose deposition in *mpk3* with respect to disease resistance, we challenged *mpk3* and *mpk6* plants with the virulent bacteria *P. syringae* pv. *tomato* DC3000 (*Pst* DC3000). The analyses were performed on adult plants and therefore *mpk4* mutant plants were not included due to their dwarf phenotype. In spray inoculated plants, *mpk3* showed significantly lower bacterial titres while *mpk6* behaved like Col-0 (Figure 7C). Given that both *mpk3* and *mpk6* are impaired in flg22-induced stomatal closure [32], we assessed whether the enhanced resistance in *mpk3* was related to post-invasive resistance by quantifying bacterial growth after inoculation via syringe infiltration that surpasses the stomatal barrier. Using this infection method, *mpk3* plants showed weak but significant reduced susceptibility to *P. syringae* as compared to Col-0 and *mpk6* (Figure 7D), indicating that MPK3 also plays a role in modulating post-invasive disease resistance in mesophyll cells.

Discussion

We performed a comprehensive analysis of early and late responses triggered by flg22 in *mpk3*, *mpk4* and *mpk6*, which revealed new roles for these immune-related MAPKs in stress signalling but also in unchallenged tissues. In untreated conditions, *mpk6* displayed minor transcriptional changes, while we unexpectedly found that *mpk3* and *mpk4* shared the differential regulation of an important set of genes. Several of the genes differentially regulated principally in *mpk4* but also in *mpk3*, were identified as 'late' flg22-regulated genes in previous reports [40]. This suggests that MPK4 and MPK3 function together in unstressed conditions to prevent misregulation of defence genes and to inhibit a premature reprogramming of flg22-regulated genes. The fact that MPK4 shares 56% phosphorylation targets with MPK3 and 28% with MPK6 [36], further supports this hypothesis. Nevertheless, even if *mpk3* and *mpk4* share common differentially regulated genes under normal growth conditions, *mpk3* does not show the developmental defects observed in *mpk4* at the adult stage. Therefore, these data reveal similarities but also fundamental differences in the roles of MPK3 and

MPK4. In contrast to the role of MPK4 in repressing basal and pathogen-induced SA and ROS accumulation [16,38], *MPK3* seems to dampen SA and ROS production only after pathogen challenge. The cause of this difference may rely on the nature of the genes constitutively upregulated in *mpk4* (related to SA and cell death) and the MPK4-mediated negative regulation of the MAPKKK MEKK2 and the NB-LRR SUMM2 [11-13]. In response to flg22 treatment, we observed that one-third of early flg22-regulated genes are differentially regulated in at least one of the three *MAPK* mutants. Among these genes, two-thirds are downregulated and are equally controlled by *MPK3*, *MPK4* and *MPK6* suggesting a cooperative activity of the three kinases in gene repression. With respect to the flg22-induced genes, we unexpectedly found an important proportion of flg22-induced genes showing compromised regulation in *mpk4* in response to flg22, which were not differentially regulated in untreated *mpk4*. This reveals that *MPK4*, usually considered as a negative regulator of defence responses, is also a master regulator of early flg22-induced transcriptional activation. In summary, these observations indicate the existence of MAPK specific and cooperative functions in gene regulation. In agreement with this concept, there are transcription factors known to be regulated by one (that is, ERF104 [34]), two (that is, ERF6 [26]) or the three MAPKs (that is, WRKY33 [67,68]).

The clustering and cis-element analyses and the construction of interaction networks, allowed us to identify putative regulators of the MAPK-dependent transcriptional responses. The clustering and cis-element analyses revealed several WRKY and MYB transcription factors as putative downstream factors controlling MAPK-dependent transcriptional reprogramming. This is in agreement with WRKY and MYB factors being known MAPK targets involved in stress responses [36,67,68]. Interestingly, the clustering analysis also revealed EIN3 binding sites in clusters of flg22-upregulated or downregulated genes. EIN3 is involved in the induction and repression of important immune components such as *FLS2* and *SID2*, respectively [55,56]. As EIN3 is a phosphorylation target of MPK3 and MPK6 involved in the regulation of the flg22-induced transcriptional reprogramming [54,56], it seems possible that EIN3 mediates the *MPK3*-dependent *SID2* repression as well as other *MAPK*-dependent transcriptional changes. As a complementary approach, we used the interaction network analysis that reveals putative regulatory hubs that are not transcriptionally regulated. Rapid transcriptional reprogramming responses are usually controlled by preformed transcription factors that are post-translationally regulated. Such key transcriptional regulators are expected to be present in the FLS2-mediated signalling pathway but are still unknown. In our analysis, we found two light-regulated transcription factors, HY5 and PIF1 as putative preformed transcription factors that may be involved in the FLS2 pathway. Indeed, HY5 is an important regulator of photomorphogenesis that is primarily regulated post-transcriptionally by protein degradation in response to light [63] and PIF1 is regulated by phosphorylation and other post-translational modifications in response to blue light [64]. Interestingly, a recent report showed that HY5 and PIF1 interact in Arabidopsis nuclei and coordinately regulate ROS and stress-related genes in response to light [69], suggesting that HY5 and PIF1 could modulate MAPK-dependent gene regulation of stress-related genes.

The transcriptome analysis revealed cooperative roles for the three MAPKs and prompted us to analyse whether the absence of one MAPK could influence the functioning of the other two MAPKs. Our biochemical analysis indeed revealed that MPK3 and MPK6 influence the activities of the other two stress-related MAPKs. In *mpk3,* there is longer and stronger activation of MPK4 and MPK6, whereby in *mpk6* there is longer and stronger activation of MPK3 and MPK4. In contrast, *mpk4* did not show differential MAPK regulation. Despite the enhanced MPK3 and MPK6 activities observed in a mutant of the upstream kinase MEKK1 [15], no differential regulation of flg22-induced MPK3 and MPK6 was detected in the double mutant of the downstream MKK1 and MKK2 [9]. These data indicate that the differential regulation of MPK3 and MPK6 observed in *mekk1* is independent of the downstream kinases. This unexpected finding sheds light on the complex cross-talk between MPK3, MPK4 and MPK6 during FLS2-mediated signalling. The intensity and duration of MAPK activities are key signatures, which can trigger different responses. Indeed, plant immune responses lead to transient MAPK activation during PTI and sustained MAPK activities during ETI [8,70]. Thus, it is possible that *mpk3* and *mpk6* phenotypes are not only due to the loss of function of one MAPK but also to the prolonged activities of the two other stress-induced MAPKs. In light of these results, it seems necessary to reconsider previous data obtained with *mpk3* and *mpk6* mutants, as certain phenotypes attributed to the loss of function of one MAPK may be due to the increased activity of other stress-induced MAPKs. The enhanced kinase activities in the respective MAPK knock out mutants may alter a number of properties of the affected MAPKs, such as their subcellular localization, substrate specificity, stability or complex formation. These changed properties may in turn compensate for the knocked out MAPK protein or lead to different responses.

MAPK activities are regulated by the concerted action of kinases and phosphatases. In our study we did not observe direct interaction in yeast (data not shown) or phosphorylation between the three MAPKs, which favours an indirect cross-talk mechanism. One such indirect

mechanism may be mediated by protein phosphatases. Phosphatases are the major negative regulators of MAPKs and, indeed, the dual specificity phosphatase MAPK Phosphatase 1 (MKP1) and the Ser/Thr PP2C-type phosphatase AP2C1 are known regulators of the activation of MPK3, MPK4 and MPK6 in response to MAMPs or DAMPs (damage-associated molecular patterns) [27,28,71]. In animal cells, the MAPK Phosphatase 3 (MKP3) regulates the activity of the MAPKs p38 and ERK2 and forms a ternary complex with the two kinases that mediates cross regulation between both MAPK pathways [72]. A similar situation could explain the differential regulation of MAPK activities we observed in the three *MAPK* mutants. A plausible hypothesis would be that MPK3 and MPK6 regulate the activity of phosphatases that in turn regulate the activation of the other MAPKs. Indeed, MKP1 was shown to be a target of MPK6 [73] and MPK6 inactivation observed in AP2C1 overexpressing lines was partially suppressed by *mpk3* [27], suggesting that MPK3-mediated activation of AP2C1 is necessary for its phosphatase activity. In a recent phosphoproteome analysis, we identified two MKP1 phosphopeptides, with a pSP motif, whose abundance increases in response to 15 min flg22 treatment [74]. These sites are important for the regulation of MKP1 phosphatase activity and were shown to be phosphorylated by MPK6 and presumably also by MPK3 [73,75]. Our phosphoproteome analysis also identified the phosphopeptides corresponding to the MPK4 and MPK6 activation loops both in the dual and single phosphorylated states [74], supporting the idea that MKP1 and other phosphatases could play a role in the regulation of these MAPKs in response to flg22. Alternatively, the transcriptional regulation and protein turnover of the flg22 receptor FLS2 are also important determinants of the activation of the pathway [55,76]. Although the transcriptome analysis did not reveal important differences in *FLS2* expression, we cannot exclude that FLS2 transcript or protein accumulation could be differentially regulated in the *MAPK* mutants by transcriptional regulation through EIN3 or other factors.

We found that *mpk6* shows minor changes in the transcriptome and no changes in SA accumulation, callose deposition and *Pst* DC3000 susceptibility, while displaying stronger and prolonged MPK3 and MPK4 activities than wild type plants. These results suggest that either MPK6 plays minor functions in FLS2-mediated signalling or that the enhanced activities of MPK3 and/or MPK4 are able to reconstitute most MPK6 functions required in these conditions. In contrast, *mpk3* mutant displayed important transcriptome changes, enhanced flg22-triggered SA accumulation, reduced callose accumulation and reduced susceptibility to *Pst* DC3000, despite presenting enhanced MPK4 and MPK6 activities. We therefore conclude that MPK4 and MPK6 lack unique features of MPK3. While we were surprised by the phenotypes observed in *mpk3*, which is usually considered as a positive regulator of PTI and disease resistance together with MPK6, we found several indications in recent reports suggesting distinct roles for the two kinases. For example, MPK3 and MPK6 play different roles in the defence response to *B. cinerea*: while MPK3 is required for basal resistance, MPK6 contributes only to elicitor-induced resistance to the fungus [25,27]. On the other hand, previous reports showed that MPK6, and not MPK3, is necessary for deregulated stress phenotypes [28,29]. Indeed, *mkp1* mutant plants show enhanced resistance to virulent *P. syringae* and enhanced MPK6 activation, and the disease resistance was suppressed in a *mkp1 mpk6* double mutant [28]. These data suggest that the enhanced activity of MPK6 may account for the enhanced stress responses observed in *mpk3*. Unfortunately, the embryo lethality of the *mpk3 mpk6* double mutant prevents the verification of this hypothesis.

Previous data on the *P. syringae* effector HopAI1, a phosphothreonine lyase that inactivates MPK3 and MPK6, suggested that these two MAPKs regulate the flg22-induced RbohD-dependent ROS production and callose accumulation [21]. These conclusions were based on the use of transgenic plants with inducible expression of $MKK5^{DD}$ and HopAI1, which respectively activate and inactivate the two MAPKs. Therefore these approaches did not allow distinguishing between the specific contributions of each kinase. Using *MAPK* single mutants, we and other groups could show that flg22-treated *mpk3* displays prolonged ROS production and increased growth inhibition but reduced callose deposition ([33] and this study). In contrast, *mpk6* behaved like wild type or had minor phenotypes in all assays. Recently, it was shown that HopAI1 is also capable of dephosphorylating and thereby inactivating MPK4 [13]. Nevertheless, current evidence indicates that while the MEKK1-MKK1/MKK2-MPK4 pathway inhibits basal callose accumulation (probably via repression of MEKK2 and SUMM2), it does not influence flg22-induced callose deposition [13,15,38]. This suggests that the reduced flg22-induced callose accumulation observed in HopAI1 transgenic plants and the constitutive callose accumulation in $MKK5^{DD}$ expressing plants is due to their regulation of MPK3. Callose deposition imposes a physical barrier to pathogen penetration but its real role in resistance is still unclear. Indeed, *pmr4* mutant plants, impaired in stress-induced callose deposition, results in an overactivation of SA-mediated defence responses leading to enhanced resistance [77,78]. Therefore, the existence of a feedback regulatory mechanism was proposed, where normal activation of pathogen-induced cell wall modifications stops the activation of downstream defences, and in contrast defects in the initial defence barrier lead to overactivation of the downstream SA defence pathway. A

recent network modelling approach studying Arabidopsis immune signalling, revealed an inhibitory effect of SA-signalling on flg22-induced *PMR4*-dependent callose deposition [79]. These data indicate the possibility that the reduced flg22-induced callose accumulation in *mpk3* could be due to the enhanced induced SA accumulation. On the other hand, flg22-induced RBOHD-dependent ROS production was proposed to be independent of MPK3 and MPK6 [80] and flg22-induced callose deposition is mostly RBOHD-dependent [21]. It is therefore difficult to propose a model that reconciles all published data. Network analysis further revealed a negative link between SA and MPK3. Indeed, the transcriptional reprogramming induced in *mpk3* by the PTI-inducing *Pst HrcC* bacteria showed a strong correlation with the JA/ET deficient mutants *ein2*, *dde2* and *coi1* and not with mutants in SA signalling and included an increased *SID2* expression [79,81]. These data are in agreement with our observations. Taken together, our analysis identified MPK3 as a key negative regulator of defence gene expression, flg22-induced SA signalling and disease resistance to *Pseudomonas syringae*.

Conclusions

A comprehensive molecular and phenotypic analysis was performed for flg22-triggered responses in *mpk3*, *mpk4* and *mpk6*, revealing new roles for these immune-related MAPKs in stress signalling but also in unchallenged tissues. A genome-wide transcriptome analysis of untreated and flg22-challenged *MAPK* mutants coupled with model-based clustering, plus the construction of gene interaction networks, allowed us to identify putative regulators of MAPK-dependent transcriptional reprogramming. Altogether, this work provides evidence that MPK3, MPK4 and MPK6 possess both cooperative and specific functions in plant immune regulation and that the absence of one MAPK influences the activities of the other stress-induced MAPKs. The link between the three MAPK pathways provides an integrated mechanism to optimally coordinate the immune responses of plants.

Material and methods

Plant material

Arabidopsis thaliana ecotype Col-0 was used in this study. The mutants were: *mpk4-2* (SALK_056245), *mpk3* (SALK_151594) and *mpk6-2* (SALK_073907). For bacterial growth curves and callose detection assays, plants were grown on soil for 4 to 5 weeks in short day conditions (8 h light, 16 h dark), with 22°C and 65% relative humidity. For gene expression analyses, protein extraction for immunoblot analyses and SA accumulation, seedlings were grown *in vitro*. Seeds were surface sterilised and stratified for 2 days at 4°C. Seedlings were then grown for 13 days in a culture chamber at 22°C with 16 h photoperiod, on MS plates (0.5 × Murashige Skoog Basal Salts (Sigma #M6899), 1% sucrose, 0.5% agar, 0.5% MES, pH 5.7). Twenty-four hours before treatment, liquid MS (same media without agar) was added to the MS plates to facilitate the transfer of seedlings to liquid MS. Seedlings were treated with deionized water (mock) or with a final concentration of 1 μM flg22, for the required times and then frozen in liquid nitrogen. In the case of *mpk4* single mutant, *mpk3 mpk4* and *mpk6 mpk4* double mutants, the *mpk4-2* mutation was segregating. These seedlings were thus first grown vertically in MS plates with 1% agar for 7 days to isolate *mpk4*$^{-/-}$ seedlings based on their root phenotype (thickening and shortening of the primary root [82]). Selected seedlings were then transferred to liquid MS with the growth conditions previously described and treated as the other lines at 14 days old.

RNA extraction and RT-qPCR

For flg22-induced gene regulation, seedlings were treated with 1 μM flg22 for 1 h. RNA was extracted and DNA digested using the RNeasy plant mini kit and the RNase-Free DNase Set (Qiagen). Three different biological replicates were performed and 1 μg of each RNA was pooled to synthesize cDNA using the Superscript II enzyme (Invitrogen). Two microliters of a 100x dilution of the cDNA was used for each quantitative PCR, using a 7900 HT Sequence Detection System (Applied Biosystem) and MESA Green qPCR Mastermix Plus detection system (Eurogentec). RNA/cDNA variable inputs were corrected by normalisation to the housekeeping transcript ACT2. Error bars shown represent the standard deviations obtained from three technical replicates. Oligonucleotides used in this study for RT-qPCR are: ACT2-For 5′-CGTTTCTATGATGCACTTGTGTG-3′, ACT2-Rev 5′-GGGAACAAAAGGAATAAAGAGG-3′, SID2-For 5′-AGCTGGAAGTGACCCATCTT-3′, SID2-Rev 5′- TGGTGAACTGCAAAAACAACA-3′, EDS1-For 5′-CTCAATGACCTTGGAGTGAGC-3′, EDS1-Rev 5′-TCTTCCTCTAATGCAGCTTGAA-3′, PAD4-For 5′-TGGTGACGAAGAAGGAGGTT-3′, PAD4-Rev 5′-TCCATTGCGTCACTCTCATC-3′, PDF1.2-For 5′-GG ACATGGTCAGGGGTTTGCGG-3′ and PDF1.2-Rev 5′-TGTGTGCTGGGAAGACATAGTTGC-3′.

Transcriptome studies

Microarray analysis was carried out at the Unité de Recherche en Génomique Végétale (Evry, France), using the CATMAv6.2 array based on Roche-NimbleGen technology. CATMAv6.2 microarray slides contain 12 chambers, each containing 219,684 primers representing all the *Arabidopsis thaliana* genes: 37,309 probes corresponding to TAIRv8 annotation (including 476 probes of mitochondrial and chloroplast genes) and 1,796 probes

corresponding to EUGENE software predictions. The slides also include 5,328 probes corresponding to repeat elements, 1,322 probes for miRNA/MIR, 329 probes for other RNAs (rRNA,tRNA, snRNA, soRNA) and several controls. In each chamber, probes are present in triplicates and in both strands. Three independent biological replicates of the microarray analysis were produced. For each biological repetition and each point, 14-day-old seedlings grown in long day conditions were collected and RNA samples were obtained by pooling more than 50 plants. Total RNA was extracted using Qiagen RNAeasy according to the supplier's instructions. For each comparison, one technical replicate with fluorochrome reversal was performed for each biological replicate (that is, six hybridisations per comparison). The labelling of cRNAs with Cy3-dUTP or Cy5-dUTP (Perkin-Elmer-NEN Life Science Products) and the hybridisation to the slides were performed as previously described [83]. Two micron scanning was performed with InnoScan900 scanner (Innopsys^R, Carbonne, FRANCE) and raw data were extracted using Mapix^R software (Innopsys^R, Carbonne, FRANCE).

Differential analysis of microarray data

For each array, the raw data comprised the logarithm of median feature pixel intensity at wavelengths 635 nm (red) and 532 nm (green). For each array, a global intensity-dependent normalisation using the loess procedure [84] was performed to correct the dye bias. The differential analysis is based on the log-ratios averaging over the duplicate probes and over the technical replicates. Hence the numbers of available data for each gene equals the number of biological replicates and are used to calculate the moderated *t*-test [85]. Under the null hypothesis, no evidence that the specific variances vary between probes is highlighted by Limma and consequently the moderated t-statistic is assumed to follow a standard normal distribution. To control the false discovery rate, we calculated adjusted *P* values using the optimised FDR approach [86]. We considered as being differentially expressed the probes with an adjusted *P* value ≤0.05. Analysis was done with the R software. The function SqueezeVar of the library limma has been used to smooth the specific variances by computing empirical Bayes posterior means. The library kerfdr has been used to calculate the adjusted *P* values. The overlap between different sets of genes was generated by the Venn diagram generator Venny [87]. The analysis to find over-represented categories in the gene sets was obtained with AmiGO [88], which is based on a hypergeometric test. Co-expression analysis was performed with ATTEDII version 6.1 using the Network Drawer tool and 'add a few genes' settings for co-expression and Protein Protein Interaction options [89,90]. The thickness is representative of the rank of correlation between two genes of interest via the calculation of a geometric averaged rank (MR).

Data availability

Microarray data from this article were deposited at CATdb [91] (Project RA12-05_mut_flg_II) and GEO (Project GSE52587) according to the 'Minimum Information About a Microarray Experiment' standards.

Clustering of microarray data

The dataset for the co-expression analysis was built from the results of the differential analyses. Probes with at least one Bonferroni pvalue lower than 0.05 were considered. It leads to a dataset of 4,378 probes described by seven expression differences, each one being the average of the three biological replicates. The clustering was performed with a multidimensional Gaussian mixture with unequal proportions and a component number varying from 2 to 40. Covariance matrices are constrained so that their volumes differ and their orientation and shape are equal. Estimations were done with the MIXMOD software [92] and a mixture of 29 components was selected according to the BIC criterion. Probes were assigned in the cluster for which the conditional probability is the highest and interpretation was done only for probes for which this probability is greater than 0.878. This threshold was fixed so that as many observations as possible were classified, under the constraint that the proportion of misclassified observations is controlled at a level of 5%. It is an extension of the BFDR previously described [93]. In our analysis based on 4,378 probes 1,928 probes were classified, which means that in average 96 probes were badly assigned. Cluster profiles are represented as boxplots. The bottom and top of the box are the first and third quartiles, denoted respectively Q1 and Q3. The band inside the box is the median. The ends of the whiskers represent, respectively, Q1 - 1.5 x (Q3-Q1) and Q3 + 1.5 × (Q3-Q1). Data not included between the whiskers are represented by a dot. On top is indicated the percentage of genes differentially regulated (Bonferroni *P* value <0.05) in the different comparisons.

Detection of the cis-elements

We analysed the presence of conserved motifs in the 5′ region of genes, also known as cis-elements. The Arabidopsis promoter dataset was downloaded from FLAGdb++ based on TAIRv8 [94]. The dataset includes 27,025 promoters containing 1,000 base pairs upstream known transcription starting sites (TSSs) or upstream the ATG start otherwise. A list of 140 motifs known to be involved in stress responses was extracted from the databases PLACE [95] and AGRIS [96]. For each cluster, the presence of these motifs was identified by the Preferentially Located Motifs (PLMs) method [97]. This method determines the preferential location of each motif relative to the TSS and a functional window derived from the peak boundaries of the region in which the transcription factor binding site is

over-represented. Taking into account the position of the binding site with respect to the TSS limits the rate of false positives. A motif identified by this method is a motif over-represented at a given place regarding the TSS and is named PLM. 29 motifs were declared as PLMs among 140 motifs tested. To evaluate whether a given PLM was over-represented in a cluster with respect to the whole genome, a binomial test was performed by comparing the gene number of this cluster containing this PLM to the gene number containing this PLM in the same functional window at the genome level. PLMs with a *P* value lower than 0.01 were considered as significantly over-represented.

GO analysis for the co-expression clusters

Gene function annotation was downloaded from TAIRv10 and the GO Slim classification for the three branches of the GO vocabulary (biological process, molecular function and cellular component) was considered. The enrichment analysis was performed by comparing the ratio of the relative occurrence of a GO term into the cluster to its relative occurrence in the genome by a hypergeometric test. A GO term was declared significantly over-represented if its *P* value was lower than 0.05.

Gene interaction network construction

A total of 12,741 protein-protein interactions (PPI) data were extracted from: (1) Arabidopsis Interactome Consortium [98], where a matrix of 9 k × 9 k full length protein encoding ORFs were tested by yeast 2-hybrid assay and a total of 6,475 positive interactions were detected; (2) public databases: BioGRId, IntAct, TAIR and BIND (6365 experimental PPI data). Concerning the TF-target data, 769 confirmed interactions were downloaded from AtRegNet database [96]. These interaction information on protein-protein interactions (PPI) and transcription factor-target interactions were combined to the co-expression clusters using a home-made Perl program leading to a gene interaction network of 839 genes linked with 983 edges. Network visualisation and analysis were done with Cytoscape [99]. The node degree varied between 1 and 115 with a median equal to 1 and a third quartile equal to 2, meaning that the majority of the genes were few connected, and the 10 most connected genes had a degree greater than 19. For this reason we defined as regulatory hubs those proteins displaying more than 19 edges, with each edge representing a validated interaction.

Immunoblotting

Protein extractions: approximately 100 mg of frozen samples were ground in liquid nitrogen using a tissue lyser (Qiagen) and metal beads. The ground material was resuspended in 200 μL of a extraction buffer containing 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% NP40, 5 mM EGTA, 0.1 mM DTT (Sigma-Aldrich chemicals), protease inhibitors (Complete cocktail, Roche, and 1 mM PMSF, Sigma-Aldrich) and phosphatase inhibitors (1 mM NaF, 0.5 mM Na_3VO_4, 15 mM beta-glycerophosphate, 15 mM 4-nitrophenyl phosphate, Sigma-Aldrich chemicals). The suspension was centrifuged at 20,000 g for 15 min at 4°C and the supernatant was collected. Protein quantification was carried out with Bradford (Sigma-Aldrich) and BSA standard (Thermo Scientific), and the normalised protein amounts of all the samples were denatured by boiling in SDS-sample buffer at 95°C. When specified, the ground material was directly boiled at 95°C in 2× SDS-sample buffer and centrifuged at 20,000 g for 2 min. The supernatant was recovered and proteins were quantified with Amido Black 10B (Sigma-Aldrich). *Immunoblottings*: Protein samples were separated on 10% SDS-PAGE gels and transferred onto PVDF membranes (GE Healthcare). *Anti-pTpY antibody*: blots were blocked with 5% (w/v) BSA (Sigma-Aldrich) in TBST and incubated overnight at 4°C with the rabbit anti-phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) monoclonal antibody (Cell Signalling) at a dilution of 1/1,500. *Anti-MAPK antibodies*: Blots were blocked in 5% (w/v) non-fat dry milk in TBST and incubated overnight at 4°C with anti-MPK3 and anti-MPK4 antibodies previously described [100] at a dilution of 1/3,000, or with anti-MPK6 antibody (Sigma-Aldrich) at a dilution of 1/5,000. As secondary antibody we used the goat anti-rabbit horseradish peroxidase (HRP)-conjugated (Sigma-Aldrich) diluted to 1/20,000. HRP activity was detected with a chemiluminescent reagent (GE Healthcare) using the GeneGnome imaging system (Syngene) or clear-blue X-ray films (Thermo Scientific). Blots were stained with Coomassie blue for protein visualization. Each immunoblotting analysis shown is representative of at least two independent biological repeats.

Purification and activity assays of recombinant MAPKs

MAPK protein expression in *E. coli*, purification, and activity assays were performed as previously described [38]. Wild type and constitutive active (Y and DE variants) variants of the MAPK proteins were His-tagged in the case of MPK4 and MPK6 and fused to peri-His-MBP in the case of MPK3 as previously described [38]. GST-tagged kinase dead variants of MPK3 and MPK6 carry mutations in the ATP binding site and were previously described [101].

Salicylic acid quantification

Total SA was extracted as previously described [102] with the following modifications. [^{14}C]SA (50 Bq, 2 GBq mmol^{-1}, NEN, UK) was added to each sample to correct for losses. Samples were dried in a SC 110A Speed-Vac (Savant Instrument Inc., New York, NY, USA) and subjected to acidic hydrolysis in order to determine total SA. SA was identified and quantified by HPLC based

on a comparison with the standard. SA standard was purchased from Sigma-Aldrich (Saint-Quentin Fallavier, France).

Pseudomonas assays

Infections with *Pseudomonas syringae* pv. *tomato* (*Pst*) DC3000 were done by spray inoculation with bacterial solution at 1×10^8 cfu/mL or by syringe-infiltration at 1×10^5 cfu/mL. Bacterial titers were determined as previously described [103].

Callose assays

Callose assay was performed as previously described after infiltration of leaves of adult plants with H_20 (mock) or 1 μM flg22 solution [40].

Additional files

Additional file 1: Table S1. List of genes affected in untreated *mpk3*, *mpk4* and *mpk6*or showing differential expression after flg22 treatment. Only genes showing a differential expression (*P* value <0.05) in at least one of the seven comparisons is retained in the table. Complete expression data can be downloaded from CATdb ([91]; Project: RA12-05_mut_flg_II).

Additional file 2: Table S2. Differentially expressed genes and GO term enrichment observed in the mock-treated samples from the comparison between Col-0 and *mpk3*, *mpk4* or *mpk6*.

Additional file 3: Table S3. Analysis of genes commonly or specifically misregulated in *mpk3* and *mpk4* in mock-treated samples. GO term enrichments associated to the different gene classes are also mentioned.

Additional file 4: Figure S1. *mpk3*, *mpk4* and *mpk6* do not mimic the flg22-induced transcriptional reprogramming. (A) Venn diagram of upregulated genes observed in Col-0 after flg22 treatment and in *mpk3*, *mpk4* and *mpk6* in comparison with Col-0. (B) Venn diagram of downregulated genes observed in Col-0 after flg22 treatment and in *mpk3*, *mpk4*and *mpk6* in comparison with Col-0. Note that few genes misregulated in the *MAPK* mutants follow the same misregulation in Col-0 treated with flg22.

Additional file 5: Table S4. Differentially expressed genes and GO term enrichment observed in response to flg22 in Col-0, *mpk3*, *mpk4* and *mpk6*. Genes for which the flg22-induced regulation is affected by at least 1 log in *mpk3*, *mpk4* and *mpk6* are mentioned, together with their associated GO term enrichments.

Additional file 6: Table S5. Analysis of genes commonly or specifically misregulated in *mpk3*, *mpk4* and *mpk6* after flg22 treatment. GO term enrichments associated to the different gene classes are also mentioned.

Additional file 7: Figure S2. Twenty-four percent of the flg22-upregulated *MPK4*-dependent genes are upregulated in mock-treated *mpk4*. (A) Expression profiles of the 89 genes that are upregulated in mock-treated *mpk4* and show reduced flg22-induced upregulation in *mpk4* as compared with Col-0. (B) Expression profiles of the 342 genes that are unmodified in mock-treated *mpk4* and show reduced flg22-induced upregulation in *mpk4* as compared with Col-0. Profiles are represented as boxplots, where the bottom and top of the box are the first and third quartiles and the band inside the box is the median. Data not included between the whiskers are represented by a dot.

Additional file 8: Figure S3. Analysis of Gene Ontology (GO) enrichment in flg22-regulated MPK4-dependent genes. (A) Venn diagram analysis of GO families in the two gene groups described in Additional file 7: Figure S2. Numbers inside the Venn diagram correspond to GO categories. (B) Throughout other enrichments, GOs for ethylene signalling and synthesisgenes show MPK4-dependency upon flg22 treatment but not under standard conditions. (C) GOs associated to cell death regulation and immune responses are present in MPK4-dependent genes upon flg22 treatment, but only partially upregulated under standard conditions. (D) GOs related to SA, JA, ROS, cell death and immune responses are present in genes upregulated in *mpk4* in standard conditions, but still show MPK4-dependency upon flg22 treatment. SA: salicylic acid, JA: jasmonic acid, ROS: reactive oxygen species, GO: Gene Ontology, HR: Hypersensitive response, N: number of genes. Note that less genes than previously indicated (Additional file 7: Figure S2) are described here since databases displaying GO enrichment do not contain data for all genes present on CATMA V6.0 chips.

Additional file 9: Figure S4. Overview of the clusters obtained from the coexpression analysis. The y-axis shows log ratios. The x-axis shows the following comparisons: Col-0 + flg22 *vs.* Col-0, *mpk3 vs.* Col-0, *mpk4 vs.* Col-0, *mpk6 vs.* Col-0, *mpk3* + flg22 *vs. mpk*3, *mpk4* + flg22 *vs. mpk4*, *mpk6* + flg22 *vs. mpk6*. Profiles are represented as boxplots, where the bottom and top of the box are the first and third quartiles and the band inside the box is the median. Data not included between the whiskers are represented by a dot. On top is indicated the percentage of genes differentially regulated (*P* value <0.05) in the different comparisons.

Additional file 10: Table S6. Tables containing the gene lists, ATTED network representations and GO term enrichment for each clusters.

Additional file 11: Figure S5. Description of the 10 gene classes defined from the kinetic study performed by Denoux et al. [40]. In the upregulated genes, the discrimination between the classes H, I, J is based on at least a two-fold difference in the fold change observed at 1 h and 3 h. For example, a gene induced 10 times at 1 h and 15 times at 3 h will belong to class H (the difference between 1 h and 3 h is less than two-fold), but a gene induced 10 times at 1 h and induced 50 times at 3 h will belong to class I (the difference in fold change between 1 h and 3 h is greater than 2). Similar analysis is made to build the classes of downregulated genes (Classes A-E). hpt: hours post treatment, a.u.: arbitrary units.

Additional file 12: Figure S6. Genes not affected by flg22 in Col-0 after 30 min and upregulated under standard conditions in *mpk4* are enriched in 'late' flg22-induced genes. Cluster 4 and 7 do not show enrichment for up- or downregulated genes classes in data from Denoux et al. [40]. Profiles are represented as boxplots, where the bottom and top of the box are the first and third quartiles and the band inside the box is the median. Data not included between the whiskers are represented by a dot.

Additional file 13: Figure S7. Flg22-induced *MPK4*-dependent genes are enriched in early and transiently induced genes, as indicated by the comparison with data from Denoux et al. [40]. Profiles are represented as boxplots, where the bottom and top of the box are the first and third quartiles and the band inside the box is the median. Data not included between the whiskers are represented by a dot.

Additional file 14: Figure S8. ATTED2 representation of gene co-expression observed in cluster 23. White coloured genes are present in the cluster. Grey coloured genes are out of the cluster but contribute the network. Transcription factors are indicated by octagonal shapes. Coloured dots indicate metabolic pathways. Red: biosynthesis of secondary metabolites (KEGG ID: ath01110), yellow: glucosinolate biosynthesis (KEGG ID: ath00966), green: flavonoid biosynthesis (KEGG ID: ath00941), light blue: glutathione metabolism (KEGG ID: ath00480), blue: valine, leucine and isoleucine biosynthesis (KEGG ID: ath00290). Thickness of lines linking two genes indicates the strength of the co-expression. Orange lines indicate protein-protein interaction. Large circles with dashed lines highlight gene clusters involved in processes of interest for our study.

Additional file 15: Figure S9. ATTED2 representation of gene co-expression observed in cluster 16. White coloured genes are present in the cluster, grey coloured genes are outside of the cluster but contribute to the network. Transcription factors are indicated by octagonal shapes. Coloured dots indicate metabolic pathways. Red: valine, leucine and isoleucine degradation (KEGG ID: ath00280), yellow: biosynthesis of secondary metabolites (KEGG ID: ath01110), green: propanoate metabolism (KEGG ID: ath00640), light blue: alanine, aspartate and glutamate metabolism (KEGG ID: ath00250), blue: arginine and proline metabolism (KEGG ID: ath00330). Thickness of lines linking two genes

indicates the strength of the co-expression. Orange lines indicate protein-protein interaction. Large circles with dashed lines highlight gene clusters involved in processes of interest for our study.

Additional file 16: Figure S10. Clusters 8, 13 and 24 group genes more rapidly regulated by flg22 in *mpk3* and *mpk6*, as indicated by the comparison with 'late' flg22-regulated genes (from Denoux et al. [40]). Profiles are represented as boxplots, where the bottom and top of the box are the first and third quartiles and the band inside the box is the median. Data not included between the whiskers are represented by a dot.

Additional file 17: Table S7. List of CIS elements enriched in the promoters of the different clusters.

Additional file 18: Figure S11. Construction of gene interaction networks. (A) Cytoscape representation of the gene interaction network highlighting the identified regulatory hubs. Zooms into the interaction networks of the regulatory hubs CML9 (B), CPK4 and CPK11 (C), PIF1 (D), HY5 (E).

Additional file 19: Figure S12. *mpk3 mpk4* and *mpk6 mpk4* double mutant plants resemble phenotypically single *mpk4* mutant plants. Pictures of 5-week-old soil grown plants of the indicated genotypes. Arrows indicate *mpk4*, *mpk3 mpk4* and *mpk6 mpk4* dwarf plants.

Additional file 20: Figure S13. Immunoblot analysis of the protein abundance of MPK3, MPK4 and MPK6 in Col-0, in *mpk3*, *mpk4* and *mpk6* single and in *mpk3 mpk4* and *mpk6 mpk4* double mutants treated with flg22. Western blot analysis of Col-0, *mpk3* and *mpk6* (A) and Col-0, *mpk4*, *mpk3 mpk4* and *mpk6 mpk4* (B) at the indicated time-points after flg22 treatment, using anti-MPK antibodies to detect MPK3, MPK4 and MPK6 abundance. Arrows indicate the protein bands corresponding to MPK3, MPK4 and MPK6. The size of the molecular weight (MW) markers is indicated in kDa on the left. Blots were stained with Coomassie blue for protein visualization; the lower panels in A and B show the protein band corresponding to the RuBisCO large subunit.

Additional file 21: Figure S14. MPK3, MPK4 and MPK6 do not phosphorylate each other in *in vitro* kinase assays. (A) Kinase activity of recombinant wild type and constitutive active (Y and DE variants) MPK3, MPK4 and MPK6 towards MBP. (B) Kinase activity of recombinant wild type and constitutive active MPK3, MPK4 and MPK6 towards kinase dead MPK3 and MPK6 variants fused to GST. Upper panels indicate kinase activities (autoradiographs) and lower panels show Coomassie blue staining of the gels to indicate equal loading. Upper panels indicate kinase activities (autoradiographs) and lower panels show Coomassie blue staining of the gels to indicate equal loading.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
NFdF, AVG and HH designed research. NFdF and AVG performed callose staining, RT-qPCR and *P. syringae* infections. AVG and MG quantified SA. NFdF, JB and MLdT performed western blots. JC performed kinase assays. SP and SB did the microarray hybridisations. RZ, VB, SA, MLMM and NFdF performed the differential and clustering analysis, prediction of cis-regulatory elements and gene interaction network. NFdF, AVG, JB, EB, MLMM and HH analysed the data. NFdF and AVG wrote the paper with contributions from all authors. All authors read and approved the final manuscript.

Acknowledgements
Funding of the project was provided by the French Agency of Research ANR.

Author details
[1]Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165 - Université d'Evry Val d'Essonne - ERL CNRS 8196 - Saclay Plant Sciences, 2 rue Gaston Crémieux, Evry 91057, France. [2]Present address: Laboratoire de Recherche en Sciences Végétales (LRSV), UMR 5546, Université Paul Sabatier/CNRS, 24, chemin de Borde Rouge B.P. 42617 Auzeville, Castanet-Tolosan 31326, France. [3]Institut de Biologie des Plantes (IBP), CNRS-Université Paris-Sud - UMR 8618 - Saclay Plant Sciences, Orsay, Cedex 91405, France. [4]Unité de Recherche en Génomique Végétale (URGV), Plateforme Transcriptome, UMR INRA 1165 - Université d'Evry Val d'Essonne - ERL CNRS 8196, 2 rue Gaston Crémieux, Evry 91057, France. [5]AgroParisTech, UMR 518 MIA, Paris 75005, France. [6]INRA, UMR 518 MIA, Paris 75005, France. [7]Center for Desert Agriculture, 4700 King Abdullah University of Sciences and Technology, Thuwal 23955-6900, Saudi Arabia.

Received: 15 November 2013 Accepted: 30 June 2014
Published: 30 June 2014

References
1. Monaghan J, Zipfel C: **Plant pattern recognition receptor complexes at the plasma membrane.** *Curr Opin Plant Biol* 2012, **15:**349–357.
2. Gomez-Gomez L, Boller T: **FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis.** *Mol Cell* 2000, **5:**1003–1011.
3. Chinchilla D, Zipfel C, Robatzek S, Kemmerling B, Nurnberger T, Jones JD, Felix G, Boller T: **A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence.** *Nature* 2007, **448:**497–500.
4. Zipfel C, Robatzek S, Navarro L, Oakeley EJ, Jones JD, Felix G, Boller T: **Bacterial disease resistance in Arabidopsis through flagellin perception.** *Nature* 2004, **428:**764–767.
5. Jones JD, Dangl JL: **The plant immune system.** *Nature* 2006, **444:**323–329.
6. Robert-Seilaniantz A, Grant M, Jones JD: **Hormone crosstalk in plant disease and defense: more than just jasmonate-salicylate antagonism.** *Annu Rev Phytopathol* 2011, **49:**317–343.
7. Meng X, Zhang S: **MAPK cascades in plant disease resistance signaling.** *Annu Rev Phytopathol* 2013, **51:**245–266.
8. Asai T, Tena G, Plotnikova J, Willmann MR, Chiu WL, Gomez-Gomez L, Boller T, Ausubel FM, Sheen J: **MAP kinase signalling cascade in Arabidopsis innate immunity.** *Nature* 2002, **415:**977–983.
9. Gao M, Liu J, Bi D, Zhang Z, Cheng F, Chen S, Zhang Y: **MEKK1, MKK1/MKK2 and MPK4 function together in a mitogen-activated protein kinase cascade to regulate innate immunity in plants.** *Cell Res* 2008, **18:**1190–1198.
10. Qiu JL, Zhou L, Yun BW, Nielsen HB, Fiil BK, Petersen K, Mackinlay J, Loake GJ, Mundy J, Morris PC: **Arabidopsis mitogen-activated protein kinase kinases MKK1 and MKK2 have overlapping functions in defense signaling mediated by MEKK1, MPK4, and MKS1.** *Plant Physiol* 2008, **148:**212–222.
11. Kong Q, Qu N, Gao M, Zhang Z, Ding X, Yang F, Li Y, Dong OX, Chen S, Li X, Zhang Y: **The MEKK1-MKK1/MKK2-MPK4 kinase cascade negatively regulates immunity mediated by a mitogen-activated protein kinase kinase kinase in Arabidopsis.** *Plant Cell* 2012, **24:**2225–2236.
12. Su SH, Bush SM, Zaman N, Stecker K, Sussman MR, Krysan P: **Deletion of a tandem gene family in Arabidopsis: increased MEKK2 abundance triggers autoimmunity when the MEKK1-MKK1/2-MPK4 signaling cascade is disrupted.** *Plant Cell* 2013, **25:**1895–1910.
13. Zhang Z, Wu Y, Gao M, Zhang J, Kong Q, Liu Y, Ba H, Zhou J, Zhang Y: **Disruption of PAMP-induced MAP kinase cascade by a Pseudomonas syringae effector activates plant immunity mediated by the NB-LRR protein SUMM2.** *Cell Host Microbe* 2012, **11:**253–263.
14. Suarez-Rodriguez MC, Adams-Phillips L, Liu Y, Wang H, Su SH, Jester PJ, Zhang S, Bent AF, Krysan PJ: **MEKK1 is required for flg22-induced MPK4 activation in Arabidopsis plants.** *Plant Physiol* 2007, **143:**661–669.
15. Ichimura K, Casais C, Peck SC, Shinozaki K, Shirasu K: **MEKK1 is required for MPK4 activation and regulates tissue-specific and temperature-dependent cell death in Arabidopsis.** *J Biol Chem* 2006, **281:**36969–36976.
16. Petersen M, Brodersen P, Naested H, Andreasson E, Lindhart U, Johansen B, Nielsen HB, Lacy M, Austin MJ, Parker JE, Sharma SB, Klessig DF, Martienssen R, Mattsson O, Jensen AB, Mundy J: **Arabidopsis map kinase 4 negatively regulates systemic acquired resistance.** *Cell* 2000, **103:**1111–1120.
17. Brodersen P, Petersen M, Bjorn Nielsen H, Zhu S, Newman MA, Shokat KM, Rietz S, Parker J, Mundy J: **Arabidopsis MAP kinase 4 regulates salicylic acid- and jasmonic acid/ethylene-dependent responses via EDS1 and PAD4.** *Plant J* 2006, **47:**532–546.
18. Bethke G, Pecher P, Eschen-Lippold L, Tsuda K, Katagiri F, Glazebrook J, Scheel D, Lee J: **Activation of the Arabidopsis thaliana mitogen-activated protein kinase MPK11 by the flagellin-derived elicitor peptide, flg22.** *Mol Plant Microbe Interact* 2012, **25:**471–480.
19. Eschen-Lippold L, Bethke G, Palm-Forster MA, Pecher P, Bauer N, Glazebrook J, Scheel D, Lee J: **MPK11-a fourth elicitor-responsive mitogen-activated**

protein kinase in Arabidopsis thaliana. *Plant Signal Behav* 2012, **7:**1203–1205.
20. Wang Y, Li J, Hou S, Wang X, Li Y, Ren D, Chen S, Tang X, Zhou JM: **A Pseudomonas syringae ADP-ribosyltransferase inhibits Arabidopsis mitogen-activated protein kinase kinases.** *Plant Cell* 2010, **22:**2033–2044.
21. Zhang J, Shao F, Li Y, Cui H, Chen L, Li H, Zou Y, Long C, Lan L, Chai J, Chen S, Tang X, Zhou JM: **A Pseudomonas syringae effector inactivates MAPKs to suppress PAMP-induced immunity in plants.** *Cell Host Microbe* 2007, **1:**175–185.
22. Wang H, Ngwenyama N, Liu Y, Walker JC, Zhang S: **Stomatal development and patterning are regulated by environmentally responsive mitogen-activated protein kinases in Arabidopsis.** *Plant Cell* 2007, **19:**63–73.
23. Han L, Li GJ, Yang KY, Mao G, Wang R, Liu Y, Zhang S: **Mitogen-activated protein kinase 3 and 6 regulate Botrytis cinerea-induced ethylene production in Arabidopsis.** *Plant J* 2010, **64:**114–127.
24. Mao G, Meng X, Liu Y, Zheng Z, Chen Z, Zhang S: **Phosphorylation of a WRKY transcription factor by two pathogen-responsive MAPKs drives phytoalexin biosynthesis in Arabidopsis.** *Plant Cell* 2011, **23:**1639–1653.
25. Ren D, Liu Y, Yang KY, Han L, Mao G, Glazebrook J, Zhang S: **A fungal-responsive MAPK cascade regulates phytoalexin biosynthesis in Arabidopsis.** *Proc Natl Acad Sci U S A* 2008, **105:**5638–5643.
26. Meng X, Xu J, He Y, Yang KY, Mordorski B, Liu Y, Zhang S: **Phosphorylation of an ERF transcription factor by Arabidopsis MPK3/MPK6 regulates plant defense gene induction and fungal resistance.** *Plant Cell* 2013, **25:**1126–1142.
27. Galletti R, Ferrari S, De Lorenzo G: **Arabidopsis MPK3 and MPK6 play different roles in basal and oligogalacturonide- or flagellin-induced resistance against Botrytis cinerea.** *Plant Physiol* 2011, **157:**804–814.
28. Anderson JC, Bartels S, Gonzalez Besteiro MA, Shahollari B, Ulm R, Peck SC: **Arabidopsis MAP Kinase Phosphatase 1 (AtMKP1) negatively regulates MPK6-mediated PAMP responses and resistance against bacteria.** *Plant J* 2011, **67:**258–268.
29. Kohorn BD, Kohorn SL, Todorova T, Baptiste G, Stansky K, McCullough M: **A dominant allele of Arabidopsis pectin-binding wall-associated kinase induces a stress response suppressed by MPK6 but not MPK3 mutations.** *Mol Plant* 2012, **5:**841–851.
30. Igarashi D, Bethke G, Xu Y, Tsuda K, Glazebrook J, Katagiri F: **Pattern-triggered immunity suppresses programmed cell death triggered by fumonisin b1.** *PLoS One* 2013, **8:**e60769.
31. Saucedo-Garcia M, Guevara-Garcia A, Gonzalez-Solis A, Cruz-Garcia F, Vazquez-Santana S, Markham JE, Lozano-Rosas MG, Dietrich CR, Ramos-Vega M, Cahoon EB, Gavilanes-Ruiz M: **MPK6, sphinganine and the LCB2a gene from serine palmitoyltransferase are required in the signaling pathway that mediates cell death induced by long chain bases in Arabidopsis.** *New Phytol* 2011, **191:**943–957.
32. Montillet JL, Leonhardt N, Mondy S, Tranchimand S, Rumeau D, Boudsocq M, Garcia AV, Douki T, Bigeard J, Lauriere C, Chevalier A, Castresana C, Hirt H: **An abscisic acid-independent oxylipin pathway controls stomatal closure and immune defense in Arabidopsis.** *PLoS Biol* 2013, **11:**e1001513.
33. Ranf S, Eschen-Lippold L, Pecher P, Lee J, Scheel D: **Interplay between calcium signalling and early signalling elements during defence responses to microbe- or damage-associated molecular patterns.** *Plant J* 2011, **68:**100–113.
34. Bethke G, Unthan T, Uhrig JF, Poschl Y, Gust AA, Scheel D, Lee J: **Flg22 regulates the release of an ethylene response factor substrate from MAP kinase 6 in Arabidopsis thaliana via ethylene signaling.** *Proc Natl Acad Sci U S A* 2009, **106:**8067–8072.
35. Feilner T, Hultschig C, Lee J, Meyer S, Immink RG, Koenig A, Possling A, Seitz H, Beveridge A, Scheel D, Cahill DJ, Lehrach H, Kreutzberger J, Kersten B: **High throughput identification of potential Arabidopsis mitogen-activated protein kinases substrates.** *Mol Cell Proteomics* 2005, **4:**1558–1568.
36. Popescu SC, Popescu GV, Bachan S, Zhang Z, Gerstein M, Snyder M, Dinesh-Kumar SP: **MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays.** *Genes Dev* 2009, **23:**80–92.
37. Sorensson C, Lenman M, Veide-Vilg J, Schopper S, Ljungdahl T, Grotli M, Tamas MJ, Peck SC, Andreasson E: **Determination of primary sequence specificity of Arabidopsis MAPKs MPK3 and MPK6 leads to identification of new substrates.** *Biochem J* 2012, **446:**271–278.
38. Berriri S, Garcia AV, Dit Frey NF, Rozhon W, Pateyron S, Leonhardt N, Montillet JL, Leung J, Hirt H, Colcombet J: **Constitutively active mitogen-activated protein kinase versions reveal functions of Arabidopsis MPK4 in pathogen defense signaling.** *Plant Cell* 2012, **24:**4281–4293.
39. Yang SH, Sharrocks AD, Whitmarsh AJ: **MAP kinase signalling cascades and transcriptional regulation.** *Gene* 2013, **513:**1–13.
40. Denoux C, Galletti R, Mammarella N, Gopalan S, Werck D, De Lorenzo G, Ferrari S, Ausubel FM, Dewdney J: **Activation of defense response pathways by OGs and Flg22 elicitors in Arabidopsis seedlings.** *Mol Plant* 2008, **1:**423–445.
41. Navarro L, Zipfel C, Rowland O, Keller I, Robatzek S, Boller T, Jones JD: **The transcriptional innate immune response to flg22. Interplay and overlap with Avr gene-dependent defense responses and bacterial pathogenesis.** *Plant Physiol* 2004, **135:**1113–1128.
42. Cui H, Wang Y, Xue L, Chu J, Yan C, Fu J, Chen M, Innes RW, Zhou JM: **Pseudomonas syringae effector protein AvrB perturbs Arabidopsis hormone signaling by activating MAP kinase 4.** *Cell Host Microbe* 2010, **7:**164–175.
43. Yang H, Yang S, Li Y, Hua J: **The Arabidopsis BAP1 and BAP2 genes are general inhibitors of programmed cell death.** *Plant Physiol* 2007, **145:**135–146.
44. Dubiella U, Seybold H, Durian G, Komander E, Lassig R, Witte CP, Schulze WX, Romeis T: **Calcium-dependent protein kinase/NADPH oxidase activation circuit is required for rapid defense signal propagation.** *Proc Natl Acad Sci U S A* 2013, **110:**8744–8749.
45. Pecenkova T, Hala M, Kulich I, Kocourkova D, Drdova E, Fendrych M, Toupalova H, Zarsky V: **The role for the exocyst complex subunits Exo70B2 and Exo70H1 in the plant-pathogen interaction.** *J Exp Bot* 2011, **62:**2107–2116.
46. Yoshioka K, Moeder W, Kang HG, Kachroo P, Masmoudi K, Berkowitz G, Klessig DF: **The chimeric Arabidopsis CYCLIC NUCLEOTIDE-GATED ION CHANNEL11/12 activates multiple pathogen resistance responses.** *Plant Cell* 2006, **18:**747–763.
47. Roux M, Schwessinger B, Albrecht C, Chinchilla D, Jones A, Holton N, Malinovsky FG, Tor M, de Vries S, Zipfel C: **The Arabidopsis leucine-rich repeat receptor-like kinases BAK1/SERK3 and BKK1/SERK4 are required for innate immunity to hemibiotrophic and biotrophic pathogens.** *Plant Cell* 2011, **23:**2440–2455.
48. von Saint PV, Zhang W, Kanawati B, Geist B, Faus-Kessler T, Schmitt-Kopplin P, Schaffner AR: **The Arabidopsis glucosyltransferase UGT76B1 conjugates isoleucic acid and modulates plant defense and senescence.** *Plant Cell* 2011, **23:**4124–4145.
49. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data.** *Bioinformatics* 2001, **17:**977–987.
50. Bernard V, Lecharny A, Brunaud V: **Improved detection of motifs with preferential location in promoters.** *Genome* 2010, **53:**739–752.
51. Pandey SP, Somssich IE: **The role of WRKY transcription factors in plant immunity.** *Plant Physiol* 2009, **150:**1648–1655.
52. Clay NK, Adio AM, Denoux C, Jander G, Ausubel FM: **Glucosinolate metabolites required for an Arabidopsis innate immune response.** *Science* 2009, **323:**95–101.
53. van Verk MC, Gatz C, Linthorst HJM: **Transcriptional regulation of plant defense responses.** In *Adv Bot Res*, Volume Volume 51. Edited by van Loon LC. Elsevier; 2009:397–438.
54. Yoo SD, Cho YH, Tena G, Xiong Y, Sheen J: **Dual control of nuclear EIN3 by bifurcate MAPK cascades in C2H4 signalling.** *Nature* 2008, **451:**789–795.
55. Boutrot F, Segonzac C, Chang KN, Qiao H, Ecker JR, Zipfel C, Rathjen JP: **Direct transcriptional control of the Arabidopsis immune receptor FLS2 by the ethylene-dependent transcription factors EIN3 and EIL1.** *Proc Natl Acad Sci U S A* 2010, **107:**14502–14507.
56. Chen H, Xue L, Chintamanani S, Germain H, Lin H, Cui H, Cai R, Zuo J, Tang X, Li X, Guo H, Zhou JM: **ETHYLENE INSENSITIVE3 and ETHYLENE INSENSITIVE3-LIKE1 repress SALICYLIC ACID INDUCTION DEFICIENT2 expression to negatively regulate plant innate immunity in Arabidopsis.** *Plant Cell* 2009, **21:**2527–2540.
57. Guan Y, Ren H, Xie H, Ma Z, Chen F: **Identification and characterization of bZIP-type transcription factors involved in carrot (Daucus carota L.) somatic embryogenesis.** *Plant J* 2009, **60:**207–217.
58. Kagaya Y, Hattori T: **Arabidopsis transcription factors, RAV1 and RAV2, are regulated by touch-related stimuli in a dose-dependent and biphasic manner.** *Genes Genet Syst* 2009, **84:**95–99.

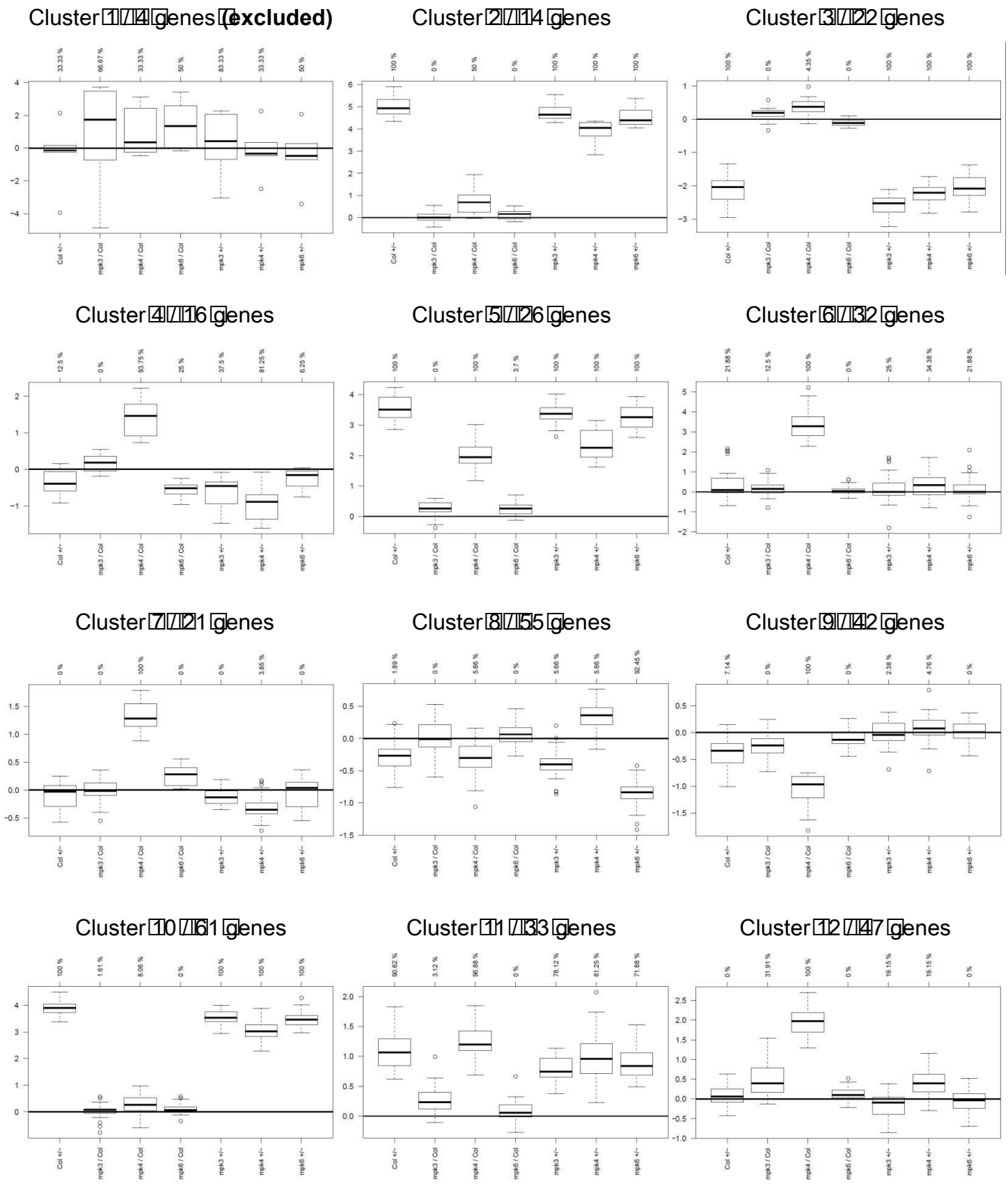
59. Leba LJ, Cheval C, Ortiz-Martin I, Ranty B, Beuzon CR, Galaud JP, Aldon D: **CML9, an Arabidopsis calmodulin-like protein, contributes to plant innate immunity through a flagellin-dependent signalling pathway.** *Plant J* 2012, **71:**976–989.
60. Boudsocq M, Willmann MR, McCormack M, Lee H, Shan L, He P, Bush J, Cheng SH, Sheen J: **Differential innate immune signalling via Ca(2+) sensor protein kinases.** *Nature* 2010, **464:**418–422.
61. Popescu SC, Popescu GV, Bachan S, Zhang Z, Seay M, Gerstein M, Snyder M, Dinesh-Kumar SP: **Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays.** *Proc Natl Acad Sci U S A* 2007, **104:**4730–4735.
62. Burr CA, Leslie ME, Orlowski SK, Chen I, Wright CE, Daniels MJ, Liljegren SJ: **CAST AWAY, a membrane-associated receptor-like kinase, inhibits organ abscission in Arabidopsis.** *Plant Physiol* 2011, **156:**1837–1850.
63. Osterlund MT, Hardtke CS, Wei N, Deng XW: **Targeted destabilization of HY5 during light-regulated development of Arabidopsis.** *Nature* 2000, **405:**462–466.
64. Bu Q, Castillon A, Chen F, Zhu L, Huq E: **Dimerization and blue light regulation of PIF1 interacting bHLH proteins in Arabidopsis.** *Plant Mol Biol* 2011, **77:**501–511.
65. Lu H: **Dissection of salicylic acid-mediated defense signaling networks.** *Plant Signal Behav* 2009, **4:**713–717.
66. Tsuda K, Sato M, Stoddard T, Glazebrook J, Katagiri F: **Network properties of robust immunity in plants.** *PLoS Genet* 2009, **5:**e1000772.
67. Li G, Meng X, Wang R, Mao G, Han L, Liu Y, Zhang S: **Dual-level regulation of ACC synthase activity by MPK3/MPK6 cascade and its downstream WRKY transcription factor during ethylene induction in Arabidopsis.** *PLoS Genet* 2012, **8:**e1002767.
68. Qiu JL, Fiil BK, Petersen K, Nielsen HB, Botanga CJ, Thorgrimsen S, Palma K, Suarez-Rodriguez MC, Sandbech-Clausen S, Lichota J, Brodersen P, Grasser KD, Mattsson O, Glazebrook J, Mundy J, Petersen M: **Arabidopsis MAP kinase 4 regulates gene expression through transcription factor release in the nucleus.** *EMBO J* 2008, **27:**2214–2221.
69. Chen D, Xu G, Tang W, Jing Y, Ji Q, Fei Z, Lin R: **Antagonistic basic helix-loop-helix/bZIP transcription factors form transcriptional modules that integrate light and reactive oxygen species signaling in Arabidopsis.** *Plant Cell* 2013, **25:**1657–1673.
70. Underwood W, Zhang S, He SY: **The Pseudomonas syringae type III effector tyrosine phosphatase HopAO1 suppresses innate immunity in Arabidopsis thaliana.** *Plant J* 2007, **52:**658–672.
71. Schweighofer A, Kazanaviciute V, Scheikl E, Teige M, Doczi R, Hirt H, Schwanninger M, Kant M, Schuurink R, Mauch F, Buchala A, Cardinale F, Meskiene I: **The PP2C-type phosphatase AP2C1, which negatively regulates MPK4 and MPK6, modulates innate immunity, jasmonic acid, and ethylene levels in Arabidopsis.** *Plant Cell* 2007, **19:**2213–2224.
72. Zhang YY, Wu JW, Wang ZX: **Mitogen-activated protein kinase (MAPK) phosphatase 3-mediated cross-talk between MAPKs ERK2 and p38alpha.** *J Biol Chem* 2011, **286:**16150–16162.
73. Park HC, Song EH, Nguyen XC, Lee K, Kim KE, Kim HS, Lee SM, Kim SH, Bae DW, Yun DJ, Chung WS: **Arabidopsis MAP kinase phosphatase 1 is phosphorylated and activated by its substrate AtMPK6.** *Plant Cell Rep* 2011, **30:**1523–1531.
74. Rayapuram N, Bonhomme L, Bigeard J, Haddadou K, Przybylski C, Hirt H, Pflieger D: **Identification of Novel PAMP-Triggered Phosphorylation and Dephosphorylation Events in Arabidopsis thaliana by Quantitative Phosphoproteomic Analysis.** *J Proteome Res* 2014, **13:**2137–2151.
75. Gonzalez Besteiro MA, Ulm R: **Phosphorylation and stabilization of Arabidopsis MAP kinase phosphatase 1 in response to UV-B stress.** *J Biol Chem* 2013, **288:**480–486.
76. Lu D, Lin W, Gao X, Wu S, Cheng C, Avila J, Heese A, Devarenne TP, He P, Shan L: **Direct ubiquitination of pattern recognition receptor FLS2 attenuates plant innate immunity.** *Science* 2011, **332:**1439–1442.
77. Ellinger D, Naumann M, Falter C, Zwikowics C, Jamrow T, Manisseri C, Somerville SC, Voigt CA: **Elevated early callose deposition results in complete penetration resistance to powdery mildew in Arabidopsis.** *Plant Physiol* 2013, **161:**1433–1444.
78. Nishimura MT, Stein M, Hou BH, Vogel JP, Edwards H, Somerville SC: **Loss of a callose synthase results in salicylic acid-dependent disease resistance.** *Science* 2003, **301:**969–972.
79. Sato M, Tsuda K, Wang L, Coller J, Watanabe Y, Glazebrook J, Katagiri F: **Network modeling reveals prevalent negative regulatory relationships between signaling sectors in Arabidopsis immune signaling.** *PLoS Pathog* 2010, **6:**e1001011.
80. Xu J, Xie J, Yan C, Zou X, Ren D, Zhang S: **A chemical genetic approach demonstrates that MPK3/MPK6 activation and NADPH oxidase-mediated oxidative burst are two independent signaling events in plant immunity.** *Plant J* 2014, **77:**222–234.
81. Wang L, Tsuda K, Sato M, Cohen JD, Katagiri F, Glazebrook J: **Arabidopsis CaM binding protein CBP60g contributes to MAMP-induced SA accumulation and is involved in disease resistance against *Pseudomonas syringae*.** *PLoS Pathog* 2009, **5:**e1000301.
82. Kosetsu K, Matsunaga S, Nakagami H, Colcombet J, Sasabe M, Soyano T, Takahashi Y, Hirt H, Machida Y: **The MAP kinase MPK4 is required for cytokinesis in Arabidopsis thaliana.** *Plant Cell* 2010, **22:**3778–3790.
83. Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette ML, Mireau H, Peeters N, Renou JP, Szurek B, Taconnat L, Small I: **Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis.** *Plant Cell* 2004, **16:**2089–2103.
84. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15.
85. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**Article3.
86. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100:**9440–9445.
87. **VENNY.** [http://bioinfogp.cnb.csic.es/tools/venny/]
88. **AmiGO.** [http://amigo.geneontology.org/]
89. **ATTEDII.** [http://atted.jp/]
90. Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H: **ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis.** *Nucleic Acids Res* 2007, **35:**D863–D869.
91. **CATdb: a Complete Arabidopsis Transcriptome database.** [http://urgv.evry.inra.fr/CATdb/]
92. Biernacki C, Celeux G, Govaert G, Langrognet F: **Model-Based Cluster and Discriminant Analysis with the MIXMOD Software.** *Comput Stat Data Anal* 2006, **51:**587–600.
93. Efron B, Tibshirani R: **Empirical bayes methods and false discovery rates for microarrays.** *Genet Epidemiol* 2002, **23:**70–86.
94. Derozier S, Samson F, Tamby JP, Guichard C, Brunaud V, Grevet P, Gagnot S, Label P, Leple JC, Lecharny A, Aubourg S: **Exploration of plant genomes in the FLAGdb++ environment.** *Plant Methods* 2011, **7:**8.
95. Higo K, Ugawa Y, Iwamoto M, Korenaga T: **Plant cis-acting regulatory DNA elements (PLACE) database: 1999.** *Nucleic Acids Res* 1999, **27:**297–300.
96. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E: **AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks.** *Plant Physiol* 2006, **140:**818–829.
97. 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* 2010, **11:**166.
98. Braun P, Carvunis AR, Charloteaux B, Dreze M, Ecker JR, Hill DE, Roth FP, Vidal M, Galli M, Balumuri P, Bautista V, Chesnut JD, Kim RC, de LosReyes C, Gilles P, Kim CJ, Matrubutham U, Mirchandani J, Olivares E, Patnaik S, Quan R, Ramaswamy G, Shinn P, Swamilingiah GM, Wu S, Ecker JR, Dreze M, Byrdsong D, Dricot A, Duarte M, *et al*: **Evidence for network evolution in an Arabidopsis interactome map.** *Science* 2011, **333:**601–607.
99. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, *et al*: **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc* 2007, **2:**2366–2382.
100. Nakagami H, Soukupova H, Schikora A, Zarsky V, Hirt H: **A Mitogen-activated protein kinase kinase kinase mediates reactive oxygen species homeostasis in Arabidopsis.** *J Biol Chem* 2006, **281:**38697–38704.
101. Forzani C, Carreri A, de la Fuente van Bentem S, Lecourieux D, Lecourieux F, Hirt H: **The Arabidopsis protein kinase Pto-interacting 1–4 is a common target of the oxidative signal-inducible 1 and mitogen-activated protein kinases.** *Febs J* 2011, **278:**1126–1136.

102. Simon C, Langlois-Meurinne M, Bellvert F, Garmier M, Didierlaurent L, Massoud K, Chaouch S, Marie A, Bodo B, Kauffmann S, Noctor G, Saindrenan P: **The differential spatial distribution of secondary metabolites in Arabidopsis leaves reacting hypersensitively to Pseudomonas syringae pv. tomato is dependent on the oxidative burst.** *J Exp Bot* 2010, **61:**3355–3370.
103. Garcia AV, Blanvillain-Baufume S, Huibers RP, Wiermer M, Li G, Gobbato E, Rietz S, Parker JE: **Balanced nuclear and cytoplasmic activities of EDS1 are required for a complete plant innate immune response.** *PLoS Pathog* 2010, **6:**e1000970.

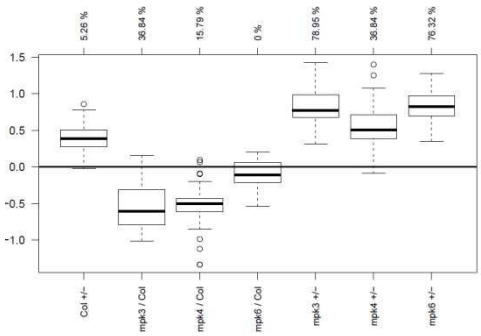
doi:10.1186/gb-2014-15-6-r87
Cite this article as: Frei dit Frey *et al.*: Functional analysis of *Arabidopsis* immune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defences. *Genome Biology* 2014 **15**:R87.

Annexe B
Figure supplémentaire S4 de l'article en annexe A :
Vue globale des profils d'expressions des 29 clusters de coexpression

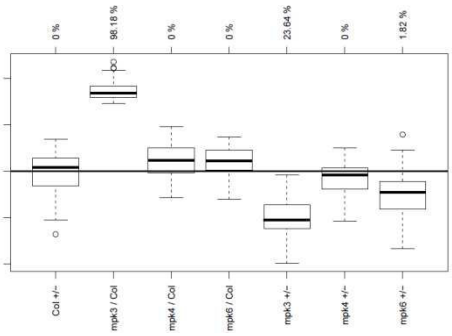
Figure S4


Cluster 1 (174 genes) (excluded)
Cluster 2 (114 genes)
Cluster 3 (122 genes)
Cluster 4 (116 genes)
Cluster 5 (126 genes)
Cluster 6 (132 genes)
Cluster 7 (121 genes)
Cluster 8 (155 genes)
Cluster 9 (142 genes)
Cluster 10 (161 genes)
Cluster 11 (133 genes)
Cluster 12 (147 genes)
Col +/-
mpk3 / Col
mpk4 / Col
mpk6 / Col
mpk3 +/-
mpk4 +/-
mpk6 +/-

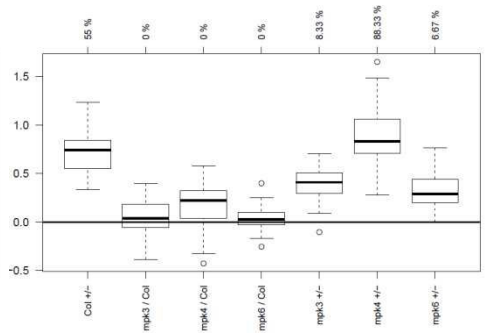
Cluster 13 / 1138 genes



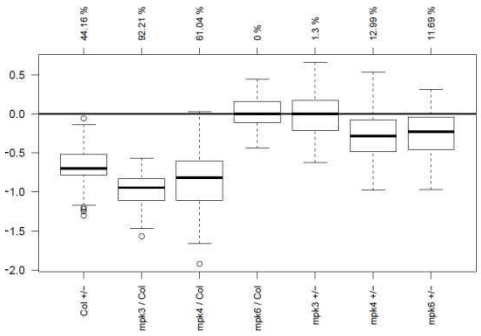
Cluster 14 / 155 genes



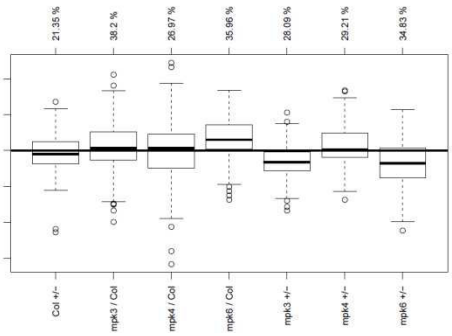
Cluster 15 / 158 genes



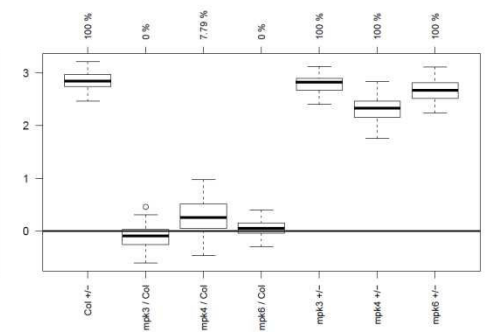
Cluster 16 / 174 genes



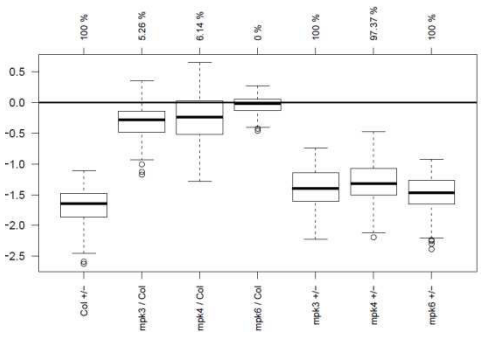
Cluster 17 / 184 genes **(excluded)**



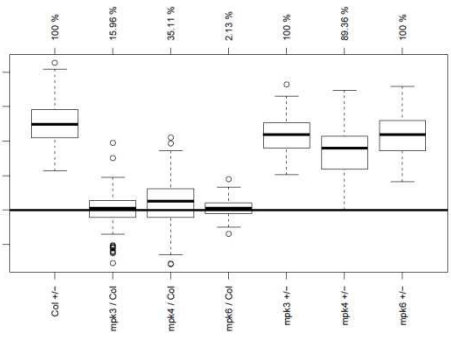
Cluster 18 / 174 genes



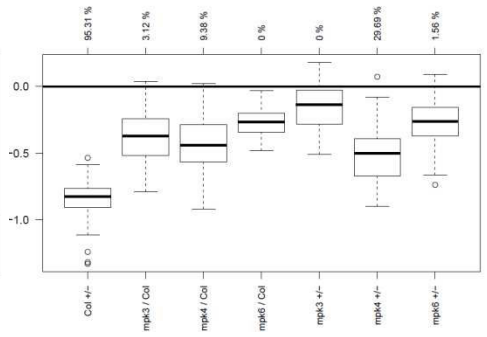
Cluster 19 / 1109 genes



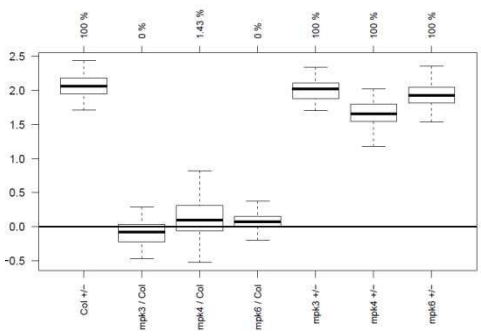
Cluster 20 / 192 genes



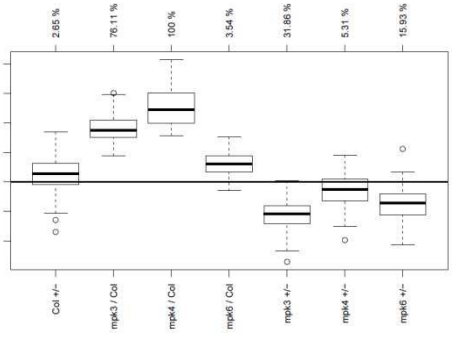
Cluster 21 / 163 genes



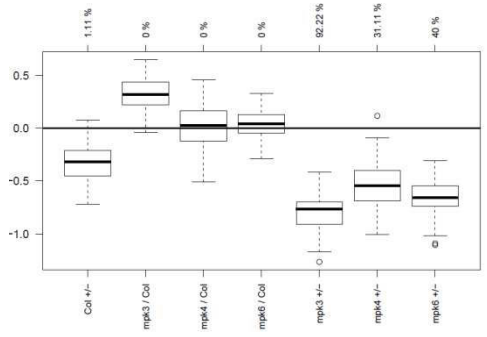
Cluster 22 / 166 genes



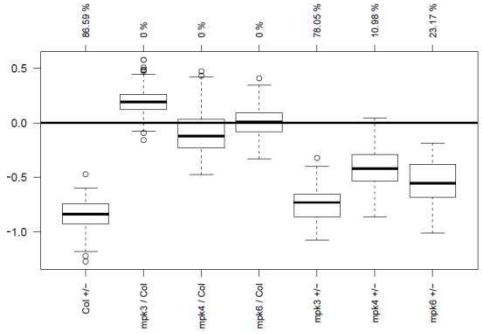
Cluster 23 / 111 genes



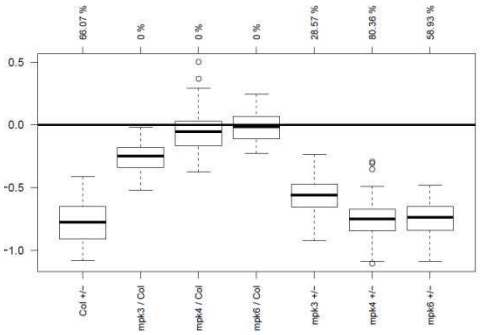
Cluster 24 / 191 genes



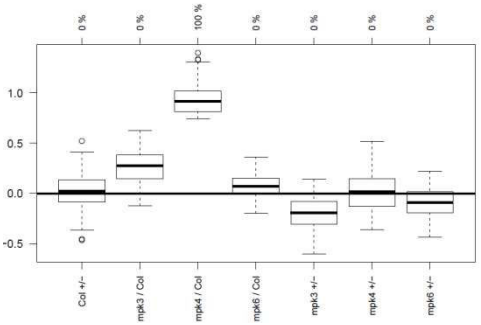
Cluster 25 / 181 genes



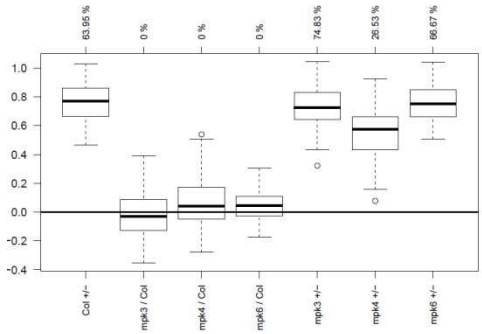
Cluster 26 / 156 genes



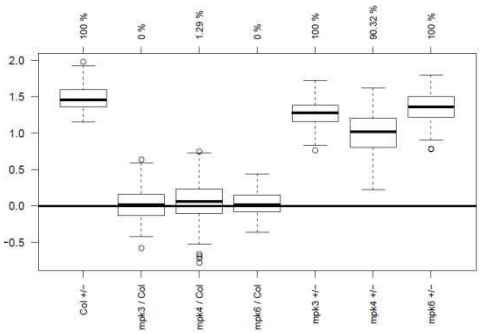
Cluster 27 / 162 genes



Cluster 28 / 146 genes



Cluster 29 / 153 genes



Annexe C

GEM2Net: from gene expression modeling to –omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response

Rim Zaag, Jean Philippe Tamby, Cécile Guichard, Zakia Tariq,
Guillem Rigaill, Etienne Delannoy, Jean-Pierre Renou, Sandrine Balzergue, Tristan Mary-Huard,
Sébastien Aubourg, Marie-Laure Martin-Magniette and Véronique Brunaud

Nucleic Acids Research Advance Access published November 11, 2014

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku1155

GEM2Net: from gene expression modeling to -omics networks, a new CATdb module to investigate *Arabidopsis thaliana* genes involved in stress response

Rim Zaag[1,2,†], Jean Philippe Tamby[1,2,†], Cécile Guichard[1,2,†], Zakia Tariq[1,2], Guillem Rigaill[1,2], Etienne Delannoy[1,2], Jean-Pierre Renou[1,2], Sandrine Balzergue[1,2], Tristan Mary-Huard[3,4,5], Sébastien Aubourg[1,2], Marie-Laure Martin-Magniette[1,2,3,4] and Véronique Brunaud[1,2,*]

[1]INRA, Unité de Recherche en Génomique Végétale, UMR 1165, ERL CNRS 8196, Saclay Plant Sciences, CP 5708, F-91057 Evry, France, [2]UEVE, Unité de Recherche en Génomique Végétale, UMR 1165, ERL CNRS 8196, Saclay Plant Sciences, CP 5708, F-91057 Evry, France, [3]INRA, UMR 518 MIA, 75005 Paris, France, [4]AgroParisTech, UMR 518 MIA, 75005 Paris, France and [5]UMRGV, INRA, Université Paris-Sud, CNRS, F-91190 Gif-sur-Yvette, Paris, France

Received September 12, 2014; Revised October 17, 2014; Accepted October 29, 2014

ABSTRACT

CATdb (http://urgv.evry.inra.fr/CATdb) is a database providing a public access to a large collection of transcriptomic data, mainly for Arabidopsis but also for other plants. This resource has the rare advantage to contain several thousands of microarray experiments obtained with the same technical protocol and analyzed by the same statistical pipelines. In this paper, we present GEM2Net, a new module of CATdb that takes advantage of this homogeneous dataset to mine co-expression units and decipher Arabidopsis gene functions. GEM2Net explores 387 stress conditions organized into 18 biotic and abiotic stress categories. For each one, a model-based clustering is applied on expression differences to identify clusters of co-expressed genes. To characterize functions associated with these clusters, various resources are analyzed and integrated: Gene Ontology, subcellular localization of proteins, Hormone Families, Transcription Factor Families and a refined stress-related gene list associated to publications. Exploiting protein–protein interactions and transcription factors-targets interactions enables to display gene networks. GEM2Net presents the analysis of the 18 stress categories, in which 17 264 genes are involved and organized within 681 co-expression clusters. The meta-data analyses were stored and organized to compose a dynamic Web resource.

INTRODUCTION

Although complete genome sequences are available for various organisms and despite the fact that it is now relatively easy to sequence a whole new genome and then to localize its genes, the functional annotation of these genes remains a big challenge. Hanson *et al.* (1) estimated that for eukaryotic organisms, whose genomes were completely sequenced, 20–40% of predicted genes do not have an assigned function. Even for the *Arabidopsis thaliana*, the first plant genome to be sequenced in 2000 (2) and for which a wealth of gene annotation is available, only 16% of all genes have a validated function and in fact 5105 genes are still orphan i.e. without any annotation about their function (see 'Orphan gene definition' in Materials and Methods). Arguably, the first functional annotation procedures, based on sequence similarities, have reached their limit due to the high complexity and heterogeneity of genomes activity. For the last 20 years, high-throughput technologies have made it possible to assess the behaviors of genes in a broader context and co-expression studies based on transcriptomic data are now considered to be a relevant approach to characterize and decipher the function of genes (3–6).

CATdb (7) can now contribute to take up this challenge. Originally CATdb was developed to manage the Arabidopsis microarrays data generated by the URGV

*To whom correspondence should be addressed. Tel: +33 1 60 87 45 14; Fax: +33 1 60 87 45 49; Email: brunaud@evry.inra.fr
† The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
Present address: Jean-Pierre Renou, IRHS, INRA, Beaucouze F-49071, France.

transcriptomic platform (http://www-urgv.versailles.inra.fr/microarray/). CATdb provides access to a large collection of transcriptomic data for *A. thaliana*, which were all obtained with the two-color CATMA (Complete Arabidopsis Transcriptome Micro Array) microarrays (8). The design of the CATMA microarrays has been regularly updated and includes almost all the annotated genes of Arabidopsis, combining the official annotation of TAIR (The Arabidopsis Information Resource (9)) with gene predictions from the EuGene software tool (10). Thus, the 'home made' CATMA microarrays are distinct from other Arabidopsis microarrays, for example CATMA includes 5095 genes that are not targeted by the commonly used ATH1 GeneChip® microarray from Affymetrix and this specificity has been successfully exploited in the past to discover new genes (11). The latest version of CATMA (CATMAv7) targets 35 656 genes. Finally, since its first release, the information system of CATdb has been upgraded to follow the technological developments of the platform. Currently, CATdb stores transcriptomic data for 20 distinct species obtained with four DNA chip technologies and totalizing 4838 CATMA, 1464 Affymetrix, 1208 NimbleGen and 48 Agilent microarrays. At present, the database includes 282 projects (231 for Arabidopsis and 51 for other species) and contains almost 10 000 hybridized samples (five times more than in 2008 for the first CATdb publication (1)), which are publicly accessible at http://urgv.evry.inra.fr/CATdb. For each project, a complete description going from the experimental design to the detailed features of the samples, the normalized intensity ratios and the results of differential expression analysis are available. The 4613 distinct samples represent 28 organs of the plant harvested at 132 development stages. Out of all samples, 1657 are extracted from plants treated with one of the 186 factors described in CATdb. Hence, the CATdb content increases regularly and includes a large range of experiments around the transcriptomic data for various plants.

In this article, we present a new module of CATdb called GEM2Net. Its main objective is to provide a global and comprehensive overview of the co-expression units of genes responding to a panel of stress stimuli. Our approach relies on a model-based clustering method applied to a carefully selected set of CATMA experiments dealing with the transcriptome of Arabidopsis under various stress conditions. Compared to other Arabidopsis resources like Genevestigator (12), the Stanford Microarray Database (13), GeneMania (14), MapMan (15) or ATTED-II (16,17), GEM2Net contains complementary and original features. As pointed out by Horan *et al.* (18), the results of such studies might suffer from the heterogeneous origins of the data. Moreover, in these latter resources, the co-expression is measured by computing the correlation between all pairs of gene across microarray datasets. In contrast our model-based clustering approach is meant to detect groups of genes and not only pairs. Besides, the discovery of new genes involved in plant response to biotic and abiotic stresses constitutes an important challenge in plant biology with relevance to agriculture and ecology since it could represent a potential starting point for plant breeding. To date, few databases related to plant stresses have been developed: the Stress-Responsive Transcription Factor Database (19), Arabidopsis Stress Responsive Gene Database (20) or Plant Stress Gene Database (21). These databases provide access to a list of curated stress genes extracted from literature, but this quality of information is incompatible with a global analysis. At a genome scale, Lan *et al.* (22) predicted new stress response genes by combining machine learning methods on genes with known functions and described by transcriptome data.

In this article, we developed the main original features of GEM2Net: (i) the global analysis of a homogeneous and dedicated transcriptomic dataset, (ii) the use of a model-based clustering approach to study gene co-expression and (iii) a set of bioinformatic developments and tools to integrate, analyze and visualize the meta-data that characterize each gene co-expression unit.

MATERIALS AND METHODS

Gene annotations from various resources are described in Supplementary Table S1.

Orphan gene definition

A Perl script was developed to identify genes that are orphan of function using TAIR (genome release R10) functional description of the 33 602 genes. A gene was considered as orphan of function if its description complies with these criteria: (i) no Gene Ontology (GO) annotation, or (ii) the terms 'unknown protein' or 'hypothetical protein' are set in biological process and molecular function from Gene Ontology and (iii) no known protein motif as defined by InterPro is associated. Following these criteria, 5105 genes have been determined as orphan of function in the Arabidopsis Reference set. We point out that this definition of orphans is restrictive and focuses on genes that are completely unknown.

Gene set enrichment tests

The enrichments of clusters in GO Slim terms, subcellular localization terms, orphan genes, transcription factors (TFs), hormones or stress-triggered genes in literature were assessed using a hypergeometric test to compare the number of genes in each cluster associated to the studied meta-data to its expected value in the genome (34 042 genes). Over-representation was declared statistically significant when the *P*-value was lower than 0.01. In the case of protein–protein interactions (PPI), we used a permutation approach to assess the significance of the number of PPI in a cluster. In brief, for every cluster, we counted the number of PPI and denoted this number k. Then we randomly sampled 1000 clusters of the same size and counted the number of PPI in each of those random clusters. We retrieved a *P*-value by computing the proportion of random clusters having a number of PPI larger than or equal to k. We considered *P*-values smaller than 5% to be significant.

Database and web implementation

CATdb was implemented using the PostgreSQL (v9.1.2) RDBMS. Attention was brought to the normalization process during conception, whereas pertinent indexes were created to improve performances with tables that need

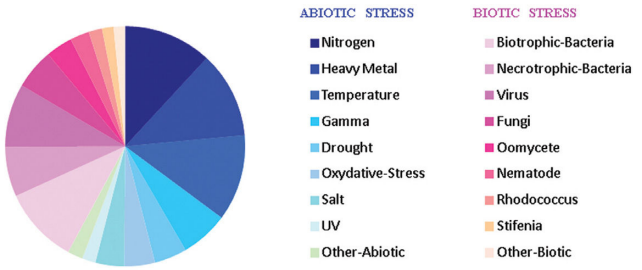


Figure 1. Stress categories. Pie chart representing the classification of the CATdb experimental comparisons into 18 stress categories, nine biotic and nine abiotic stresses.

to be intensively queried. A Web interface for accessing GEM2Net content was developed with the PHP (release 5.3.3) language to benefit from its integration as a module of Apache HTTP server (https://httpd.apache.org/). Object-oriented programming concepts were applied to take advantage of the easy maintainability qualities in development of a complex interface. Dynamic rendering was added using Javascript libraries (Jquery, Json) and functions, in order to facilitate the navigation between pages and then improve the user experience. The Cytoscape Web software tool (v1.0.4) was downloaded from http://cytoscapeweb.cytoscape.org and was integrated to the interface. It provides the useful interactivity for the visualization of networks obtained from PPI data or target genes links to TF data.

RESULTS

Transcriptomic dataset

To provide a global insight into the plant response to an environmental change or biological attack, a set of CATMA microarrays (387 dye-swaps) dedicated to biotic or abiotic stresses were extracted from CATdb. Their normalized expression differences were the starting point of this meta-analysis project. All these data have the rare advantage to have been generated with the same technical protocols and the same statistical pipelines including the normalization and the differential analysis (refer to Gagnot *et al.* (1) for details). Overall, GEM2Net explores 18 stress categories (Figure 1) describing nine biotic and nine abiotic stresses.

To define the set of genes to be considered in each stress category, some criteria were taken into account: (i) only genes for which a probe with a good specificity and without missing values were mined, (ii) raw *P*-values of these genes for each comparison were adjusted to control a global FWER (corrective term equals the number of genes by 387) and only genes for which at least one of the adjusted *P*-values was lower than 0.05 were kept, (iii) genes were removed if they were declared differentially expressed only once when the transcriptome comparison was made on two or more biological replicates. Finally a total of 17 264 genes were found to be differentially expressed in at least one stress experiment. By stress category, this number ranges from 1565 to 13 807.

Co-expression analysis

To directly identify co-expression units of several genes, the co-expression analysis was performed with a model-based clustering approach. The idea is that there exist unknown subpopulations that are observable through the expression differences. Model-based clustering aims at identifying this underlying structure in modeling the unknown distribution of transcription profiles by a mixture of parametric distributions, each one representing a subpopulation of genes. Since observations are the expression differences, a subpopulation represents a cluster of genes having the same dynamic of response across all the comparisons. Practically, multi-dimensional Gaussian mixtures of 2–100 subpopulations with unequal proportions were estimated with the MIX-MOD software (23). Covariance matrices were constrained so that their volumes were differing and their orientation and shape were equal. Among this collection of mixtures, the Bayesian Information Criterion (BIC) curve as a convex function of the number of subpopulations was checked to guarantee that the modeling fitted well the data. This assessment of the BIC behavior was the origin of the definition of the stress categories, since an analysis based on the 387 comparisons considered together led to an unstable behavior of BIC, proof of an issue of modeling. The best mixture according to BIC was selected to find the number of co-expression units and to perform the classification task. Two classification rules were applied: (i) all genes were classified according to the Maximum *A Posteriori* (MAP) rule by assigning each gene into the cluster for which the conditional probability is the highest and (ii) only genes with a highest conditional probability greater than a threshold were classified. This threshold was fixed for each analysis so that as many genes as possible were classified, under the constraint that the proportion of misclassified genes is controlled at a level of 5%. This classification rule is called Multi-class False Discovery Rate (MFDR) (24) and is an extension of the previously described BFDR (25).

Following this procedure, a total of 681 clusters equal to co-expression units were identified for the 18 stress categories. According to the MFDR rule, the percentage of classified genes with high confidence ranged from 20 to 67% and out of the 17 264 analyzed genes, only 8% of them (1469 genes) were never classified with high confidence. An example of a co-expression profile is presented in Supplementary Figure S1 and all the co-expression profiles are accessible in GEM2Net module. The quality of a cluster can be evaluated through the size of the boxplots: a boxplot with a reasonably small size means that all the genes have the same dynamic of response. Although the biological replicates are coherent at the level of the whole set of genes, the clustering method highlights that some biological comparisons may behave differently for a given subset of genes. Aubourg *et al.* (11) have already observed this phenomenon. On the top of the profile, the percentage of genes differentially expressed is indicated for each comparison. It can be useful to identify the comparisons that are relevant for the cluster under study.

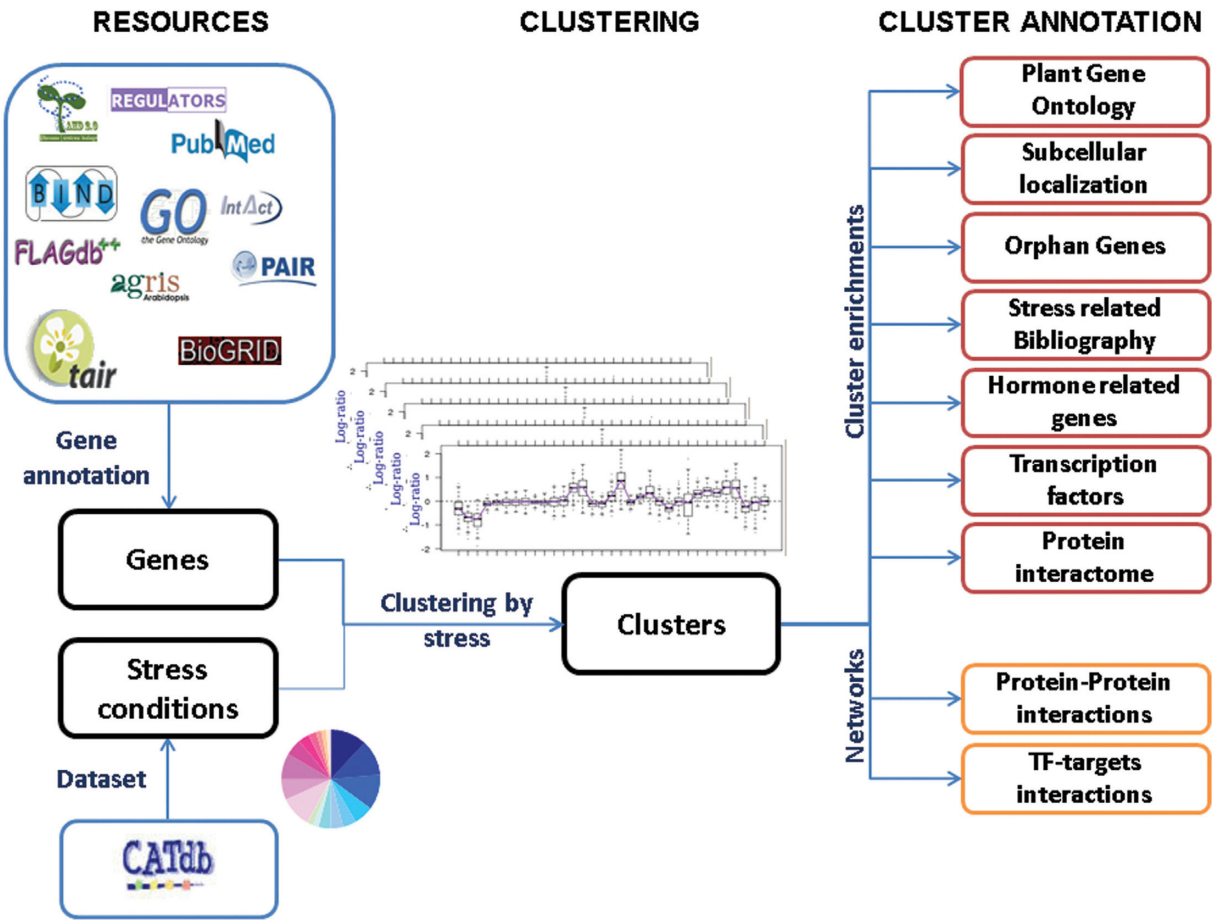


Figure 2. Workflow of GEM2Net. This workflow describes the bioinformatics steps required from the classification of CATdb experimental comparisons toward cluster annotation and gene interaction networks, with integration of the various meta-data.

Table 1. Number of genes by meta-data in GEM2Net gene set and Arabidopsis genome Reference

	Total	Orphan	BP stress	Bibliostress	TF	Hormone
Arabidopsis Reference	34 042	5105 (15%)	5106 (15%)	2580 (7.5%)	2260 (6.5%)	695 (2%)
GEM2Net dataset	17 264	2165 (13%)	**4003 (23%)**	**2064 (12%)**	**1578 (9%)**	487 (3%)

Comparison of the number of genes between Reference (all Arabidopsis genes) and GEM2Net dataset for the following meta-data: Orphan genes; BP stress gathers two terms of Biological Process from GO ('response to stress' and 'response to abiotic or biotic stress'); Bibliostress lists the stress-responsive genes with related bibliography extracted from GO; TF is a list of genes characterized as TFs in the Regulators project; Hormone is a list of genes having a link with hormone response as annotated in the AHD2.0 database. Numbers in bold highlight significant gene set enrichments of the GEM2Net compared to the Reference datasets (binomial test with P-value < 0.05).

From cluster annotation to gene function

Clustering of gene expression profiles has long been considered a fruitful approach to gain insight into gene function (26,27). It is based on the 'guilt by association' concept, which assumed that genes with similar expression profiles are likely to have similar functions. Once a cluster has been identified, it is common to perform an enrichment analysis to detect whether annotated genes share the same characteristics to give more clues about function of orphan genes of this cluster (28). Our model-based clusters were annotated with meta-data gained from multiple resources dedicated to Arabidopsis (see Supplementary Table S1 for more details on resources), the main one being TAIR, which is the reference site for the Arabidopsis genome. The latest release of Arabidopsis genome (TAIR version 10) was used in addition to the Plant Slim Gene Ontology (29), which gives a broad overview of the ontology content without the detail of the specific fine grained terms (40 terms). From this classification, we extracted a list of genes known to be involved in stress response and having an associated publication (we named this list Bibliostress). Afterward, the meta-data were refined by considering databases specialized in families of Arabidopsis genes for both significant functional groups involved in stress, i.e. the TFs from the Regulators project (30) and the hormone-related genes from the Arabidopsis Hormone Database (AHD 2.0) (31). Whatever the source of an-

Figure 3. GO Biological Process analyses for the 'Virus' stress category. The GEM2Net web page representing the GO 'Biological Process' pie charts for all the clusters of Virus stress category. Statistically significant results of gene set enrichment tests are displayed with colored sections, and gene counts and *P*-values are mentioned in the information frame on the right side. In the same frame, all analysis results are summarized with blue points for the cluster being hovered over with the mouse.

notation chosen, a cluster was qualified as 'enriched for an annotation' if the result of a hypergeometric test was significant for this cluster (P-value < 0.01). The main steps of the analyses workflow are described in Figure 2. A global view of the numbers of annotated genes regarding a series of meta-data is summarized in Table 1, and highlights some characteristics of the GEM2Net dataset compared to the Arabidopsis genome (the Reference). Out of the 5105 orphan genes from the Reference, 2165 (42%) were found to be impacted by stress in the GEM2Net dataset; this was close to the expected value of 50% in view of the proportion in all genes (17 264/34 042). Moreover, an equivalent distribution of orphan genes was found in biotic, abiotic and the individual 18 stress categories (Supplementary Table S2). The stability of the number of orphan genes points out a regular distribution of the knowledge level throughout stress categories, even if the number of experiments was not equivalent. In regard to the gene annotation related to stress, an expected enrichment is found for the GEM2Net dataset with 23% of genes annotated with GO terms linked to stress (BP stress column) against 15% in the Reference set, and 12% against 7.5% for the Bibliostress meta-data. An original characteristic of the GEM2Net gene set, which is visible in Table 1, is the enrichment in TFs (9 against 6.5% for the Reference), a class of genes that are essential to regulate transcription of other genes, especially in the response to stress.

Concerning the cluster annotation enrichment analyses, 98% of the clusters have a functional bias in at least one GO term and 80% are associated to the stress term 'response to stress' or 'response to abiotic or biotic stress'. Despite the expected bias in stress category, other biases are found and make it possible to decipher the functions of orphan genes. For instance, 63% of the clusters are enriched in the GO term 'transport' and 39% are enriched in 'plastid' as a prediction of the subcellular localization of proteins. These numerous enrichments indicate that our large-scale co-expression study generates biologically meaningful clusters and performs favorably as compared to those obtained with correlation-based approaches by Heyndrickx *et al.* (32).

Visualization of meta-data

The GEM2Net Web interface (http://urgv.evry.inra.fr/GEM2NET) allows users to query the database by stress category or to submit a list of genes of interest to retrieve the stress categories they are implicated in. It is possible to answer questions like 'are my genes of interest involved in the same co-expression unit?' by exploring graphically the results of meta-data analyses. To identify at a glance and summarize the potential functional biases of each cluster, GEM2Net proposes an original representation and interactive visualization, using pie charts and graphs. For each stress category, cluster annotation analyses are divided into several tabs, one per meta-data type, which allow the or-

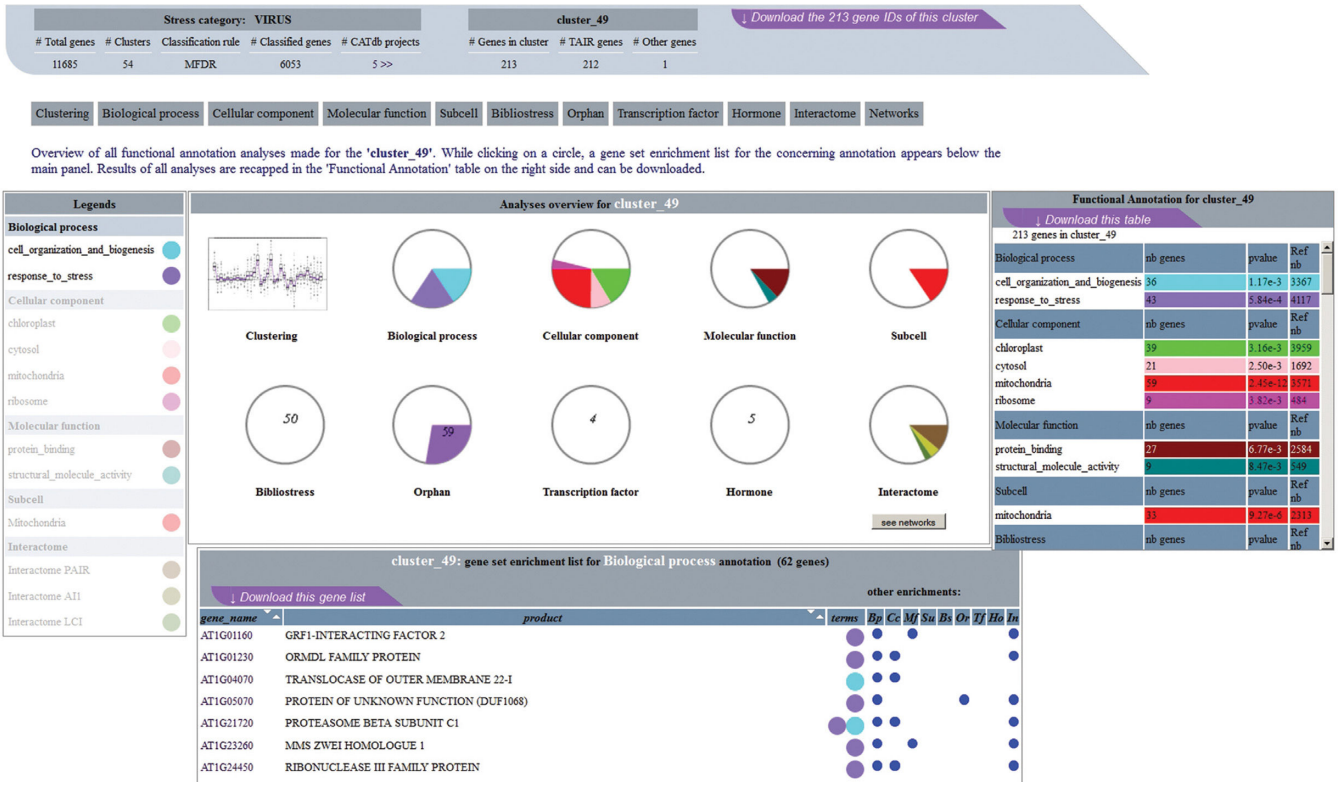


Figure 4. Meta-analyses overview for Cluster_49 of the 'Virus' stress category. Synoptic view of the meta-data analyses performed on the cluster_49 is shown in the central panel and results of all analyses are summarized in the frame table on the upper right side. Part list of the genes involved in the Biological Process bias is seen below the central panel. In this table, each gene accession is tagged with colored circle(s) (legend table on the left) and other meta-data enrichments are indicated on the right with blue points when appropriate.

ganization of complex and abundant results. In such tabs, a central panel displays every cluster as a pie chart whose diameter is directly related to the total number of genes in the cluster. Colored sections within a pie chart materialize the numbers of genes of the corresponding meta-data when over-represented. For example, Figure 3 presents a view of clusters enrichment analyses for the GO Slim branch 'Biological Process' of the 'Virus' stress category. Complementary information related to each cluster is also available in a table nearby the central panel, which appears when moving the mouse over a given cluster. In Figure 4, a synthetic view of one cluster, the cluster_49 from the 'Virus' stress category, shows a summary of all the meta-data analyses done for this cluster. In each pie chart, significant enrichments are seen in color for few meta-data analyses, e.g. in the GO terms 'cell organization and biogenesis' and 'response to stress' of the 'Biological Process' ontology. The complete list of 213 genes of cluster_49 appears in a table sorted by gene roles in the selected meta-data type and highlights their implication in other biases by colored circles.

GEM2Net integrates protein interactomes and targets of TF data from external resources (Supplementary Table S1), so gene clusters can be viewed as interactive biological networks (Figure 5), thanks to the embedded Cytoscape Web software tool (33). Adding gene annotation, e.g. TFs families, GO terms or choosing experimentally confirmed PPI, makes it easier to predict regulatory networks of biological relevance and to identify new functional partners. Therefore, by combining gene annotation corresponding to relevant meta-analyses with the available interaction networks in GEM2Net, it is possible to gather clues to infer the biological functions of some orphan genes in a co-expression unit.

DISCUSSION

The new CATdb module, GEM2Net, was developed to summarize the transcriptomic responses of Arabidopsis to various stress conditions. The goal is to provide new information that will give new insights into plant stress responses and orphan genes involved in these responses when they are crossed with other experimental data or knowledge. The major outputs of GEM2Net are (i) the classification of several CATdb projects in biotic and abiotic stress categories, (ii) a global co-expression analysis using a model-based clustering approach that is not so often used in genomic analyses, (iii) a visualization system that summarizes rapidly the cluster annotation enrichments in terms of Gene Ontology, genes cross-referenced in stress-related bibliography, hormone and TF families, (iv) the gene interaction networks constructed with protein interactome data or TF-target interactions that are involved in each cluster. To progress further, we plan to integrate other types of meta-data, such as the *cis*-regulatory motifs detected in promoters of genes belonging to a same cluster (34). This may give the possibility to associate a *cis*-regulatory motif to a par-

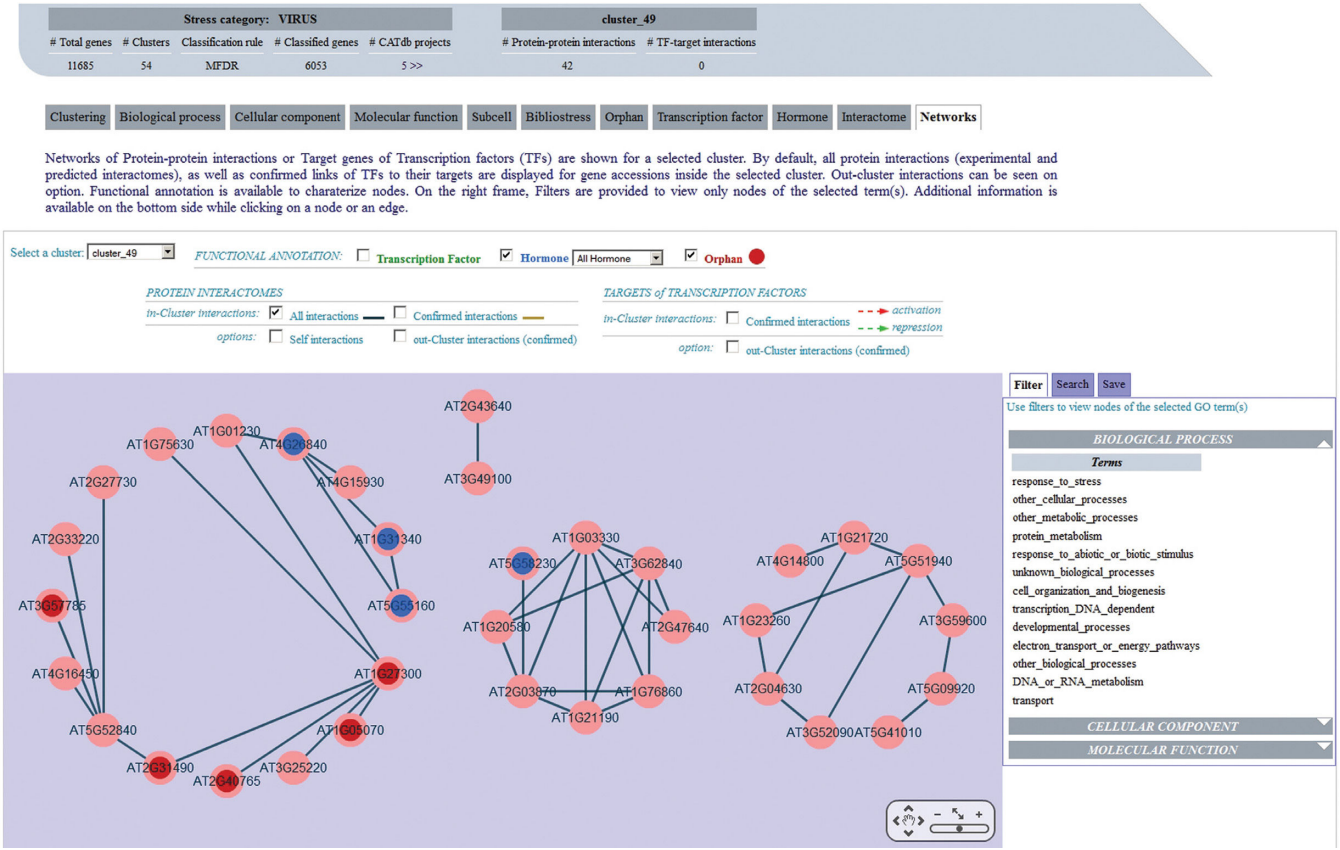


Figure 5. Protein Interactome Network for Cluster_49 of the 'Virus' stress category. In the central panel, all PPI (edges) between gene accessions (nodes) within the cluster_49 are represented with dark blue lines, using the Cytoscape Web software tool. Functional annotation is superimposed on nodes by selecting the corresponding checkbox above, hormone families (in blue) and orphans (in red) here. On the right frame, filters on GO categories can be applied to the network to view only nodes of the selected annotation. In addition, a 'Targets of Transcription Factors' option is available to display this type of interaction in the same network.

ticular stress stimulus, thus providing a valuable resource to complete interactions between TFs and their targets. To refine the co-expression units, we would update data with new transcriptome comparisons in each stress category. Thanks to the model-based clustering approach, it does not necessitate restarting completely the co-expression analysis. For instance, it was recently applied to investigate the roles of MAP kinases in Arabidopsis immune response to a microbial stress (35). Therefore, GEM2Net is an ongoing project and new meta-analyses will be released in the future to share the results with a large scientific community.

To conclude, GEM2Net aims at taking advantage of using controlled data associated to consistent meta-data to define relevant co-expression clusters, to improve their annotation and, thus to enhance the predictive power of assigning the right functions to orphan genes. Besides, since the plant responses to abiotic and biotic stresses are regulated by complex signaling networks and are associated with massive changes in gene expression, the GEM2Net results also allow building new hypotheses that might be the starting points for future biological projects.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

The authors thank G. Celeux (INRIA), C. Maugis-Rabusseau (INSA Toulouse), S. Huet (INRA), C. Keribin (University of Paris-Sud), N. Verzelen (INRA) and C. Giraud (University of Paris-Sud) for helpful discussions within the working group SONATAStat. The authors thank Philippe Grevet for computer system administration and PostgreSQL management at URGV.

FUNDING

Funding for open access charge: National Institute of Agricultural Research (INRA): Plant Biology and Breeding division and Applied Mathematics and Informatics division.
Conflict of interest statement. None declared.

REFERENCES

1. Hanson,A.D., Pribat,A., Waller,J.C. and de Crécy-Lagard,V. (2009) 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list—and how to find it. *Biochem. J.*, **425**, 1–11.
2. Genome Initiative,Arabidopsis. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. (2000) *Nature*, **408**, 796–815.
3. Shaik,R. and Ramakrishna,W. (2013) Genes and co-expression modules common to drought and bacterial stress responses in Arabidopsis and rice. *PLoS One*, **8**, e77261.

4. Dozmorov,M.G., Giles,C.B. and Wren,J.D. (2011) Predicting gene ontology from a global meta-analysis of 1-color microarray experiments. *BMC Bioinformatics*, **12**, S14.
5. Wren,J.D. (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics*, **25**, 1694–1701.
6. Usadel,B., Obayashi,T., Mutwil,M., Giorgi,F.M., Bassel,G.W., Tanimoto,M., Chow,A., Steinhauser,D., Persson,S. and Provart,N.J. (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ.*, **32**, 1633–1651.
7. 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 Res.*, **36**, D986–D990.
8. Crowe,M.L., Serizet,C., Thareau,V., Aubourg,S., Rouzé,P., Hilson,P., Beynon,J., Weisbeek,P., van Hummelen,P., Reymond,P. *et al.* (2003) CATMA—a complete Arabidopsis GST database. *Nucleic Acids Res.*, **31**, 156–158.
9. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
10. Foissac,S., Gouzy,J., Rombauts,S., Mathe,C., Amselem,J., Sterck,L., Van de Peer,Y., Rouzé,P. and Schiex,T. (2008) Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinformatics*, **3**, 87–97.
11. Aubourg,S., Martin-Magniette,M.L., Brunaud,V., Taconnat,L., Bitton,F., Balzergue,S., Jullien,P.E., Ingouff,M., Thareau,V.,, Schiex,T. *et al.* (2007) Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. *BMC Genomics*, **8**, 401–410.
12. Hruz,T., Laule,O., Szabo,G., Wessendorp,F., Bleuler,S., Oertle,L., Widmayer,P., Gruissem,W. and Zimmermann,P. (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinformatics*, **2008** , 420747.
13. Hubble,J., Demeter,J., Jin,H., Mao,M., Nitzberg,M., Reddy,T.B., Wymore,F., Zachariah,Z.K., Sherlock,G. and Ball,C.A. (2009) Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Res.*, **37**, D898–D901.
14. Zuberi,K., Franz,M., Rodriguez,H., Montojo,J., Lopes,C.T., Bader,G.D. and Morris,Q. (2013) GeneMANIA prediction server 2013 update. *Nucleic Acids Res.*, **41**, W115–W122.
15. Klie,S. and Nikoloski,Z. (2012) The choice between MapMan and Gene Ontology for automated gene function prediction in plant science. *Front Genet.*, **3**, 115.
16. Obayashi,T., Kinoshita,K., Nakai,K., Shibaoka,M., Hayashi,S., Saeki,M., Shibata,D., Saito,K. and Ohta,H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res.*, **35**, D863–D869.
17. Obayashi,T., Nishida,K., Kasahara,K. and Kinoshita,K. (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell Physiol.*, **52**, 213–219.
18. Horan,K., Jang,C., Bailey-Serres,J., Mittler,R., Shelton,C., Harper,J.F., Zhu,J.K., Cushman,J.J.C., Gollery,M. and Girke,T. (2008) Annotating genes of known and unknown function by large-scale co-expression analysis. *Plant Physiol.*, **147**, 41–57.
19. Naika,M., Shameer,K., Mathew,O.K., Gowda,R. and Sowdhamini,R. (2013) STIFDB2: an updated version of plant stress-responsive transcription factor database with additional stress-signals, stress-responsive transcription factor binding sites and stress-responsive genes in Arabidopsis and rice. *Plant Cell Physiol.*, **54**, e8.
20. Borkotoky,S., Saravanan,V., Jaiswal,A., Das,B., Selvaraj,S., Murali,A. and Lakshmi,P.T. (2013) The Arabidopsis Stress Responsive Gene Database. *Int. J. Plant Genomics*, **2013**, 949564.
21. Prabha,R., Ghosh,I. and Singh,D.S. (2011) Plant Stress Gene Database: a collection of plant genes responding to stress condition. *ARPN J. Sci. Technol.*, **1**, 28–31.
22. Lan,H., Carson,R., Provart,N.J. and Bonner,A.J. (2007) Combining classifiers to predict gene function in Arabidopsis thaliana using large-scale gene expression measurements. *BMC Bioinformatics*, **8**, 358–374.
23. Biernacki,C., Celeux,G., Govaert,G. and Langrognet,F. (2006) Model-based cluster and discriminant analysis with the MIXMOD software. *Comput. Stat. Data Anal.*, **51**, 587–600.
24. Mary-Huard,T., Perduca,V., Martin-Magniette,M.L. and Blanchard,G. (2013) Error rate control for classification rules in multi-class mixture models. In: *45e, J. Statistique, SFDS Proceedings*. Toulouse.
25. Efron,B. and Tibshirani,R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.*, **23**, 70–86.
26. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 14863–14868.
27. Schöner,D., Barkow,S., Bleuler,S., Wille,A., Zimmermann,P., Bühlmann,P., Gruissem,W. and Zitzler,E. (2007) Network analysis of systems elements. *EXS*, **97**, 331–351.
28. Atias,O., Chor,B. and Chamovitz,D.A. (2009) Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network. *BMC Syst. Biol.*, **3**, 86–107.
29. Berardini,T.Z., Mundodi,S., Reiser,R., Huala,E., Garcia-Hernandez,M., Zhang,P., Mueller,L.M., Yoon,J., Doyle,A., Lander,G. *et al.* (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol.*, **135**, 1–11.
30. Castrillo,G., Turck,F., Leveugle,M., Lecharny,A., Carbonero,P., Coupland,G., Paz-Ares,J. and Oñate-Sanchez,L. (2011) Speeding cis-trans regulation discovery by phylogenomic analyses coupled with screenings of an arrayed library of Arabidopsis transcription factors. *PLoS One*, **6**, e21524.
31. Jiang,Z., Liu,X., Peng,Z., Wan,Y., Ji,Y., He,W., Wan,W., Luo,J. and Guo,H. (2011) AHD2.0: an update version of Arabidopsis Hormone Database for plant systematic studies. *Nucleic Acids Res.*, **39**, D1123–D1129.
32. Heyndrickx,K.S. and Vandepoele,K. (2012) Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiol.*, **159**, 884–901.
33. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
34. Bernard,V., Lecharny,A. and Brunaud,V. (2010) Improved detection of motifs with preferential location in promoters. *Genome*, **9**, 739–752.
35. 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. *et al.* (2014) Functional analysis of Arabidopsis immune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defenses. *Genome Biol.*, **15**, R87.

Annexe F

Liste des gènes prédits positifs au moins 80 fois pour les 8 termes spécifiques retenus par la méthode d'apprentissage

Liste des gènes prédits positifs au moins 80 fois pour les 8 termes spécifiques retenus par la méthode d'apprentissage. Pour chaque terme caractérisé l'ontologie, le terme et les paramètres (seuilCorégulation, classement, décompte) sont précisés. Pour chaque identifiant de gène, les informations indiquant si le gène est orphelin (0 ou 1) ainsi que le nombre de fois où ce gène est déclaré positif sont fournies.

CC, cytosol, (2, Independant, Occurence)
AT1G73940 1 100
CC, chloroplast, (2, Dépendant, Occurence)
AT2G31725 1 100
AT4G02790 0 100
AT1G50732 1 100
AT5G55650 1 100
AT1G78995 1 100
AT3G01660 0 100
AT3G28460 0 100
AT5G57345 1 100
AT5G26960 0 100
AT1G17200 1 100
AT1G16320 1 100
AT5G14970 1 100
AT1G02205 0 100
AT5G27730 1 100
AT3G27180 0 100
AT2G30150 0 99
AT3G50790 0 99
AT1G15260 1 99
AT5G61412 1 99
AT3G54600 0 96
CC, ribosome, (2, IndépendantAmélioré, Occurence)
AT1G73940 1 100
CC, plastid, (3, IndépendantAmélioré, Occurence)
AT2G31725 1 100
AT4G02790 0 82
BP, response_to_abiotic_or_biotic_stimulus, (7, IndépendantAmélioré, Occurence)
AT4G27652 1 100
AT4G08540 0 100
AT1G07135 0 100
AT3G57450 1 100
AT2G26530 1 100
AT3G10930 1 100
BP, response_to_stress, (7, IndépendantAmélioré, Occurence)
AT4G27652 1 100
AT4G08540 0 100
AT1G07135 0 100
AT3G57450 1 100
AT2G26530 1 100

AT3G10930	1	100
AT5G42090	0	82
AT5G64460	0	82

BP, protein_metabolism, (4, Independant, Occurence)

AT3G49040	0	100
AT1G29250	0	100
AT1G73940	1	100
AT5G07020	0	100
AT5G53490	0	100
AT5G58250	1	100
AT3G56010	1	100
AT3G08920	0	100
AT4G02530	0	100
AT3G59840	1	100
AT2G01755	1	82

MF, structural_molecule_activity, (3, Dépendant, Occurence)

AT3G49040	0	100
AT1G52930	0	100
AT1G73940	1	100
AT5G19650	0	100
AT2G43780	1	100
AT2G41650	1	100
AT2G30990	1	100
AT1G56110	0	100
AT3G59650	0	100
AT4G30220	0	100
AT5G58250	1	95
AT3G59840	1	85

université PARIS-SACLAY | ÉCOLE DOCTORALE Structure et dynamique des systèmes vivants (SDSV)

Titre : Enrichissement de profils transcriptomiques par intégration de données hétérogènes : annotation fonctionnelle de gènes d'*Arabidopsis thaliana* impliqués dans la réponse aux stress.

Mots clés : annotation fonctionnelle, données hétérogènes, intégration, transcriptome, réseaux de gènes et apprentissage supervisé

Résumé : À l'ère de la biologie computationnelle, l'annotation fonctionnelle reste un défi central. Les méthodes d'annotation récentes reposent sur l'hypothèse d'association par culpabilité et s'appuient sur l'intégration de données pour la recherche de partenaires fonctionnels. Cependant, la majorité de ces méthodes souffrent de l'hétérogénéité des données et du manque de spécificité du contexte biologique, ce qui expliquerait le taux élevé de faux positifs parmi les prédictions.
Ce travail de thèse développe une approche intégrative de données moléculaires contrôlant leur hétérogénéité pour annoter des gènes d'*Arabidopsis thaliana* impliqués dans la réponse aux stress.
Les contributions majeures de cette thèse sont: (1) l'annotation fonctionnelle de groupes de gènes coexprimés par l'intégration de données omiques. (2) la construction d'un réseau de corégulation par une analyse transversale des groupes coexprimés qui renforce les liens fonctionnels entre les gènes. (3) le développement d'une méthode d'apprentissage supervisé pour l'inférence de fonction centrée sur les termes de la GO Slim en contrôlant le FDR. En identifiant une règle de décision par terme, cette méthode a permis de prédire la fonction de 47 gènes partiellement annotés ou orphelins.

Title : Enrichment of transcription profiles by integration of heterogeneous data: functional annotation of *Arabidospis thaliana* genes involved in stress responses.

Keywords : functional annotation, heterogeneous data, integration, transcriptome, gene Networks and supervised machine learning.

Abstract : In the era of computational biology, functional annotation remains a major challenge. Recent annotation methods are based on the guilt by association assumption and rely on data integration to identify functional partners. However, most of these methods suffer from data heterogeneity and a lack of biological context specificity which would probably explain the high rate of false positives among predictions.
This thesis develops an approach of molecular data integration controlling their heterogeneity in order to annotate *Arabidopsis thaliana* genes involved in stress response. The major contributions of this thesis are:
: (1) functional annotation of groups of co-expressed genes by omics data integration (2) the construction of a coregulatory gene network through a cross-analysis of the coexpressed groups strengthening the functional links between genes (3) the development of a supervised learning method for the inference of gene function centered on the GO Slim terms with a control of the FDR. By identifying a decision rule by term, this method was used to predict the function of 47 orphan or partially annotated genes.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France