

# UNIVERSITÉ DE TOURS

**ÉCOLE DOCTORALE SSBCV**

**EA2106 – Biomolécules et Biotechnologies Végétales**

**THÈSE** présentée par :

**Franziska LIESECKE**

soutenue le : 2 octobre 2018

pour obtenir le grade de : **Docteur de l'Université de Tours**

Discipline/ Spécialité : Sciences de la Vie et de la Santé / Bio-informatique

**"Coupable par association"  
Exploitation de ressources  
transcriptomiques pour la construction  
de réseaux de co-expression de gènes  
dédiés à l'élucidation de voies  
cellulaires**

**THÈSE dirigée par :**

**Dr PICHON Olivier**

Maître de conférences (HDR), Université de Tours

**Co-encadrée par :**

**Dr DUGE DE BERNONVILLE Thomas**

Maître de conférences, Université de Tours

**RAPPORTEURS :**

**Dr AUBOURG Sébastien  
Pr DELAVAULT Philippe**

Directeur de recherche, INRA Angers  
Professeur, Université de Nantes

**JURY :**

**Dr AUBOURG Sébastien  
Dr CARPIN Sabine  
Pr DELAVAULT Philippe  
Dr DUGE DE BERNONVILLE Thomas  
Pr HUGUET Elisabeth  
Dr PICHON Olivier**

Directeur de recherche, INRA Angers  
Maître de conférences (HDR), Université d'Orléans  
Professeur, Université de Nantes  
Maître de conférences, Université de Tours  
Professeur, Université de Tours  
Maître de conférences (HDR), Université de Tours



A ma famille  
A Hélène

« There is a way out of every box, a solution to every puzzle ; it's just a matter of finding it. »  
Captain Jean-Luc Picard



# Remerciements

« Science is not only a discipline of reason but, also one of romance and passion. »

Stephen Hawking

La thèse est une véritable aventure; et comme tout périple, il ne peut s'effectuer seul. Je suis heureuse qu'au-delà de l'expérience scientifique que j'ai pu acquérir durant ses années, cette période a également été marquée par de nombreuses rencontres qui me marqueront longtemps. Je sais que je n'aurais jamais pu mener à bien ce projet sans vous tous, et il serait impensable de ne pas vous consacrer une petite partie de ce manuscrit.

En premier lieu, je tiens à remercier les membres de mon jury de thèse de m'avoir fait l'honneur d'accepter d'évaluer ce travail.

Le Professeur Elisabeth Huguet pour avoir accepté de présider mon jury de thèse. Je profite de ces lignes pour vous remercier sincèrement de toute la bienveillance que vous avez pu avoir à mon égard durant mes études de licence et de master au sein de l'université.

Le Professeur Philippe Delavault ainsi que le Dr Sébastien Aubourg pour avoir accepté d'être rapporteurs de ce travail de thèse. Je vous remercie chaleureusement pour toutes les critiques constructives et les conseils au cours de ce projet et qui ont grandement contribué à améliorer ce travail.

Le Dr Sabine Carpin pour les discussions intéressantes au Biotechnocentre 2015 à Seillac, et d'avoir accepté d'examiner ce travail de thèse.

J'exprime toute ma gratitude au Professeur Loïc Vaillant, ancien Président de l'Université de Tours ainsi qu'au Professeur Emmanuel Lesigne pour le financement de ce doctorat et sa prolongation partielle, rendant ce travail possible.

Je remercie le Professeur Nathalie Guivarc'h pour son accueil au sein du laboratoire Biomolécules et Biotechnologies Végétales de l'Université de Tours.

Je voudrais bien sûr sincèrement remercier mon directeur de thèse, Dr Olivier Pichon. Non seulement pour l'encadrement de mon doctorat, mais également pour ses conseils et sa



disponibilité tout au long de mes études. Je te remercie également pour ta bonne humeur tout au long de ce travail.

Je remercie également le Professeur Joël Crèche d'avoir accepté d'encadrer cette thèse à ses débuts ainsi que pour ses nombreuses anecdotes scientifiques et sur la vie du laboratoire.

Il serait totalement inimaginable de ne pas remercier mon encadrant de thèse, Dr Thomas Dugé de Bernonville. Je tiens sincèrement à te remercier du fond du cœur pour ces années de travail commun et je dois dire que cela a été pour moi un immense plaisir d'être ton disciple ! Cette expérience de la thèse est de loin la plus importante de ma vie jusqu'à maintenant, et je suis heureuse d'avoir pu effectuer ce bout de chemin avec toi. Je te remercie d'avoir accepté d'encadrer ce travail malgré les contraintes que cela amenait. Tu as su et toujours pris la peine de t'adapter à mes besoins parfois un peu particuliers (merci pour les innombrables cafés !). Tu as également été toujours présent pour me guider lorsque cela a été nécessaire tout en me laissant l'autonomie nécessaire pour évoluer par moi-même et ainsi m'apporter la confiance pour m'aventurer sur de nouveaux chemins. Je suis persuadée que cela a fait de moi quelqu'un de bien plus audacieux et constituera un atout précieux pour mes futurs travaux de recherche ; je te remercie sincèrement pour cela. Ton investissement, ton sens du détail, ta volonté d'aller toujours plus loin et d'apporter toujours une dimension supplémentaire à notre travail ont été une grande inspiration pour moi. Tu as, avec ton infinie patience et ton enthousiasme, su me transmettre le virus de la bioinfo, ce qui n'était tout de même pas gagné au départ ! Enfin, je ne peux qu'être reconnaissante pour ta gentillesse et ta bonne humeur (et les joutes verbales !) qui ont largement contribué au bon déroulement de cette thèse. C'est non sans fierté que je regarde le chemin parcouru pendant ce projet de recherche et son évolution colossale (ha, casé!), de ses balbutiements en Master 2 à ce manuscrit de thèse. Je te souhaite le meilleur pour ton avenir tant professionnel que personnel et je compte sur toi pour voir Tetrark sur scène au Hellfest d'ici quelques années ! \m/

Puisqu'il serait un peu trop long de nommer tout le monde (il y a quand même un manuscrit à lire après), je tiens à chaleureusement remercier ici l'ensemble des membres de BBV, permanents ou de passage pour une période plus ou moins longue, pour leur accueil chaleureux et leur disponibilité. Il va falloir que je trouve de nouveaux cobayes pour la pâtisserie ! :-) Je profite de ces pages pour remercier sincèrement l'ensemble de l'équipe pédagogique « Sciences et Techniques » du laboratoire. Je vous remercie d'avoir toujours fait



preuve de bonne volonté et d'avoir été si disponible pendant mon parcours d'études, me permettant ainsi d'acquérir les bases nécessaires pour ce doctorat. Un remerciement particulier à Dr Nadine Imbault pour m'avoir guidée et écoutée durant toutes ces années ainsi qu'au Professeur Benoit Saint Pierre pour son soutien pendant mes années de Master !

Je voudrais également adresser mes plus sincères remerciements à Alexandra Elbakyan pour son travail inestimable et son courage qui ont grandement permis d'établir le socle de ce travail. « Science is better when it's open ! »

Un grand merci à Yann Jullian qui maîtrise le code informatique aussi bien que la raquette de tennis. Ton aide lors de mes premiers pas sur le cluster de calcul « Neptune », ta disponibilité et ta pédagogie ont été réellement précieux pour ce travail. Je remercie également les membres de CaSciModOT, d'abord pour l'espace de calcul, mais également pour les régulières journées de conférences et les discussions variées et toujours passionnantes !

Je voudrais particulièrement remercier Dr Christelle Dutilleul et Dr Eric Ducos pour leur encadrement de stage en Master 1 qui a fait germer en moi ma véritable passion pour la recherche ; et rendre la prénylation sexy, ce n'est pas donné à tout le monde ! Je remercie Dr Johan-Owen de Craene alias « Soleil couchant » pour les longues argumentations sur des sujets aussi variés que les maladies ciliaires, la bio-informatique manuelle ou des questions plus fondamentales du type « l'être humain a-t-il besoin de se nourrir ? » (étant d'un affligeant ennui, je soutenais la thèse du oui). Je suis heureuse d'avoir pu profiter de ton expérience scientifique et cosmopolite. Dankeschön Dr Marc Clastre pour ton ouverture d'esprit extraordinaire et tes points de vue anticonformistes. Je te remercie pour les échanges d'idées réellement passionnants et enrichissants pour ma part. Tu resteras toujours mon scientifique « out of the box » favori ! Je remercie également Dr Arnaud Lanoue pour les nombreuses conversations et les anecdotes sur ma patrie. Cela a toujours été un plaisir de partager un moment avec toi. Je remercie Dr Vincent Courdavault pour sa pédagogie et sa disponibilité lors mes études ainsi que pour les animations musicales dans les couloirs du labo. Merci au Dr Sébastien Bessau pour les discussions vidéoludiques ! Un merci également au Dr Gaëlle Glevarec pour les nombreux moments de partages et les acides gras bretons ! Je remercie Nathalie Riche et Catherine Roisin pour leur disponibilité, leur efficacité redoutable et leur bonne humeur. Je profite de ces pages pour remercier Guillaume Pied-Trotignon pour sa bonne volonté et son implication dans ce projet lors de son stage de Master 1 ; le multi-



espèces, même pas peur ! ;-) Un grand merci également à Céline Melin pour son aide à la paillasse (car oui, j'ai fait de la paillasse, quelqu'un en doutait?!). Merci à Emeline Marais d'avoir pris soin de mes (toutes) petites mains. Merci Thibault Munsch pour toute ton énergie d'animation péri-labo et ton délicieux Kirsch ! Merci également à François et Lucile !

Je remercie les nombreux co-grouillots de l'EA2106 ! Je suis heureuse qu'au delà d'une relation collégiale, des amitiés sont nées pendant ce doctorat. Merci Benji pour tes conseils cinématographiques et la déco du bureau ! Merci Emilien pour ton extrême gentillesse dès mon arrivée dans ce labo et l'aide que tu as toujours apporté comme une évidence. Merci Dimitri pour les nombreuses heures (cela doit faire des jours voire des semaines au bout du compte) de discussion scientifique ou comicstique ! Et merci aussi pour tes cours de Master, me permettant non seulement de faire évoluer mes connaissances scientifiques, mais également mes capacités de décodage. ;-) Merci Florent pour ta bonne humeur et ton humour qui ont su illuminer le bureau des thésards. Je reste infiniment triste que l'aventure ait pris prématurément fin pour toi. Certaines portes de donjon méritent de rester fermées. Merci Tatiana alias Tatoche pour ta spontanéité, les fou-rires et les conseils précieux. Cela n'a pas été facile de te voir partir à Angers alors que nous avions commencé le périple ensemble. Merci Kevin, mon voisin de bureau pour un temps, de m'avoir fait me sentir moins seule dans cet environnement. Merci Valentin ! Cela a été un plaisir d'apprendre à te connaître et de partager sur nos loisirs communs (oui on joue encore avec des petites figurines en plastique à notre âge et c'est cool !). J'ai bon espoir que cette amitié perdure au-delà de la thèse (micro et onde resteront marqués à jamais par votre rencontre fugace) et je te souhaite beaucoup de courage pour la route qu'il te reste à parcourir en espérant que tu ne finiras pas complètement prénylé !

I would like to thank Dr Marta Ines Teto Carqueijeiro (my name isn't so difficult to spell finally) alias Tata for sharing her experience. After all these years rubbing shoulders with you, I guess that I know Portugal better than my home country! My sincere thanks to Dr « Kostas » Koudounas for bringing some greek sun in the lab, your permanent good spirit was like a breeze of fresh air in the daily work routine ! Thank you Pamela for your brazilian smiles and the geeky moments ; I wish you all the best for your future and the upcoming thesis! Also a big thank you for the latino vibes to all the Colombian interns I had the chance to meet !



Ich möchte an dieser Stelle Prof. Dr. (!) Günther Weber danken. Es war schön, einen Landsmann an meiner Seite zu haben, so habe ich mich vielleicht nicht komplett « gallisiert »! Ich danke Dir für Deine unfehlbare Unterstützung während meines Studiums und meiner Doktorarbeit. Deine unermessliche Leidenschaft für die Wissenschaft war und ist hochgradig ansteckend. Vor allem danke ich Dir von Herzen, dass Du mich immer ermutigt hast, an mich und meine Arbeit zu glauben. Es ist stets eine Freude nach Labor und Schweiss, Wein und Spiel mit Dir zu teilen. Ich möchte an dieser Stelle auch Dr. Sandra Krull für ihre wertvolle Hilfe im Labyrinth der akademischen Welt und die vielen schönen Momente danken. Du fehlst!

Selbstverständlich danke ich innig meiner Familie für Ihre Unterstützung während der letzten... 33 Jahre. Ohne Euch, würde heute niemand dieses Manuskript in den Händen halten. Ich danke meinen Eltern dafür, dass sie stets hinter mir standen und mich ermutigten, welch auch mein Weg und welch auch die Hindernisse waren. Ein besonderer Dank geht an meine grosse Schwester Frederike, die mich während dieser Doktorarbeit gleich zweimal zur Tante gemacht hat. Ich hoffe, diese Arbeit ist auch nur halb so gut gelungen wie die Projekte Valentin und Joséphine!

Probablement le plus grand soutien de tous, je tiens à remercier *Coffea arabica* pour la production de l'alcaloïde le plus précieux de tous, la méthylthéobromine, et qui a été un soutien indéfectible durant ces années de thèse et plus particulièrement durant la rédaction de ce manuscrit.

Un grand grand merci également à tous mes amis, d'ici et d'ailleurs. Je suis heureuse et infiniment chanceuse de vous avoir dans ma vie et d'avoir toujours pu compter sur votre soutien ! Je remercie en particulier Catherine et Jean-Christian Brucker (et par extension tout le clan !) de m'avoir accueillie au sein de leur famille comme s'il n'y avait rien de plus naturel. Bon courage à vous, Antoine et Martin, pour la suite de votre parcours d'études ! Je tiens également à remercier Alain Verger et Martine Thiersault, pour le soutien indéfectible dès ma licence. Un grand grand merci à mes amis parisiens pour m'avoir permis de rester en lien avec la civilisation citadine durant ces années. Merci à mon «semi-Ösi » favori Viktor, Delphine et Sigrid pour les weekends de détente et tout votre soutien qui m'ont toujours permis de revenir gonflée à bloc ! Merci Julien, Sara (et Roxane !) pour les nombreuses discussions que ce soit sur la passion scientifique ou celles plus tentaculaires ! Sara, je savais la frontière entre



paillasse et maladies psy très mince, tu as réussi à complètement l'effacer. :-D Merci Herr Inspektor Florent pour les bons moments en concert ou ailleurs ! Merci également à Jörn et Marie ! Un clin d'oeil particulier à Torti. ;-)

Parait-il qu'il faut garder le meilleur pour la fin : Hélène, comment exprimer ma gratitude pour ton soutien infaillible durant ces années ? J'ai certes bataillé pour tout cela, mais une bataille ne se gagne jamais seul (à part si tu as la Sainte Grenade éventuellement). De la danse du tube pollinique aux réseaux de co-expression ça en fait du chemin. Et voici l'épisode final de « Plus beau Grandmont » ! Tu as été mon Sam tout au long de cette aventure. Et je ne fais pas référence à ton appétit (ou tes pieds), car même si je n'ai pas eu à combattre d'araignées géantes ou de Nazguhls en tongs ; comme lui avec son acolyte, tu as toujours été à mes côtés et surtout (et c'est là que cela devient réellement fantastique) réussi à me supporter. Quel plaisir de partager le quotidien avec toi et quel soulagement d'avoir une amie comme toi à mes côtés lors de cette expérience qui peut par moment être éprouvante. Je ne pourrais te remercier assez pour ton soutien, ta patience et surtout d'avoir évité que la thèse fasse de moi (complètement) un Gollum.



## Résumé

Avec l'essor de technologies à haut débit capables de fournir une vue à grande échelle du transcriptome, une grande quantité de données de type microarray et RNAseq a été, et est toujours générée. Ce travail est axé sur la réutilisation de données publiquement disponibles comme base pour la construction de réseaux de co-expression de gènes, exploitables pour l'élucidation de voies cellulaires de type voie de biosynthèse ou de signalisation.

Nous nous sommes intéressés à différents paramètres pouvant fortement influencer la structure du réseau de co-expression final et par conséquent sa capacité à révéler de réelles associations biologiques. Dans la conduite de ce travail dont le but final est de fournir une méthodologie pour l'élucidation de voies cellulaires, trois enjeux majeurs ont été identifiés : (i) choix d'une distance appropriée pour évaluer la similarité de profils d'expression entre gènes, (ii) fixer un nombre d'échantillons minimal à inclure dans la matrice d'expression pour construire des réseaux robustes et enfin (iii) comparer des réseaux ciblés de type *Pathway Level Co-expression* (PLC) construits avec les gènes codant les acteurs de la voie *Multi Step Phosphorelaïs* (MSP) comme guides, entre différentes espèces.

**Mots-clés:** Réseau de co-expression, transcriptomique, voies métaboliques, voies de signalisation, larges données



## Abstract

With the rise of high throughput technologies able to provide a large-scale view of transcriptomes, a high amount of microarray and RNAseq data has been and is still produced. This work focuses on publicly available data reuse to construct gene co-expression networks usable to elucidate cellular pathways such as metabolic or signalling pathways.

We looked after different parameters susceptible to strongly influence the structure of the final co-expression network and consequently its ability to reveal actual biological associations. During this work, which final aim was to provide a methodology for the elucidation of cellular pathways, three main issues have been identified: (i) the choice of an appropriated distance to evaluate similarity between gene expression profiles, (ii) set a minimal number of samples to be included in the expression matrix in order to construct robust co-expression networks, and finally (iii) compare targeted co-expression networks built with the *Pathway Level Co-expression* (PLC) approach and using guide genes coding actors of the *Multi Step Phosphorelay* (MSP) among different species.

**Keywords :** **Co-expression network, transcriptomics, metabolic pathways, signaling pathways, large-scale data**



# Table des matières

Remerciements.....	3
Résumé.....	9
Abstract.....	10
Table des matières.....	11
Liste des tableaux.....	14
Liste des figures.....	15
Liste des annexes.....	17
Liste des abréviations.....	18
Introduction.....	19
Que-est-ce qu'un réseau ?.....	20
Réseaux biologiques.....	27
1. Les réseaux de co-expression.....	29
2. Construction des réseaux.....	34
2.1. Données d'expression issues de technologies à haut débit.....	34
2.1.1. Microarray.....	34
2.1.2. RNA-seq.....	36
2.1.3. Microarray <i>versus</i> RNA-seq : avantages et inconvénients.....	37
2.2. Établir la co-expression.....	40
2.2.1. Coefficients de corrélation.....	41
2.2.2. Classement de coefficients de corrélation.....	44
2.2.3. Algorithmes complexes et co-expression de gènes.....	45
2.3. Caractéristiques du jeu de données initial.....	49
2.3.1. Technologie utilisée pour la génération des données.....	49
2.3.2. Pré-traitement des données.....	49
2.3.3. Normalisation des données.....	51
2.3.3.1. Normalisation des données Microarray.....	51
2.3.3.2. Normalisation des données RNA-seq.....	52
2.3.4. Taille du jeu de données.....	55
2.3.5. Autres paramètres.....	56
3. Validation et analyse des réseaux.....	58
3.1. Validation.....	58



3.1.1. Données disponibles.....	59
3.1.2. Validation des relations mises en évidence dans le réseau.....	61
3.2. Analyse.....	62
3.2.1. Réseaux globaux et réseaux ciblés.....	62
3.2.2. Du réseau à l'information biologique : détection de communautés.....	64
Contexte, méthodologie générale et organisation de la thèse.....	67
1. Mise en place initiale.....	69
2. Méthodologie générale.....	70
2.1. Identification et préparation des données d'expression.....	70
2.2. Ressources informatiques.....	72
2.3. Construction des réseaux de co-expression.....	73
2.3.1. Choix de la matrice initiale.....	74
2.3.2. Choix de la distance.....	74
2.3.3. Choix du seuil.....	75
2.3.4. Le set de validation.....	75
2.3.5. La technique de confrontation.....	76
3. Organisation de la partie Résultats.....	77
Partie I.....	79
Partie II.....	98
Partie III.....	119
Discussion générale.....	154
Conclusion.....	175
Bibliographie.....	178
Résumé.....	204



## Liste des tableaux

Table I : Bases de données permettant des études de co-expression à partir de gènes candidats

Table II: Comparaison entre RNA-seq et microarray

Table III : Organisation des données transcriptomiques déposées dans les bases de données publiques.

Table IV : Étapes pour la récupération des données publiées à partir de bases publiques.

Table V : résumé des consommations informatiques sur les ressources Artemis et Neptune.



# Liste des figures

Figure 1 : Boom des données -omiques.

Figure 2 : Le problème des ponts de Königsberg.

Figure 3 : Différents types de réseaux.

Figure 4 : Propriétés et caractéristiques des réseaux.

Figure 5 : Un exemple de réseaux : l'unité de recherche Biomolécules et Biotechnologies Végétales (BBV EA2106, Université de Tours)

Figure 6 : Co-expression de gènes.

Figure 7 : Principe de "culpabilité par association" (Guilt By Association, GBA).

Figure 8 : Représentation schématique de l'approche microarray (puce à ADN).

Figure 9 : Représentation schématique du processus de RNAseq.

Figure 10 : Portails des bases de données SRA (NCBI) et ArrayExpress (EMBL-EBI).

Figure 11 : Exemples de différents types de corrélations entre variables.

Figure 12 : Principales étapes du workflow du paquet WGCNA

Figure 13 : Contrôle qualité de données microarray avec le paquet R 'arrayQualityMetrics'.

Figure 14 : Extrait de fichier de sortie RNA-seq au format FastQ.

Figure 15: Contrôle qualité de données RNA-seq avec FASTQC.

Figure 16 : Spécificité et sensibilité d'un test statistique.

Figure 17 : Annotation "Gene Ontology" (GO).

Figure 18 : Méthodologie pour la construction de réseaux ciblés de type Pathway Level Co-expression (PLC).

Figure 19 : Problématiques pour lesquelles une analyse de co-expression pourrait être envisagée.

Figure 20 : Exemple de code R pour la création d'un réseau de co-expression.

Figure 21 : Protocole de construction d'un réseau de co-expression.



Figure 22 : Confrontation d'un réseau de co-expression à une annotation de référence.

Figure 23 : Impact de la distance sur la performance de réseaux de co-expression.

Figure 24 : Impact de la taille de matrice sur la performance de réseaux de coexpression.

Figure 25 : Comparaison multi-espèce de réseaux de co-expression.

Figure 26 : Alignement de PLC microarrays et RNA-seq ciblant le métabolisme des phénylpropanoïdes chez *Arabidopsis thaliana*.

Figure 27 : Application des réseaux de co-expression à des différentes thématiques.

Figure 28 : Evaluation de la performance des réseaux en fonction des distances et des jeux de données initiaux.

Figure 29 : Intégration de données d'interactions protéine-protéines (PPI) dans la construction de données biologiques.



## Liste des annexes

Annexe 1 : Extrait de la table d'entrée pour la construction du réseau BBV

Annexe 2 : Publication Navarro et al., 2017

Annexe 3 : Publication Dugé de Bernonville et al., 2017

Annexe 4 : Publication Daudu et al., 2017



## Liste des abréviations

**ACP** : Analyse en Composantes Principales

**AIM** : Alcaloïdes Indoles Monoterpéniques

**AUROC** : Area Under Receiver Operating Characteristic curve

**CC** : Coefficient de corrélation

**CK** : Cytokinine

**EMBL-EBI** : European Molecular Biology Laboratory – European Bioinformatics Institute

**GO** : Gene Ontology

**HRR** : Highest Reciprocal Rank

**JA** : Jasmonic Acid

**MI** : Mutual Information

**MR** : Mutual Rank

**MSP** : Multi Step Phosphorelay

**NCBI** : National Center for Biotechnology Information

**PC** : Partial Correlation

**PCC** : Pearson Correlation Coefficient

**PLC** : Pathway Level Co-expression

**PP** : Phenylpropanoïdes

**PPI** : Protein-Protein Interaction

**SCC** : Spearman Correlation Coefficient

**WGCNA** : Weighted correlation network analysis



# Introduction



## Que-est-ce qu'un réseau ?

« Le diable est dans le détail ! » proclamait Friedrich Nietzsche dans la seconde moitié du XIXe siècle. Ce postulat illustre parfaitement les approches des sciences naturelles modernes qui, avec l'apparition de nouvelles technologies et de techniques de plus en plus précises, essaient d'expliquer les phénomènes régissant leur domaine en les réduisant à leur dimension la plus petite possible. Ainsi l'univers serait décrit par les particules élémentaires et le vivant par les molécules qui les composent, alors qu'une maladie ou un comportement serait résumé par seulement un ou quelques facteurs.

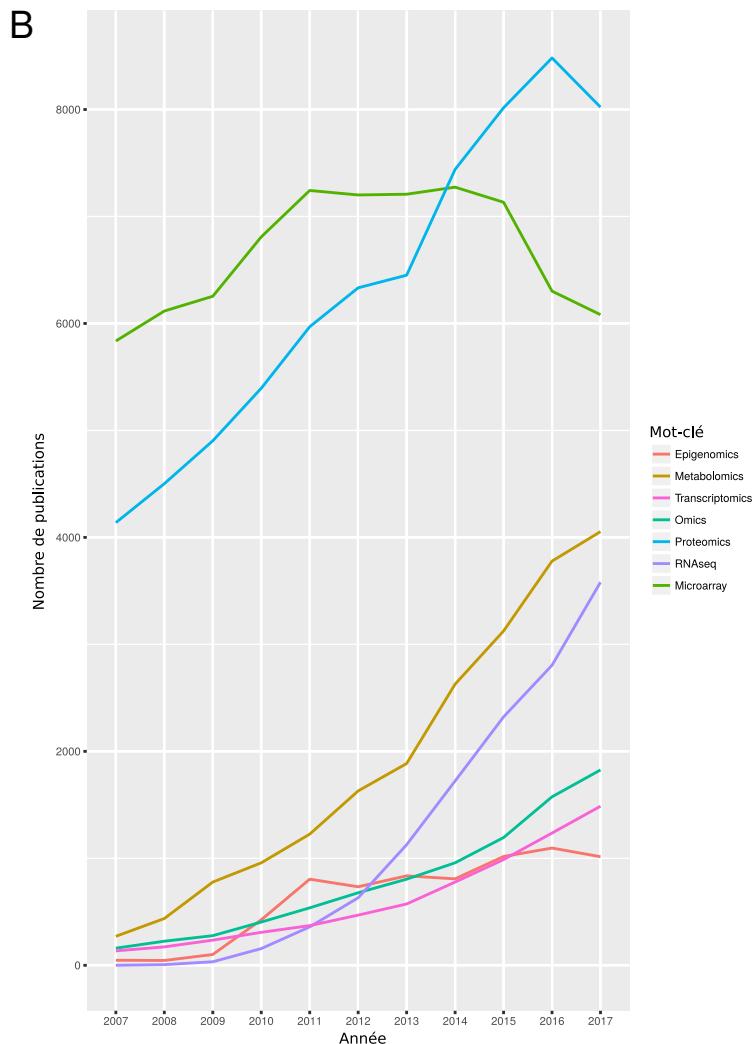
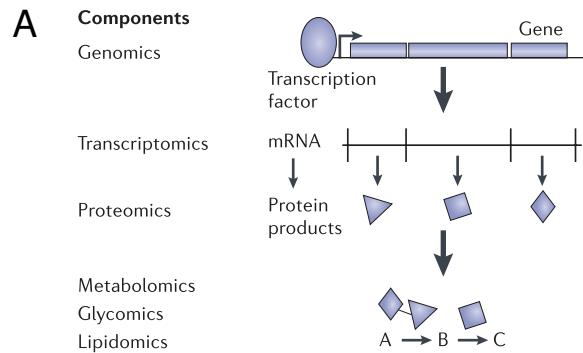
Lors de la seconde moitié du XXème siècle, l'essor de la biologie moléculaire a entraîné un réductionnisme méthodologique dans le domaine des sciences du vivant. Pour comprendre un système biologique, l'approche la plus efficace serait de se situer au plus petit niveau possible (Andersen 2017) ; les études expérimentales se limitent alors à élucider les causes moléculaires et biochimiques d'un phénomène. Ce type d'approche réductionniste repose lui-même sur le dogme du réductionnisme ontologique qui présume qu'un système biologique, comme un organisme par exemple, est constitué exclusivement de molécules et de leurs interactions. Leur étude et leur élucidation suffiraient alors à en comprendre le fonctionnement. Un exemple concret est ce qui constitue probablement un des projets scientifiques les plus ambitieux de la fin du XXème siècle : le « projet génome humain » qui a vu le jour dans les années 1980 avec pour but le déchiffrage complet du génome humain à l'aide de nouvelles technologies toujours plus performantes. En 2003, ce dernier a été séquencé dans son intégralité après une course intense entre le Consortium International de Séquençage du Génome humain et l'entreprise privée Celera Genomics. (Consortium IGHS, 2001; Venter et al., 2001). L'aboutissement de ce projet et l'obtention du « livre de la vie » constitue sans doute une prouesse scientifique et une source d'information inestimable. Mais rapidement, la communauté scientifique se heurte aux limitations de l'enchaînement du code d'environ trois milliards de lettres constituant notre ADN comme base à la résolution de problèmes scientifiques. Il apparaît en effet qu'il s'agit d'une information insuffisante pour faire le lien entre génotype et phénotype, et par conséquent pour expliquer le développement et la physiologie d'un organisme par exemple. Ce lien est étudié depuis longtemps par



l’intermédiaire de la détermination de locus de traits quantitatifs (*Quantitative Trait Loci* ou QTL) par exemple, mais l’émergence des données de séquençage génomiques à haut débit constitue une source de renseignements supplémentaires précieuse, notamment en changeant la perspective, se plaçant non plus à l’échelle de l’individu mais à celle de la population. Ce principe est mis en pratique avec les études d’association pangénomiques ou *Genome Wide Association study* (GWAS) qui visent le génotypage d’une population portant un caractère commun (une pathologie en biologie humaine ou une amélioration à intérêt agronomique chez les végétaux par exemple) dans le but de relier des variants génétiques à un ou des traits phénotypiques (Hirschhorn & Daly, 2005). Un projet monumental visant à identifier la majorité des variants génétiques occurant avec des fréquences d’au moins 1 % chez l’homme, le « projet 1000 génomes », a analysé 2 504 génomes provenant de 26 populations entre 2008 et 2015 et ainsi mis en évidence 88 millions de variants génétiques (1000 Genomes Project Consortium, 2015)<sup>1</sup>. Malgré le nombre d’échantillons faramineux produit lors de ces génotypages de population, il reste difficile de faire le lien entre variabilité du génome et phénotype. La compréhension de la nécessité de considérer un système dans son intégrité a rapidement émergé, remplaçant peu à peu la discipline de la génomique structurale par la génomique fonctionnelle. Cette dernière étudie le génome dans sa dynamique en considérant les gènes par leur expression, la régulation de cette dernière ainsi que les interactions de leurs produits (Keller, 2005).

Bien que la caractérisation des mécanismes à l’échelle moléculaire constitue une partie fondamentale dans le processus de recherche en biologie, nous pouvons fortement supposer qu’elle ne suffit pas à décrire la complexité d’un système vivant. En effet, ce dernier repose sur d’innombrables interactions entre processus physiques, chimiques et biologiques à différentes échelles ainsi que sur l’interaction de l’organisme avec son environnement.

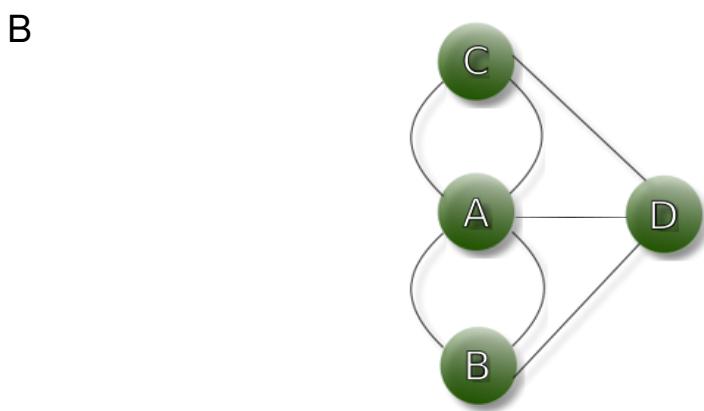
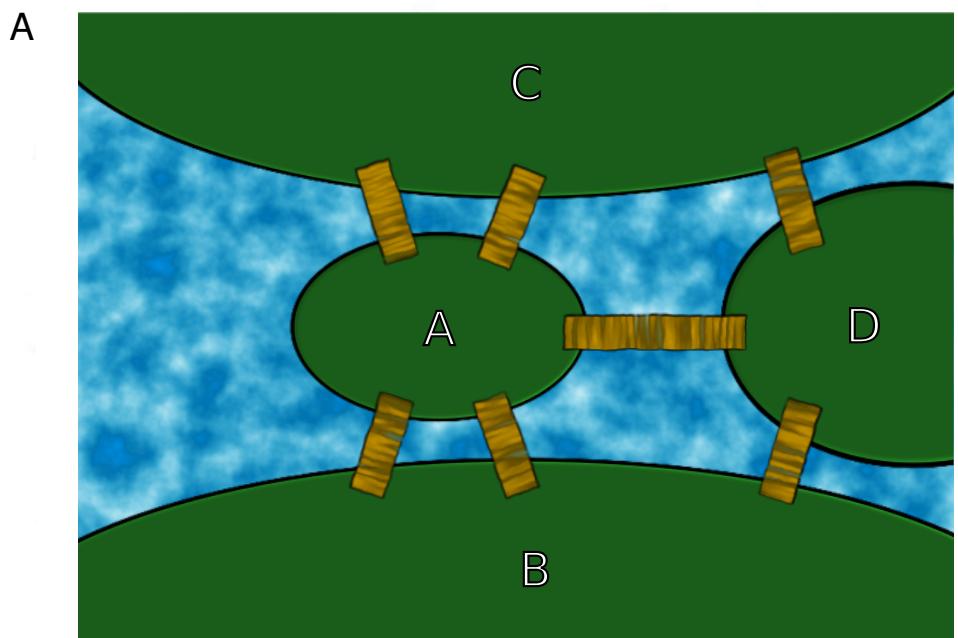
1 Les données du projet 1000 génomes est toujours étendu à l’heure actuelle. Par ailleurs, des projets encore plus ambitieux en terme de taille de population ont vu le jour comme les projets des 100 000 génomes britanniques et asiatiques (<http://www.phgfoundation.org/blog/the-global-genomics-race-asia-the-uk-and-beyond>).



**Figure 1: Boom des données -omiques.** A, présentation des différents niveaux -omiques. B, Nombre d'articles référencés sur NCBI comportant les termes « Microarray », « RNAseq » ou différentes approches «-omiques », publiés entre 2007 et 2017. On peut constater un abandon des technologies de type microarray avec le développement et la généralisation du RNAseq.

Cette vision réductionniste se retrouve *ipso facto* au niveau moléculaire lui-même ; avec des approches se focalisant sur un ou quelques gènes pour résumer un trait de caractère ou un processus biologique, décomplexifiant artificiellement ce dernier. Par ailleurs, l'investissement en temps ainsi que le coût financier représentent souvent des facteurs limitants et contraignent à contingenter les expériences entreprises à un faible nombre de gènes candidats. L'avènement et le développement de technologies à haut débit, de type séquençage et microarray, permettent d'étudier directement un nombre important de gènes voire des génomes entiers. Cet essor ne se limite pas seulement au support initial de l'information, le génome, puisque des technologies d'analyse à grande échelle sont rapidement apparues pour différents niveaux de mécanismes et composés cellulaires : l'ARN messager (transcriptomique), les protéines (protéomique), les métabolites (métabolomique), les lipides (lipodomique), les carbohydrates (glycomique) *etc.* (Joyce & Palsson, 2006). Ces approches « -omiques », basées sur l'analyse d'un grand nombre de molécules, apportent une vue plus globale des processus biologiques et peuvent s'envisager de manière complémentaire dans le cadre d'études « multi-omiques », afin d'obtenir un aperçu plus global d'un processus et ainsi une approche plus holistique (Gomez-Cabrero et al., 2014; Palsson, 2015). Les développements techniques considérables permettent l'étude rapide et quantitative de ces molécules faisant de ces études « -omiques » rapidement des outils incontournables en recherche biologique (**Figure 1**). Elles ont ainsi entraîné la génération d'une quantité de données considérable au cours des deux dernières décennies, nécessitant le développement d'approches *in silico*, dédiées au traitement et à la visualisation de ces dernières. En effet, considérer un grand nombre de candidats d'intérêt à la fois et *a fortiori* un « -ome » dans sa totalité, d'un ou plusieurs organismes, nécessite des outils spécifiques permettant de transformer ces données de mesures brutes en information biologique valorisable conduisant *in fine* à la réponse à une problématique concrète.

Se pose donc la question de la visualisation et de l'interprétation de ces données impliquant potentiellement un grand nombre d'acteurs, afin d'illustrer au mieux les interactions complexes se produisant dans le cadre d'un processus biologique. Un système peut être représenté graphiquement par un réseau décrivant l'inter-connectivité entre les différentes composantes de ce dernier, non seulement par des relations binaires, mais dans leur ensemble. Le mot « réseau » provient étymologiquement du mot latin « *retiolus* », diminutif de « *retis* »

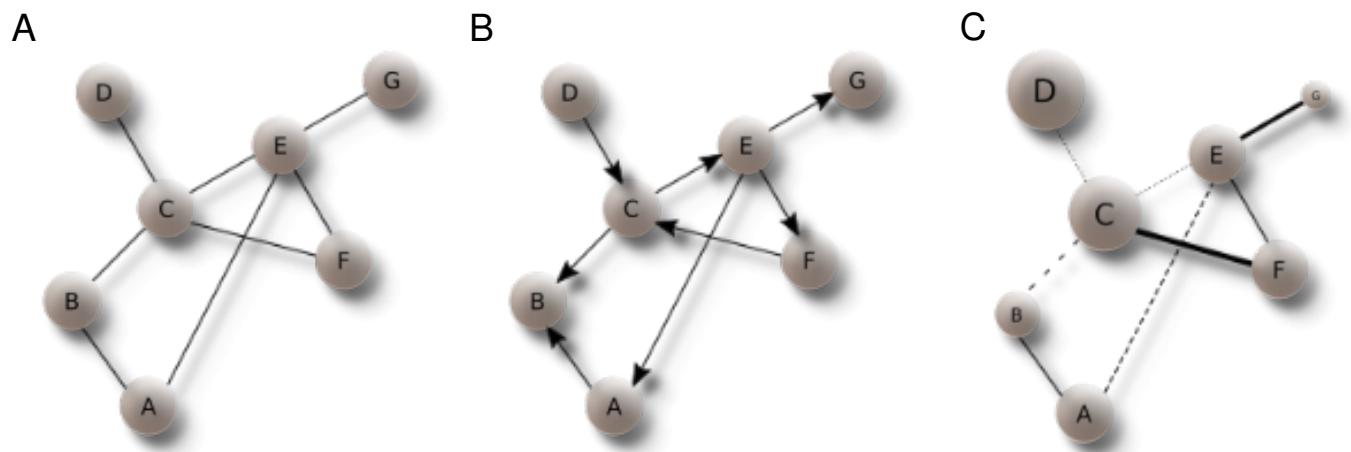


**Figure 2: Le problème des ponts de Königsberg.** A. La configuration de ponts de Königsberg : Les parties émergées sont reliées par un pont entre elles et aux rives du fleuve par six autres ponts. B. Représentation de la situation ci-dessus sous forme de graphe : les différentes parties de la ville forment les noeuds d'un réseau connectés par les arêtes formés par les ponts. Ce réseau comporte donc 4 noeuds reliés par 7 arêtes. Ce graphe ne présente ni chaîne (chaîne passant une et une seule fois par toutes les arêtes du graphe) ni cycle eulerien (chaîne eulerienne dont les extrémités sont confondues). En effet, un graphe connexe possède une chaîne eulérienne si et seulement si ses sommets sont tous de degré pair sauf au plus deux et possède un cycle eulerien si et seulement si tous ses sommets sont de degré pair. Dans l'exemple de cette démonstration, de chaque sommet part un nombre impair d'arêtes.

qui signifie « filet ». Fondamentalement, un réseau décrit un ensemble d’éléments connectés les uns aux autres. La métaphore du filet prend son sens au-delà de la représentation graphique en forme de toile, puisque ce réseau va pouvoir être utilisé pour « pêcher » des informations sur et par l’interaction des différents éléments. Les réseaux trouvent leur place dans tous les domaines scientifiques, qu’il s’agisse de sciences humaines et sociales, de sciences naturelles ou d’économie, avec par exemple les réseaux neuronaux, les réseaux sociaux ou les réseaux de distribution. Par ailleurs, ils sont nativement présents dans la nature et par extension dans tous les domaines de la vie sous forme de « réseaux réels » comme le système circulatoire sanguin ou un système de transports publics par exemple. Ils présentent une structure commune et suivent des lois mathématiques définies permettant leur analyse statistique (Newman, 2006).

L’étude des réseaux est basée sur la théorie des graphes. Les fondements de cette discipline mathématique et informatique furent posés par le mathématicien suisse Leonhard Euler en résolvant le problème des sept ponts de Königsberg (Euler, 1726). Ce problème mathématique est basé sur une situation réelle dans la ville de Königsberg du XVIIIème siècle, située non loin de la ville de résidence d’Euler, Saint Petersbourg. Cette ville est en partie construite sur l’île de Kneiphof et une surface émergée située entre deux branches du fleuve Pregel (**Figure 2A**). Ces dernières sont reliées entre elles par un pont ainsi qu’aux rives du fleuve par six autres ponts. Résoudre ce problème revient à trouver une route à travers la ville permettant de revenir à son point de départ en ne passant qu’une seule fois par chacun des sept ponts. Euler proposa une solution simple et élégante en démontrant qu’un tel chemin n’existe pas<sup>2</sup>. Ce n’est pas tant la preuve de la résolution du problème qui devint célèbre que la méthode adoptée par Euler pour y parvenir. Son idée a été de visualiser les ponts de Königsberg sous la forme d’un graphe sur lequel chacune des parties émergées est représentée par un sommet (A, B, C et D) et chaque pont par une arête (**Figure 2B**). La preuve apportée par Euler qu’un chemin passant une seule fois par chaque pont n’est pas possible est basée sur une simple observation : les nœuds présentant un nombre impair d’arêtes doivent être soit le sommet de

2 Une solution plus pragmatique au problème des sept ponts de Königsberg a été proposée : en effet, le chemin d’Euler devient possible en construisant un huitième pont ce qui fut le cas en 1875. D’ailleurs, suite à la destruction de deux ponts pendant la seconde guerre mondiale, le chemin d’Euler est aujourd’hui possible dans le Kaliningrad actuel, mais toujours pas le cycle eulérien.



**Figure 3: Différents types de réseaux.** Réseau non orienté (A), Réseau orienté (B), Réseau non orienté et pondéré (C). Chaque réseau possède 7 nœuds reliés par 7 arrêtes.

départ soit le point d'arrivée. Un chemin continu qui passe par tous les ponts ne peut avoir qu'un point de départ et un seul point d'arrivée. Donc un tel chemin ne peut pas exister sur un graphe qui a plus de deux points avec un nombre impair d'arêtes. Le graphe en comptant quatre, le chemin est impossible. Ce qui est intéressant dans la démonstration d'Euler est le fait que l'existence de ce chemin ne dépend pas de l'ingéniosité mise en œuvre pour le trouver, mais est elle une propriété intrinsèque du graphe.

L'objectif de l'étude de réseaux est justement de tirer profit de ces propriétés sous-jacentes à leur construction et leur topologie afin d'expliquer des systèmes complexes représentés sous forme de graphe (Barabási & Albert, 1999). *De facto* la représentation graphique par elle-même est un outil précieux pour l'analyse visuelle de données comme le démontre l'approche d'Euler.

Fondamentalement, un réseau est donc la représentation d'un ensemble d'éléments symbolisés par des nœuds (ou sommets) reliés entre eux par des liens (ou arêtes) illustrant la relation entre ces derniers (**Figure 3A**). Ce réseau va présenter une organisation et donc une topologie qui lui sont propres et conditionnées par les données de départ. Il existe différents types de réseaux simples : un réseau peut-être orienté (**Figure 3B**) dans quel cas les liens connectant deux nœuds comportent une flèche qui associe une direction à l'interaction (A,C). Ces nœuds et liens peuvent également être pondérés et ainsi comporter une information de la force de la relation (**Figure 3C**). En fonction des liens qu'il partage avec d'autres membres du réseau, un nœud va former des modules ou des communautés avec d'autres nœuds auxquels il est fortement connecté. Cette caractéristique va permettre de décrire les relations entre les différents nœuds. Un réseau étant un graphe, il possède des propriétés liées à sa structure et descriptibles par des approches statistiques comme illustré par la **Figure 4**.

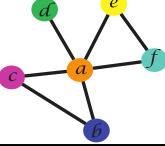
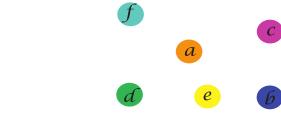
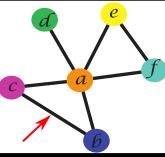
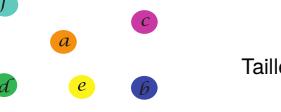
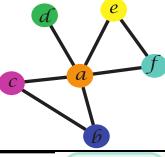
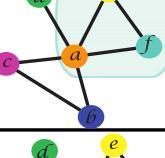
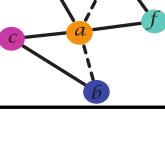
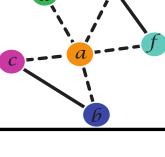
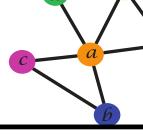
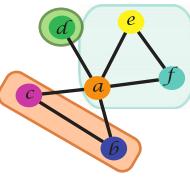
<b>Réseau:</b> Noeuds connectés par des arêtes																																											
<b>Sommet (Vertex) ou noeud (node):</b> Unité fondamentale du réseau																																											
<b>Arête (Edge) ou lien (link):</b> Segment qui relie deux noeuds																																											
<b>Taille:</b> Nombre de noeuds																																											
<b>Voisinage:</b> Nœuds connectés à un nœud N du réseau	 si $N = c$ voisinage = a et b																																										
<b>Module:</b> Nœuds densément connectés entre-eux (attribution statistique)																																											
<b>Chemin géodésique:</b> Chemin le plus court dans le réseau d'un nœud à un autre (N.B. Plusieurs chemins sont souvent possibles)	 chemin géodésique entre b et e = 2																																										
<b>Diamètre:</b> Longueur (en nombre de liens) du chemin géodésique le plus long entre deux nœuds	diamètre = 2																																										
<b>Densité:</b> Ratio entre le nombre liens et le nombre possible de liens dans un réseau avec Z nœuds	nombre possible de liens = 15 nombre de liens = 7 densité = 7/15 = 0.46																																										
<b>Degré d'un noeud:</b> Nombre de liens qu'un nœud possède avec d'autres nœuds	 degree $a = 5$																																										
<b>Degré moyen:</b> Moyenne du nombre de liens que possèdent les nœuds du réseau	degré moyen = 2.3																																										
<b>Transitivité ou coefficient de clustering:</b> Probabilité que les nœuds adjacents à un nœud N soient connectés	 <table border="1"><tr><td></td><td colspan="5">neighbors</td></tr><tr><td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td></tr><tr><td>b</td><td>-</td><td>✓</td><td>✗</td><td>✗</td><td>✗</td></tr><tr><td>c</td><td>✓</td><td>-</td><td>✗</td><td>✗</td><td>✗</td></tr><tr><td>d</td><td>✗</td><td>✗</td><td>-</td><td>✗</td><td>✗</td></tr><tr><td>e</td><td>✗</td><td>✗</td><td>✗</td><td>-</td><td>✓</td></tr><tr><td>f</td><td>✗</td><td>✗</td><td>✗</td><td>✗</td><td>-</td></tr></table> 2/10 connections - clustering coefficient of node a = 0.2		neighbors					a	b	c	d	e	f	b	-	✓	✗	✗	✗	c	✓	-	✗	✗	✗	d	✗	✗	-	✗	✗	e	✗	✗	✗	-	✓	f	✗	✗	✗	✗	-
	neighbors																																										
a	b	c	d	e	f																																						
b	-	✓	✗	✗	✗																																						
c	✓	-	✗	✗	✗																																						
d	✗	✗	-	✗	✗																																						
e	✗	✗	✗	-	✓																																						
f	✗	✗	✗	✗	-																																						
<b>Modularité:</b> Mesure la qualité d'un partitionnement d'un réseau en communautés. Cette qualité est élevée, si le nombre d'arêtes intra-communautés est fort, alors que le nombre d'arêtes inter-communauté est faible (comprise entre 0 et 1)	 modularité = 0.07																																										

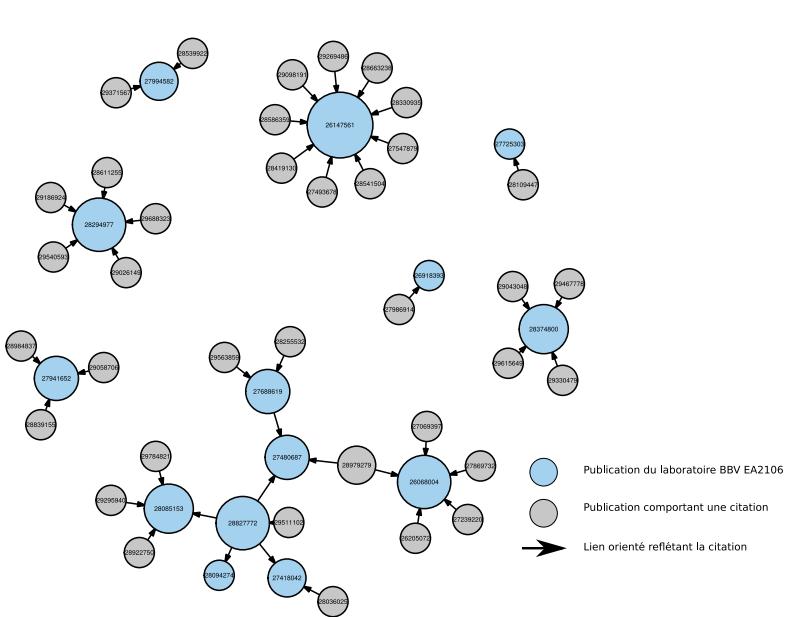
Figure 4: Propriétés et caractéristiques des réseaux. Adapté d'après Toubiana et al 2013.

**Un exemple de réseaux : l'unité de recherche Biomolécules et Biotechnologies Végétales**  
**(BBV EA2106, Université de Tours)**

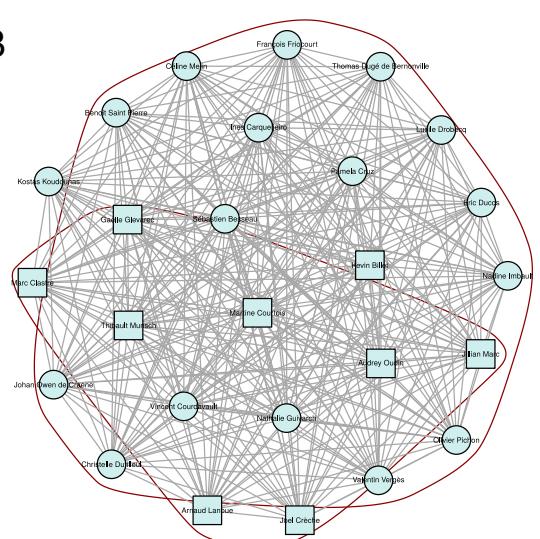
Un laboratoire de recherche est une structure sociale axée autour d'une ou plusieurs thématiques scientifiques, au sein de laquelle interagissent continuellement ses différents membres par des activités distinctes et communes (recherche, enseignement, administration etc.). Cette structure représente un système ouvert puisque au-delà des relations intra-membres, elle interagit continuellement avec le monde extérieur (publications scientifiques, congrès, demande de financements, accueil de chercheurs invités, d'étudiants etc.). Comme évoqué précédemment, un réseau décrivant les différentes interactions entre les éléments d'un système, il est possible d'utiliser cette représentation graphique afin de mettre en lumière des schémas intrinsèques à l'organisation du laboratoire.

Le moyen de communication le plus répandu et le plus pérenne au sein de la communauté scientifique est la publication dans les revues spécialisées. Une manière de représenter le lien qui existe entre une unité de recherche et ses pairs est le réseau de citation (Hummon & Dereian, 1989). Il s'agit d'un réseau social contenant des publications sources reliées à d'autres publications plus récentes qui ont cité les précédentes. La **Figure 5A** montre une représentation (non exhaustive) des publications du laboratoire BBV entre 2016 et 2018 et des publications subséquentes qui les ont citées. Il s'agit d'un exemple de réseau orienté : un auteur B qui cite un auteur A est représenté par un lien dirigé  $B \rightarrow A$ , les flèches permettant un suivi chronologique. Étant donné que logiquement un papier ne peut-être cité avant sa publication, ce type de réseau est également acyclique. En plus de fournir une vision globale des liens de collaborations intellectuelles d'une équipe ou au sein d'une thématique de recherche, ces réseaux peuvent être analysés à l'aide de modèles statistiques, afin d'en faire émerger des tendances scientifiques ou d'évaluer l'influence d'une revue spécialisée par exemple (Nerur et al., 2005; Shibata et al., 2011). Ainsi, Wagner et al. ont étudié les réseaux formés par les lauréats du prix Nobel de physiologie ou médecine entre 1969 et 2011 et les ont comparés à un groupe de scientifiques non lauréats par rapport à différents critères, comme l'année de la première publication, le domaine de recherche, le h-index etc. Par

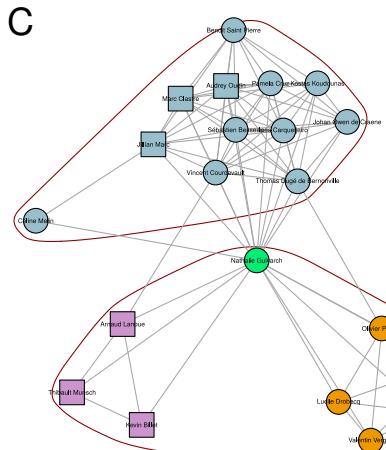
A



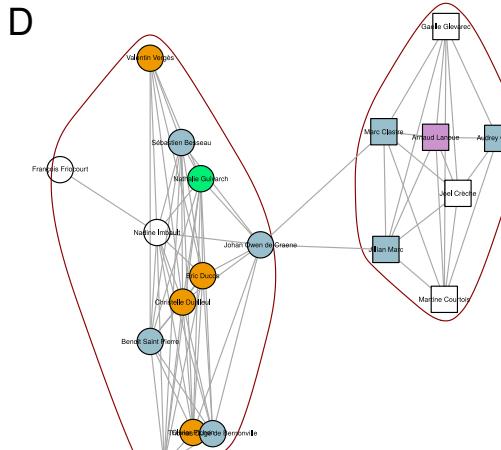
B



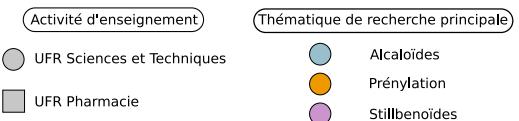
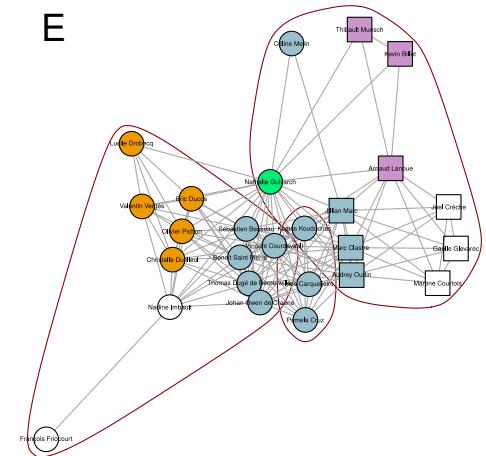
C



D



E



**Figure 5: Un exemple de réseaux : l'unité de recherche Biomolécules et Biotechnologies Végétales (BBV EA2106, Université de Tours)**

A. Interactions entre le laboratoire BBV et la communauté scientifique. Graphique de citation représentant les publications du laboratoire Biomolécules et Biotechnologies Végétales EA2106 référencées sur NCBI entre 2016 et 2018 (représentés en bleu). Les publications sont ici identifiées par leur numéro PMID.

Interactions entre membres du laboratoire.

B. Réseau formé par la globalité des interactions au sein de l'unité de recherche. Ce réseau possède 26 nœuds reliés par 326 arêtes (degré moyen=25, modularité=0, transitivité=1). C et D. Réseaux construits à partir des mêmes données mais en introduisant respectivement un niveau d'information sur les activités de recherche principales et les activités d'enseignements des membres (caractéristiques pour respectivement les réseaux C et D : nombre de nœuds = 21 et 19, nombre de liens : 86 et 70, degré moyen= 8,20 et 7,63, transitivité=0,76 et 0,83, modularité=0,28 et 0,37). E. Réseau comportant conjointement les deux informations précédentes. Il comporte 26 nœuds reliés par 135 arêtes (degré moyen=10,38, transitivité=0,72, modularité 0,20).

l'étude de la structure de ces réseaux, cette étude met en lumière les différences de « capital social » entre les deux groupes (Wagner et al., 2015).

D'une manière semblable, les relations intra-équipe du laboratoire peuvent être représentées sous forme de graphe. Lorsque l'on représente la totalité des interactions non hiérarchisées entre différents membres du laboratoire, on constate, comme on peut l'attendre d'un environnement de travail, que tout le monde semble interagir avec tous ses collègues à un moment où un autre (**Figure 5B**). Au-delà du fait, qu'un laboratoire est un haut lieu d'interaction sociale entre les différentes personnes qui y travaillent, ce graphe ne peut être transformé en information tangible ni par un examen visuel, ni par une étude statistique.

Quel serait l'effet sur la structure du réseau si nous intégrons des niveaux d'information supplémentaires ? De manière générale, dans une unité de recherche universitaire, les deux activités principales sont bien sûr la recherche scientifique, souvent divisible en différentes sous-thématiques, mais aussi l'enseignement (dans notre cas, séparé entre l'UFR Sciences et Techniques et l'UFR de Sciences Pharmaceutiques de l'Université de Tours). Ces caractéristiques spécifiques des liens et des nœuds sont intégrables dans le réseau final. La thématique principale de chaque personne ainsi que son appartenance à un UFR ou l'autre ont été insérées dans le graphe pour chaque nœud sous la forme d'une étiquette respectivement de couleur et de forme. En effet, certains membres du laboratoire, représentés par les nœuds, vont avoir des interactions spécifiques avec d'autres par l'intermédiaire des activités de recherche et/ou d'enseignement (par exemple intervention dans une même UE d'enseignement ou participation dans un même projet de recherche), symbolisées par les arêtes. Ces interactions peuvent être représentées sous la forme d'un tableau indiquant si une relation spécifique existe entre les membres pris deux-à-deux (**Annexe 1**). Les **Figures 5C** et **5D** montrent les réseaux construits en se basant sur respectivement la caractéristique enseignement ou thématique de recherche principale. Contrairement au réseau global, on peut constater que les associations fortes entre certains nœuds créent des regroupements de ces derniers sous forme de modules ou de communautés, faisant apparaître des tendances dans la topologie du réseau. On voit que dans les deux cas, respectivement les deux groupes d'enseignement « Sciences » et « Pharmacie » et les thématiques de recherche ségrégent clairement.



Le dernier réseau (**Figure 5E**) comporte les deux informations à la fois, la connexion entre deux membres du laboratoire a également été indexée en fonction du partage soit d'une activité de recherche soit d'enseignement ou bien des deux. On peut constater que les séparations de modules correspondant à nos deux différentes caractéristiques sont ici moins nettes mais toujours discernables par les différents niveaux de lecture du réseau (*i.e.* modules et code de couleur et de forme).

Retirer ou ajouter des nœuds et/ou des liens, change fondamentalement la structure du réseau. Comme dans la démonstration d'Euler, où la suppression ou l'ajout de ponts pouvait rendre possible l'existence d'un chemin unique, cet exemple nous montre qu'en fonction de la quantité et de la nature des données servant de base à la construction du réseau, des informations différentes peuvent en être interprétées. Cette notion prend tout son sens lorsqu'on travaille sur de larges données biologiques, ces dernières fournissant un nombre quasi infini d'indications potentielles. La question essentielle qui conditionne toute la méthodologie subséquente et qui surgit, est donc : au sens large, quel type, quelle quantité de données et quelle méthode allons-nous intégrer pour établir notre réseau afin de répondre *in fine* à une question de recherche donnée ?

## Réseaux biologiques

Les réseaux trouvent donc pleinement leur place dans la représentation et l'étude d'associations complexes, directes ou indirectes, entre différents acteurs, pour mieux visualiser les données à grande échelle, et enfin finalement, comprendre un système biologique. Les réseaux biologiques présentent très souvent une topologie particulière de réseau invariant d'échelle (*scale free*) impliquant que la distribution des degrés de leurs nœuds suit une loi de puissance. Autrement dit, leur structure n'est pas uniforme car la plupart des nœuds montre peu de liens (présentant donc un degré bas), alors qu'un petit nombre présente beaucoup de connexions (degré élevé, ces nœuds sont appelés *hubs*) maintenant ainsi l'unité de l'ensemble (Barabási & Oltvai, 2004). Un exemple qui illustre parfaitement cette structure est celui des réseaux métaboliques. Étant donné que beaucoup de réactions sont



irréversibles, il s'agit d'un réseau orienté. Ils présentent, comme la plupart des réseaux complexes une structure en « petit monde » (*small-world*) *i.e.* en prenant deux nœuds au hasard, ils peuvent être connectés par un chemin relativement court passant par des liens existants. Dans ce cadre, ce chemin correspond à la réaction biochimique qui lie deux substrats. La plupart des substrats participe à une ou deux réactions mais certains, comme le pyruvate ou la coenzyme A sont transformés dans une douzaine de réactions différentes, formant ainsi des hubs (Jeong et al., 2001; Barabási & Oltvai, 2004).

Les réseaux d'interaction protéine-protéine représentent les interactions spécifiques observées expérimentalement entre protéines faisant intervenir des domaines de liaisons particulier dans le cadre d'une fonction donnée. Ces interactions peuvent être stables (formation de complexes protéiques) ou transitoires (par exemple réactions impliquant des protéines kinases ou des complexes multiprotéiques régulateurs). L'interactome protéique permet d'associer des rôles à des protéines non caractérisées, de mettre en évidence des étapes de voies de signalisation ou de décrire les relations entre protéines formant des complexes. Ces réseaux présentent les mêmes caractéristiques topologiques que les réseaux métaboliques (*i.e. small-world et scale-free*). Les hubs que l'on retrouve dans ce type de réseaux correspondent en général aux protéines essentielles à la survie de l'organisme (Jeong et al., 2001).

Un autre niveau d'association peut-être représenté par les réseaux de régulation de gènes représentant l'interaction privilégiée et dirigée de protéines régulatrices avec des régions promotrices de gènes pour contrôler l'expression de gènes cibles et, *in fine* l'abondance de leurs produits. A l'image des processus qu'ils modèlent, ces réseaux sont généralement complexes, contenant des inter-dépendances inductives ou répressives ainsi que des boucles de rétro-contrôle (Karlebach & Shamir, 2008; Vijesh, Chakrabarti & Sreekumar, 2013).

Les activités cellulaires et leurs modulations face à des stimuli internes et externes sont gouvernées par des mécanismes de communication très élaborés. Les voies de transduction d'un signal, de sa perception par un récepteur jusqu'à la réponse finale sont rarement linéaires et font intervenir des inter-connexions spécifiques avec d'autres voies. Cette complexité est très bien illustrée par la signalisation hormonale chez les plantes par exemple. Ces connexions multiples sont régulées de manière extrêmement précise dans l'espace et le temps ; un même signal, selon son intensité ou sa localisation, pouvant conduire à une réponse cellulaire différente. La représentation de cette complexité aboutit à l'existence réelle de réseaux de



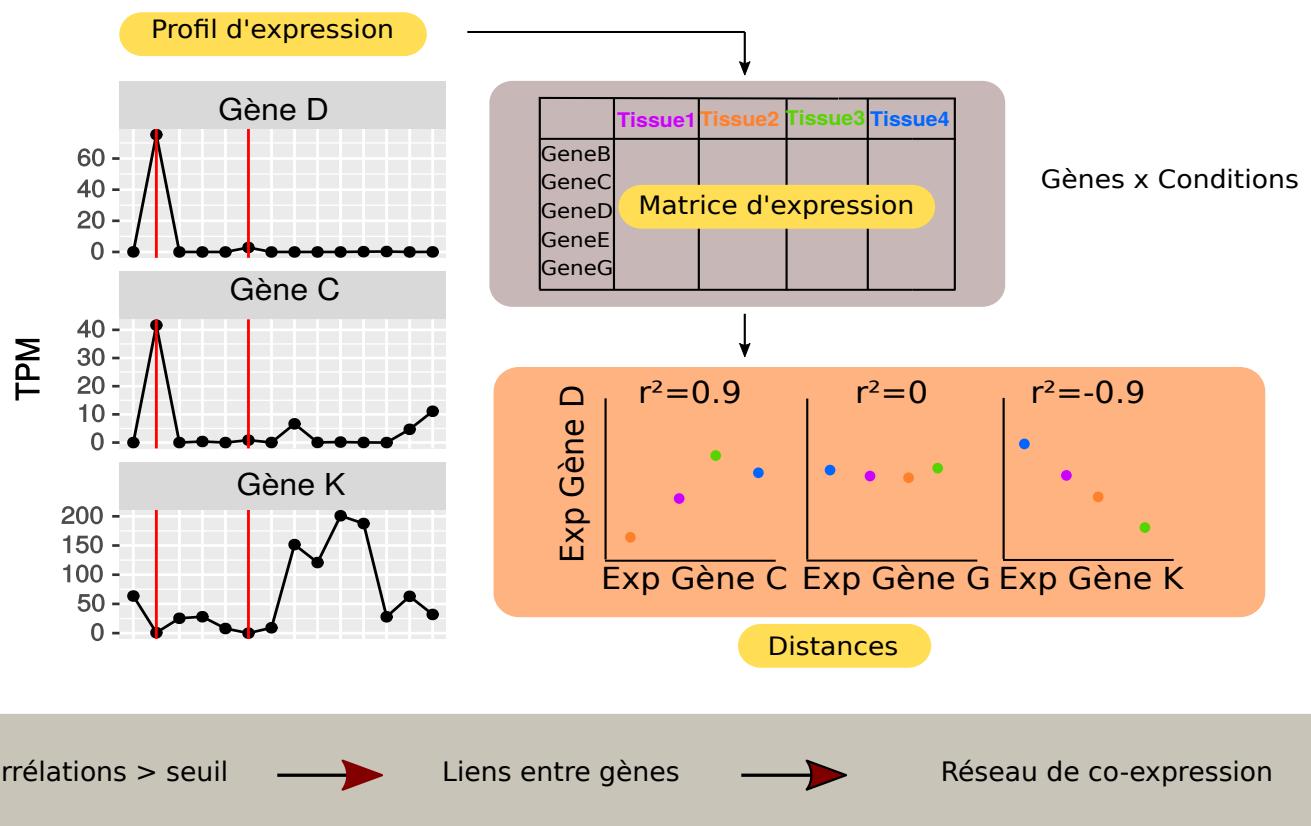
signalisation (Pawson & Saxton, 1999; Jordan, Landau & Iyengar, 2000). Mais les réseaux biologiques existent également à des niveaux supra-moléculaires reflétant des interactions entre populations par exemple (Krause, Lusseau & James, 2009).

En parallèle des réseaux de régulation génique qui impliquent des dépendances directes entre acteurs, les réseaux de co-expression associent des gènes dont l'expression conjointe dans des conditions spécifiques les prédispose à intervenir dans des processus communs. L'étude de processus biologiques par réseaux de co-expression étant l'aspect principal de ce travail de thèse, ils seront abordés de manière plus détaillée ci-dessous. Les différentes données « -omiques » sont donc parfaitement prédisposées pour être utilisées dans la construction de réseaux biologiques robustes et analysables par des approches adaptées, afin de contribuer à l'élucidation du fonctionnement de systèmes biologiques.

## 1. Les réseaux de co-expression

Le transcriptome représente l'ensemble des ARN transcrits à partir de l'ADN d'une cellule ou d'un organisme. Contrairement au génome, ce dernier est hautement dynamique, le contenu en ARN variant fortement en fonction du type cellulaire et des conditions biologiques<sup>3</sup>. Représentant une information qualitative et quantitative sur les transcrits d'une cellule, d'un tissu ou d'un organisme à un moment donné, il fournit un niveau d'information supplémentaire par rapport à l'exome. Son étude permet une meilleure compréhension du lien entre génotype et phénotype et constitue ainsi un outil précieux ouvrant de nouvelles perspectives dans l'élucidation de processus moléculaires et physiologiques complexes. Le développement de nouvelles technologies de séquençage à haut débit toujours plus performantes (voir **paragraphe 2.1**), a clairement permis l'accès à ces transcriptomes, en particulier pour des espèces non modèles même en l'absence de séquence génomique préalable. Par ailleurs, ces technologies permettent une vision plus globale et sans a priori du

<sup>3</sup> Certains gènes sont exprimés même après la mort de l'organisme dont certains qui étaient réprimés depuis le développement embryonnaire. Ils constituent ce qui est appelé le Thanathotranscriptome. Pozhitkov 2016

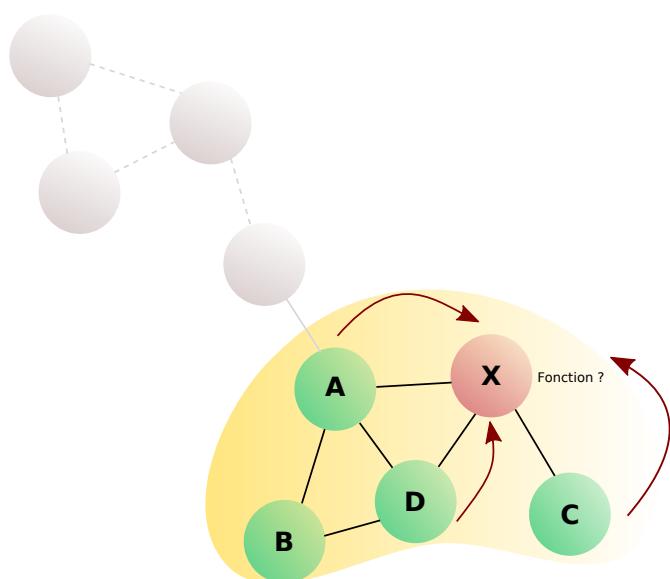


Corrélations > seuil → Liens entre gènes → Réseau de co-expression

Exemple: avec un seuil de distance de 0.7



**Figure 6: Co-expression de gènes.** Une matrice d'expression est construite à partir des profils d'expression de gènes dans différentes conditions expérimentales. A partir de ces valeurs, des corrélations entre gènes vont pouvoir être établis en fonction des similarités de ces profils. Si la valeur de distance calculée entre deux gènes se situe au-dessus d'un certain seuil, un lien peut-être établi entre ces derniers. L'ensemble des liens retenus va former le réseau de co-expression.



**Figure 7: Principe de "culpabilité par association" (Guilt By Association, GBA).** Dans une communauté de gènes densément connectés à l'intérieur du réseau, les gènes sont susceptibles d'intervenir dans un même processus biologique. Dans cet exemple, la fonction du gène X est inconnue, alors que les autres gènes de la communauté (en verts) ont été décrits préalablement. Ainsi il est probable que le gène X partage des fonctionnalités avec ces derniers.

système étudié puisqu’elles permettent de capturer le transcriptome plus ou moins dans sa globalité.

Une manière d’appréhender ce transcriptome consiste à identifier les gènes qui présentent des profils d’expression semblables *i.e.* qui s’expriment de manière similaire dans des conditions expérimentales données (**Figure 6**). En effet, cette co-expression de gènes reflète une potentielle action conjointe dans un processus donnée, décrite par le principe de « culpabilité par association » (*Guilt-By-Association* ou GBA) (Wolfe, Kohane & Butte, 2005) (**Figure 7**). Ce principe a été notamment illustré par Hughes et al (Hughes et al., 2000) qui montre que la perturbation de gènes d’un même parcours conduit à des profils d’expression similaires des gènes influencés par ce dernier. Bien que ce postulat soit parfois remis en cause (Gillis & Pavlidis, 2011), il reste une force majeure pour l’interprétation des réseaux de co-expression. En effet, des gènes d’une même voie ont tendance à montrer une concordance transcriptionnelle plus importante que ceux qui ne le sont pas (Kumari et al., 2012).

Les réseaux de co-expression sont basés sur une telle comparaison globale des profils d’expression. Ce sont des réseaux non orientés, contrairement aux réseaux de régulation, et le lien entre deux gènes ne permet pas de conclure sur la relation hiérarchique entre les deux. Leur analyse permet cependant de mettre en évidence des liens directs et indirects entre gènes *a priori* impliqués dans un même processus biologique selon le principe GBA décrit ci-dessus. Ceci implique que dans un réseau de co-expression, des gènes situés dans un même environnement, c’est-à-dire associés à un groupe se détachant du reste du réseau, auront une forte probabilité de participer au même processus cellulaire. Par conséquent, un réseau de co-expression constitue aussi une ressource importante pour l’annotation de gènes dont la fonction est encore inconnue. Si la structure de la séquence nucléique ou protéique permet d’identifier des domaines fonctionnels qui peuvent être reliés à des activités biochimiques, l’étude de co-expression permet en plus de mettre en lumière les gènes situés dans le voisinage transcriptionnel. En fonction des conditions expérimentales étudiées ainsi que des processus connus faisant intervenir les gènes co-exprimés avec un ou des candidat(s), des informations sur le processus biologique comme un stress ou une maladie par exemple impliquant ce(s) gène(s) peuvent être extraites. Ainsi, les objectifs d’une étude de co-



expression sont doubles : (i) étudier des environnements transcriptionnels pour mieux comprendre un processus cellulaire donné et (ii) attribuer une fonction moléculaire ou cellulaire et un rôle physiologique à un ou plusieurs gènes. Cette approche peut être effectuée en parallèle d'autres types d'analyses, comme la mise en évidence de gènes différentiellement exprimés en fonction d'une condition ou d'un tissu, ou un *clustering* de gènes par exemple afin de produire un échantillon réduit de gènes candidats dont le rôle restera à confirmer *in vitro* et *in vivo*.

Par ailleurs, des études phylogénétiques peuvent être menées en comparant les réseaux de co-expression de différentes espèces apparentées, afin de mettre en évidence les processus évolutifs qui se sont produits lors de la spéciation de différents organismes (Ruprecht et al., 2017). La génomique comparative, basée sur l'analyse de similarités entre séquences, permet en effet d'inférer des relations entre gènes et organismes afin d'étudier leur évolution, ainsi que, dans une certaine mesure, de faciliter l'annotation de génomes d'espèces non modèles, mais comporte toutefois des limites. Ces approches se basent sur la notion que les fonctions biologiques émergent et évoluent en tant que famille de gènes. Toutefois, les gènes et les familles de gènes ne fonctionnent que rarement comme une entité unique mais plutôt sous forme de modules fonctionnels (Hartwell et al., 1999). Ces modules comprennent des gènes et des produits de ces derniers (ARN, protéines) différents qui vont interagir pour remplir une fonction précise. Encore une fois, ces fonctions peuvent rester incomprises lorsque les gènes sont étudiés individuellement. De plus, en particulier chez les plantes, les familles de gènes peuvent être très grandes et compter de nombreux homologues qui ont pourtant des fonctions divergentes (Kliebenstein et al., 2001). Une analyse génomique basée sur les séquences de ces gènes dupliqués ne permet donc pas de mettre en évidence des événements de néofonctionnalisation par exemple (Lynch & Katju, 2004). Dans ce cas, étudier la conservation de modules de co-expression entre espèces peut conduire à une identification plus précise de gènes orthologues avec une fonction commune tout, en transférant les connaissances d'une espèce modèle vers des espèces non modèles (voir par exemple les travaux sur la conservation de gènes co-exprimés au niveau de carcinomes mammaires murins et tumeurs du sein chez l'humain (Herschkowitz et al., 2007) ).



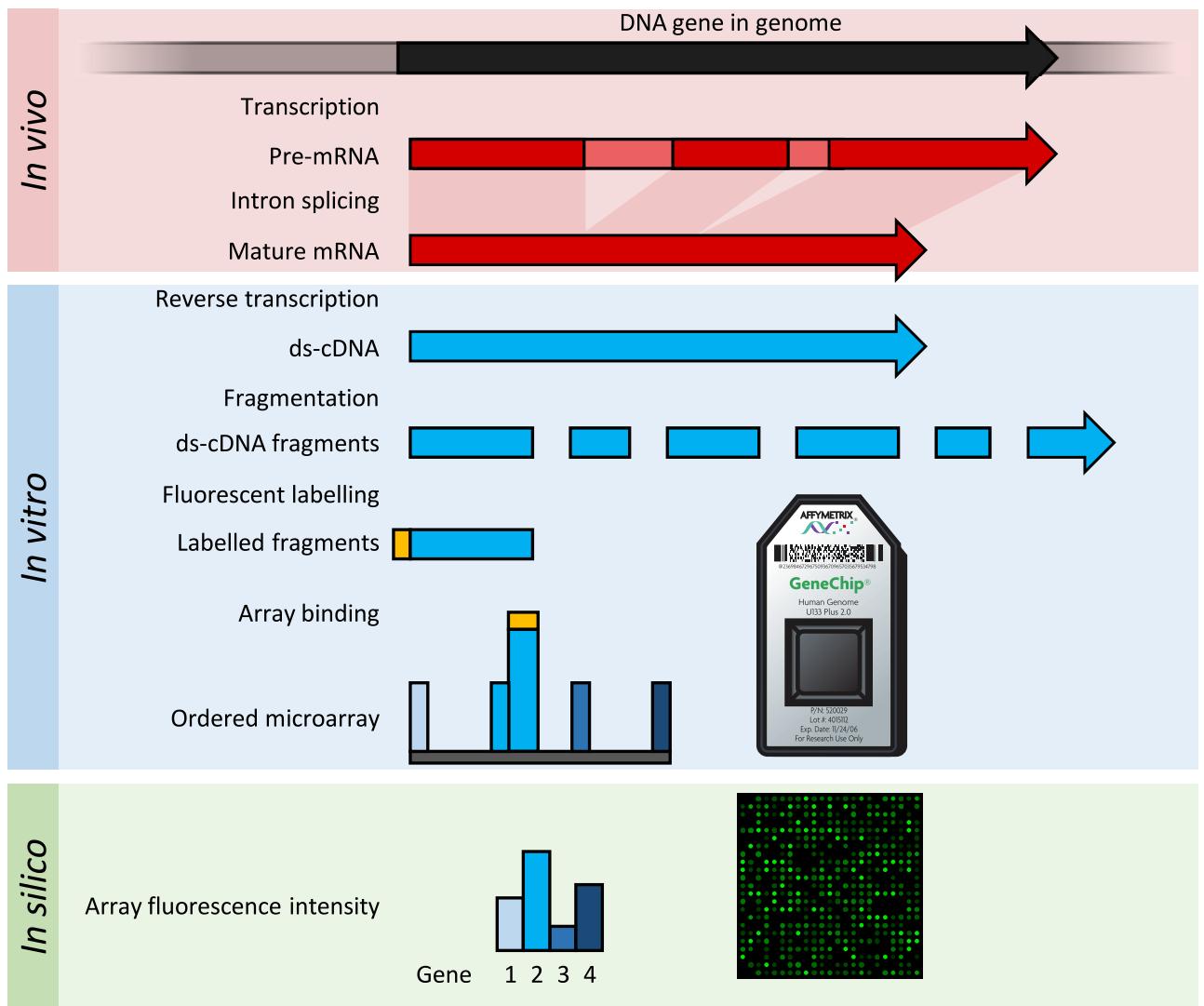
Les réseaux de co-expression de gènes sont donc un outil précieux dans l'étude de processus physiologiques permettant, du moins au niveau du transcriptome, une vue globale des différents intervenants potentiels et de leurs interactions. Un grand nombre d'outils et de bases de données, tirant profit de la quantité de données disponible toujours plus importante, a été développé au cours des dernières années pour un large éventail d'organismes, de tissus biologiques ainsi que pour l'étude de thématiques spécifiques (**Table I**) (Mutwil et al., 2011b; Obayashi et al., 2013, 2018; Yim et al., 2013; van Dam, Craig & de Magalhães, 2015; Wang et al., 2015; Aoki et al., 2016; Yang et al., 2017).

Compte tenu du nombre croissant de travaux scientifiques reposant sur des études de co-expression de gènes, la question se pose si les gènes identifiés comme co-exprimés *in silico* interviennent réellement dans un même processus biologique *in vivo*? La véracité du principe d'association a pu être démontrée à multiples reprises. En 2005, Brown et al ainsi que Persson et al. ont ainsi identifié par différentes approches des gènes co-exprimés avec des gènes connus pour être impliqués dans la synthèse de la paroi secondaire chez *Arabidopsis thaliana* (Persson et al., 2005; Brown et al., 2005). L'implication dans ce processus de certains candidats mis en évidence dans ces études a ensuite été confirmée par des approches de génétique inverse. Koo et al et Hirai et al ont respectivement mis en évidence et confirmé par ce type d'approche, de nouveaux gènes impliqués respectivement dans la voie de biosynthèse de l'acide jasmonique et des glucosinolates (Koo et al., 2006; Hirai et al., 2007). Cette approche apparaît donc comme particulièrement adaptée pour l'étude de voies cellulaires comme les voies métaboliques et de signalisation. En effet, des gènes d'une même voie ont tendance à montrer une concordance plus importante que ceux qui ne le sont pas (Kumari et al., 2012). Associer des gènes qui sont exprimés similairement dans des conditions communes permet notamment d'identifier des gènes manquants dans les différentes étapes d'une voie donnée, ou de mettre en évidence son implication dans un processus particulier.



**Table I: Bases de données permettant des études de co-expression à partir de gènes candidats**

Base de données	Organismes	Mesure de co-expression	Données	Reference
AlcoDB	Algues	Rang mutuel (MR)	Microarray/RNA-seq	Aoki et al. (2016)
Atted-II	Plantes	Rang mutuel (MR)	Microarray/RNA-seq	Obayashi et al 2018
Coexpedia	Humain et souris	Coefficient de corrélation de Pearson (PCC), algorithme bayésien, bibliographie	Microarray	Yang et al. (2017)
CoXPRESdb	Divers organismes modèles (levures, plantes, nématode, souris, humain etc.)	Rang mutuel (MR)	Microarray/RNA-seq	Okamura et al. 2015
GeneFriends	Divers organismes modèles (souris, rat, mouche, levure, humain).	Rang mutuel (MR)	Microarray/RNA-seq	Van Dam et al. 2015
ImmCo	Cellules du système immunitaire (humain et souris)	Coefficient de corrélation de Pearson (PCC)	Microarray	Wang et al (2016)
PlaNet	Plantes	Meilleur rang réciproque (HRR)	Microarray	Mutwill et al 2011
PLANEX	Plantes	Coefficient de corrélation de Pearson (PCC)	Microarray	Yim et al (2015)
PODC	Plantes	Analyse de correspondance (CA)	RNA-seq	Ohyanagi et al (2014)
CoExpNetViz	Plantes	MR/MI	RNA-seq	Tzafdia et al 2016



**Figure 8: Représentation schématique de l'approche microarray (puce à ADN)** (extrait de Lowe et al., 2017). *In vivo*, l'ADN génomique va être transcrit en ARNmessager (avec épissages des introns chez les eucaryotes). *In vitro*, cet ARNm va être extrait, rétrotranscrit en ADNcomplémentaire (cDNA), fragmenté et marqué avec des fluorophores. Ces fragments de cDNA vont s'hybrider sur des sondes complémentaires représentant un grand nombre de gènes d'intérêt sur la lame de la puce. La lame va ensuite être scannée, et *in silico*, la fluorescence représentative de l'abondance des transcrits correspondant aux sondes va pouvoir être mesurée.

## 2. Construction des réseaux

### 2.1. Données d'expression issues de technologies à haut débit

De nouvelles technologies, permettant d'appréhender le transcriptome à grande échelle, se sont rapidement développées au cours des dernières décennies. Nous aborderons ici les deux procédés principaux à l'origine des données utilisées pendant ce travail de thèse que sont les *microarrays* et le séquençage de l'ARN.

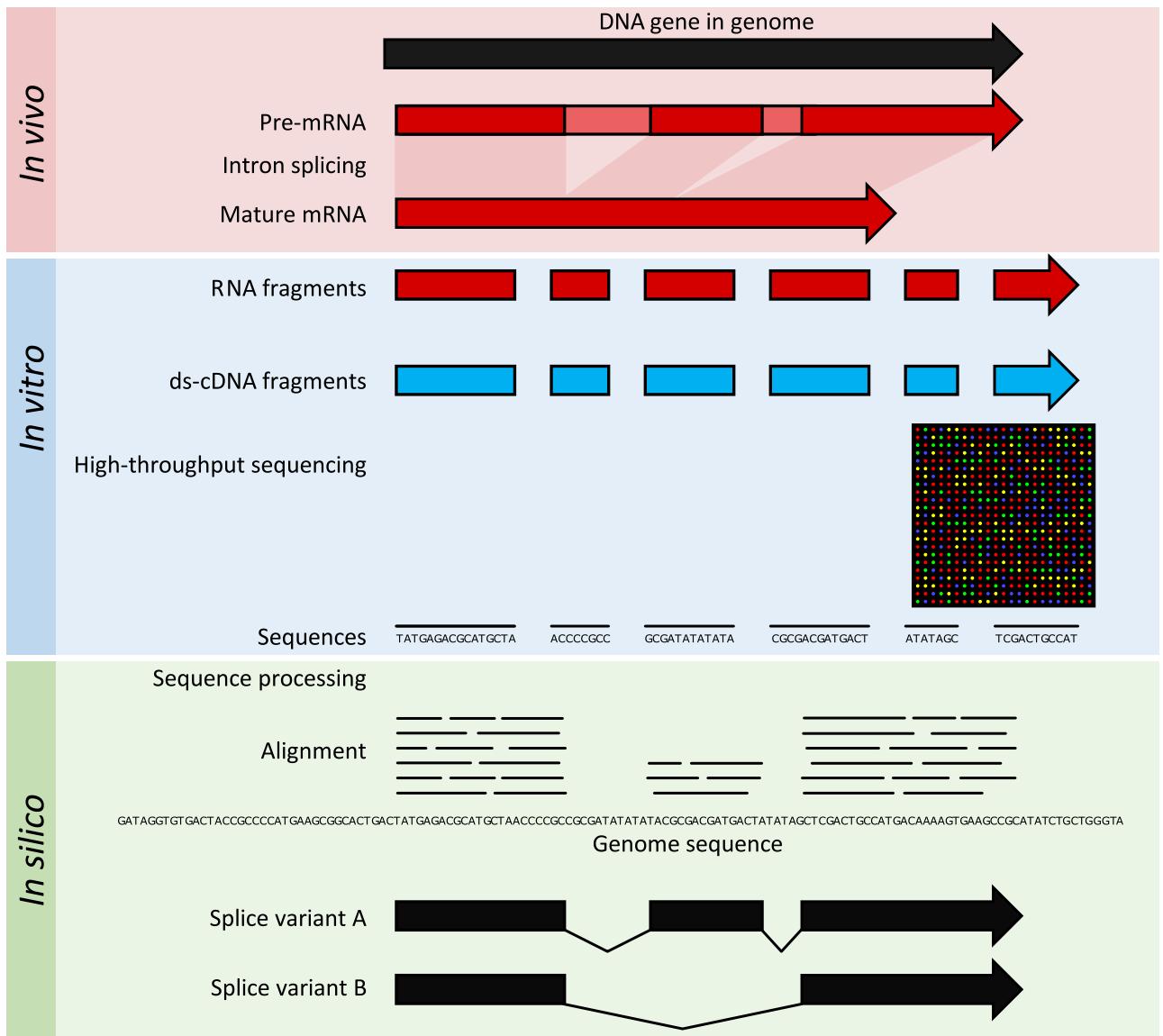
#### 2.1.1. Microarray

L'utilisation de *DNA microarray* ou puces à ADN (**Figure 8**) pour quantifier l'expression de multiples gènes a été décrite pour la première fois en 1995 par Schena et al. utilisant 45 séquences d'ADNc (ADN complémentaire) de la plante modèle *Arabidopsis thaliana* (Schena et al., 1995). Cette technique repose sur les propriétés d'hybridation des acides nucléiques pour identifier la présence et éventuellement quantifier un certain nombre de cibles. Le concept implique d'hybrider un mélange de molécules d'ADN sur une puce présentant des molécules d'ADN réceptrices ou sondes complémentaires d'une partie ou de l'ensemble des séquences contenues dans le mélange. Chaque brin cible va s'hybrider aux sondes qui lui sont complémentaires et former un duplex double brin sonde/cible. Le marqueur permet de suivre l'hybridation. Comme les coordonnées des sondes sont connues sur la lame, si un signal est détecté sur l'une d'entre elles après hybridation du mélange, cela implique que la cible est effectivement présente dans le mélange. Dans le cas d'analyses transcriptomiques, les ARN totaux sont extraits des échantillons étudiés et les ARN messagers rétrotranscrits en ADN complémentaire (ADNc) et marqués. Les marquages mono-couleur (*single channel microarray*) reposent soit sur le couple biotine/streptavidine (Affymetrix) soit sur la cyanine 3 (Cy3 ; Agilent par exemple), et permettent de mesurer approximativement le niveau d'expression absolu des gènes. Plus généralement, les microarrays *dual channel* utilisent deux molécules (ou plus (Woo et al., 2004)) qui vont être excitées par des longueurs d'ondes



différentes : la cyanine 3 (Cy3, fluorochrome vert) et la cyanine 5 (Cy5, fluorochrome rouge) permettant de comparer simultanément deux conditions expérimentales dont les ADNC auront été marqués avec l'un ou l'autre sur une même lame. Dans ce cas, un ratio d'expression est mesuré entre les deux échantillons. La lame va ensuite pouvoir être scannée et l'image résultante, comportant une grille de spots fluorescents, analysée afin d'associer des valeurs d'intensité de fluorescence (visibles sous forme de pixels) à chaque sonde de la puce, permettant de mesurer l'expression des gènes correspondant à cette dernière (**Figure 8**). Les microarrays actuels permettent d'analyser simultanément l'intégralité des transcrits d'une cellule, d'un tissu ou bien d'un organisme entier grâce à des développements techniques comme les puces à oligomères. Dans ces derniers développements, chaque gène est représenté par plusieurs oligomères pour optimiser la sensibilité et la précision de la détection.

Les résultats de scan des lames sont brutes et plusieurs étapes sont nécessaires avant de pouvoir procéder à l'analyse proprement dite des expressions de gènes. Dans un premier temps, la qualité des spots est analysée en fonction de la qualité de la fluorescence et ceux de mauvaise qualité peuvent être retirés. Les signaux bruts de fluorescence sont alors normalisés pour les rendre homogènes sur la lame et comparables entre lames. Pour cela, le bruit de fond est généralement retiré (une valeur pour chaque spot basée sur le signal environnant hors spot). Il existe de nombreuses manières de normaliser les signaux (Quackenbush, 2002; Smyth & Speed, 2003). L'une d'entre elle, qui sera utilisée par la suite dans ce travail, est la normalisation par quantile (Bolstad et al., 2003). Dans les grandes lignes, elle suggère que toutes les lames utilisées dans une étude présentent une même distribution et substitue les quantiles de chaque lame par la moyenne des quantiles de l'ensemble des lames. Par une transformation algorithmique, il est alors possible de corriger l'intensité des spots sur chaque lame par rapport à ce quantile moyen.



**Figure 9: Représentation schématique du processus de RNAseq (extrait de Lowe et al., 2017).** L'ADN génomique est transcrit en ARNmessager mature (introns épissés chez les eucaryotes) *in vivo*. *In vitro*, cet ARN va être extrait de l'échantillon, fragmenté et être rétro-transcrit en ADNcomplémentaire (cDNA) qui va pouvoir être séquencé après une étape de préparation de banques de cDNA. Ce séquençage va produire des lectures courtes qui, *in silico*, vont pouvoir être alignés sur un génome de référence afin de mesurer l'abondance des transcrits ou de détecter des variants d'épissage.

### 2.1.2. RNA-seq

Le *RNA-sequencing* ou RNA-seq repose, comme son nom l'indique, sur le séquençage à haut débit d'un pool d'ARN (généralement restreint aux ARNm dans le cadre d'une étude transcriptomique) (**Figure 9**). Comme précédemment, ce pool va être converti en une banque d'ADN complémentaire par rétro-transcription. L'ADNc est ensuite fragmenté de façon aléatoire par des méthodes mécaniques ou chimiques (cette étape peut aussi être réalisée directement sur l'ARN purifié) et flanqué d'adaptateurs constitués de séquences connues sur les extrémités 5' et 3' des fragments, nécessaires pour le séquençage. La technologie de *Next Generation Sequencing* (NGS) de 2nde génération la plus utilisée est clairement la technologie Illumina (ex Solexa) qui génère des lectures plutôt courtes (<250 pb) mais à forte profondeur (du million à la centaine de million de lectures) selon le principe de séquençage par synthèse à l'aide de terminateurs fluorescents et réversibles. Il existe cependant d'autres plateformes qui ont été également utilisées (pyroséquençage ROCHE 454, ABI SOLiD). Deux paramètres intrinsèquement liés sont à considérer pour la mise en place du séquençage : la profondeur (nombre de lectures obtenues) et la couverture (nombre de fois en moyenne que chaque nucléotide est théoriquement séquencé) (Sims et al., 2014). Après le séquençage à haut débit, de petits morceaux de fragments appelés lectures ou *reads* sont obtenus (pour un aperçu des différentes technologies disponibles, voir (Goodwin, McPherson & McCombie, 2016)). Puisque les molécules initiales ont été fragmentées, les données de séquençage doivent être traitées pour les reconstruire intégralement. La stratégie à mener est dépendante de la disponibilité de séquences de référence. Dans le cas d'un génome ou d'un transcriptome déjà connus, les lectures sont alignées sur cette référence par des algorithmes spécialisés impliquant des transformées Burrows-Wheeler (c'est le cas du très communément utilisé Bowtie2 (Langmead & Salzberg, 2012)). Puisque la profondeur de séquençage est élevée, de tels développements algorithmiques ont été nécessaires pour diminuer les temps de traitement ainsi que les besoins en ressources informatiques. Il existe d'autres approches telles que celle utilisée dans STAR (Dobin et al., 2013), reposant sur un processus biphasique de recherche d'identité de la lecture puis de l'attribution finale de la position génomique. L'intérêt de cartographier sur une référence est aussi de mettre en évidence des transcrits alternatifs (e.g. TopHat2 & Hisat2 (Pertea et al., 2016)). Lorsqu'aucune séquence de référence n'est disponible, des algorithmes spécifiques peuvent construire ce dernier *de novo* en identifiant



les chevauchements possibles entre les lectures (Nagarajan & Pop, 2013). Le RNA-seq permet une approche qualitative (annotation de transcriptomes, mise en évidence de jonctions entre introns et exons, de transcrits alternatifs, de sites d'initiation de la transcription, gènes de fusion etc.) et bien sûr quantitative, puisqu'elle permet de quantifier l'expression de gènes (Ozsolak & Milos, 2011). La difficulté d'une quantification précise est liée à l'attribution correcte des lectures sur la séquence de référence. Etant donné que l'échantillon en amont est fragmenté, des lectures courtes pourraient être attribuées à plusieurs séquences très similaires *in silico* alors qu'elles ne proviennent en réalité que d'une d'entre elles. Des algorithmes poussés utilisent les fichiers d'alignement pour optimiser l'attribution des lectures, notamment par des phases d'espérance-maximization de la probabilité d'appartenance à une séquence données dans la référence. C'est notamment le cas du programme RSEM (Li et al., 2010; Li & Dewey, 2011), précurseur dans ce type d'approche. Bien entendu, les incertitudes d'attribution sont réduites par des lectures plus longues et un design paired-end. De nouvelles approches ont été récemment développées pour diminuer les temps de quantification tout en préservant voire en augmentant la précision. Elles reposent sur le principe de pseudo-alignement (Kallisto (Bray et al., 2016)) ou de quasi-mapping (Salmon (Patro et al., 2017)). Ces méthodes génèrent des sorties beaucoup moins lourdes et nécessitent des ressources informatiques faibles, justifiant leur utilisation dans le traitement de nombreuses données.

### **2.1.3. Microarray *versus* RNA-seq : avantages et inconvénients**

Quelque soit la technique utilisée, pour estimer l'abondance des transcrits, l'objectif est de capturer les transcriptomes exhaustifs d'échantillons afin de mettre en évidence une variabilité biologique par des expressions différentielles ou communes de gènes entre différentes conditions expérimentales. Dans les deux cas, la préparation des échantillons nécessite plusieurs étapes, chacune étant susceptible d'apporter une variabilité d'origine technique et donc un biais dans les mesures subséquentes (Chen, 2007; Hitzemann et al., 2013). Les procédures de normalisation sont censées limiter ces biais.



D'un premier abord, le séquençage semble présenter un certain nombre d'avantages par rapport au microarray. Premièrement et comme vu précédemment, une approche microarrays nécessite une connaissance génomique suffisante de l'organisme pour construire des sondes complémentaires. Deuxièmement, il peut être plus difficile de discriminer des transcrits comportant des séquences très similaires ce qui résulter en des phénomènes de cross-hybridation et *in fine* des niveaux d'expression erronés.. Enfin, la gamme dynamique de détection de l'abondance est limitée par l'intensité du signal fluorescent, rendant moins sensible la détection de transcrits en très faible et dans une moindre mesure d'échelonner les différences d'expression à très forte abondance due à une saturation des sondes (T-cells, (Hitzemann et al., 2013)). Comme abordé précédemment, le RNA-seq présente une vision globale du transcriptome avec une meilleure résolution en nécessitant une quantité d'ARN moindre que le microarray (10 pg contre 200 ng respectivement (Mantione et al., 2014), et ne nécessite pas obligatoirement de référence, ce qui est une opportunité significative pour les espèces non modèles.D'ailleurs beaucoup de projets de séquençage à grande échelle de transcriptomes ont été initiés ces dernières années (cas des 1,000 Plant/Fish/Insect Transcriptomes). De plus, la réutilisation de ces données est facilement réalisable, puisque contrairement au microarray qui est limité aux connaissances prédéfinies au moment de l'élaboration des sondes de la lame, les données de séquençages peuvent être retravaillées et ré-annotées. A titre d'exemple, la puce d'hybridation Affymetrix ATH1, la plus utilisée pour la plante modèle *Arabidopsis thaliana*, comporte à l'heure actuelle 22 500 sondes représentant environ 24 000 gènes alors que l'annotation de son génome « TAIR10 » prédit un total de 33 602 gènes dont 27 416 coderaient pour des protéines (Lamesch et al., 2012; Giorgi, Del Fabbro & Licausi, 2013). Un aspect à ne pas négliger reste le coût financier, la dernière technique restant encore plus onéreuse à l'heure actuelle (environ 300\$/échantillon pour le microarray contre 1000\$/échantillon pour le RNA-seq en 2014) (Mantione et al., 2014), malgré une baisse significative des coûts dans les dernières années. Elle demande également un investissement en temps plus important pour le traitement des données puisque les données brutes comportant les lectures doivent être alignées avec une référence et être comptées, ainsi qu'en espace de stockage informatique rendant leur manipulation sur un ordinateur personnel difficile (Hitzemann et al., 2013; Mantione et al., 2014).



**Table II: Comparaison entre RNA-seq et microarray.**

	RNA-seq	Microarray
Coût moyen par échantillon	1000 €	300 €
Couverture du transcript complet	Normalement complète mais dépendant de la profondeur et du niveau d'expression	Sondes couvrant en moyenne 20 % du transcript
Couverture complète du transcriptome	Normalement complète sauf si biais technique	Limitée à la qualité du génome utilisé pour la construction de la puce
Nécessite connaissance préalable du génome	Pas obligatoire	Oui sauf pour des espèces proches (hybridation inter espèces possible)
Détection de transcrits alternatifs et isoformes	+	+-
Spécificité	+	- (Les designs actuels minorent cet aspect)
Sensibilité	+ (dépendant de la profondeur)	-
Pipelines d'analyse standardisées	+ (Tuxedo suite + edgeR/DESeq2)	+ (Limma, R)
Simplicité de l'analyse des données (temps, taille)	Nécessite des ressources infos plus conséquentes	Difficultés possibles sur la normalisation des données
Mesure de l'expression absolue	+	+-
Bruit de fond	+	-
Reproductibilité	+	-

Un « + » désigne une meilleure performance par rapport au « - ».

Globalement, les résultats obtenus avec les deux technologies sont fortement comparables, et même si à première vue, le RNA-seq semble être plus performant que le microarray, cela n'est pas toujours le cas en pratique (Kogenaru et al., 2012; Sîrbu et al., 2012; Giorgi, Del Fabbro & Licausi, 2013; Nazarov et al., 2017). Nazarov et al. démontrent, sur des données microarray et RNAseq humaines, que malgré une forte constance entre les résultats issus des deux technologies, la première montrait de meilleures performances, notamment pour la détection de transcrits alternatifs et de gènes courts ou faiblement exprimés

SRA

SRA

Advanced

Search

Help

Full ▾

[SRX3371179: GSM2843913: pwo1; Arabidopsis thaliana; RNA-Seq](#)  
 3 ILLUMINA (Illumina HiSeq 2000) runs: 79M spots, 14.2G bases, 9Gb downloads

Submitted by: NCBI (GEO)

Study: Genome-wide transcriptomic analysis of Arabidopsis mutants controlling nuclear size in comparison to the wildtype  
[PRJNA417457](#) • [SRP124476](#) • All experiments • All runs  
[show Abstract](#)

Sample: pwo1  
[SAMN07987890](#) • [SRS2668513](#) • All experiments • All runs  
 Organism: *Arabidopsis thaliana*

Library:

Instrument: Illumina HiSeq 2000  
 Strategy: RNA-Seq  
 Source: TRANSCRIPTOMIC  
 Selection: cDNA  
 Layout: PAIRED  
 Construction protocol: Material was harvested at 2 week-old seedling stage. RNA was isolated according to Rneasy Plant Mini Prep Kit (Qiagen)  
 According to TruSeq RNA Library Prep Kit v2 (Illumina)

Experiment attributes:

GEO Accession: GSM2843913

Links:

Runs: 3 runs, 79M spots, 14.2G bases, 9Gb

Run	# of Spots	# of Bases	Size	Published
<a href="#">SRR6264971</a>	26,754,546	4.8G	3.1Gb	2018-08-02
<a href="#">SRR6264972</a>	26,514,925	4.8G	3Gb	2018-08-02
<a href="#">SRR6264973</a>	25,713,266	4.6G	2.9Gb	2018-08-02

ID: 4707567

Send to: ▾

**Related information**[BioProject](#)[BioSample](#)[GEO DataSets](#)[Taxonomy](#)**Recent activity**[Turn Off](#) [Clear](#)

Your browsing activity is empty.

Home EMBL-EBI Services Research Training About us

# ArrayExpress

Search Examples: [E-MEXP-31](#), [cancer](#), [p53](#), [Geuvadis](#)

[advanced search](#)

[Contact Us](#) | [Login](#)

[ARRAYEXPRESS](#) / [BROWSE](#) / [E-MTAB-5821](#)

### E-MTAB-5821 - Microarray on wild-type and mutant Hesx1 embryonic stem (ES) cells cultured in serum/LIF

Status	<a href="#">Submitted on 7 June 2017, last updated on 13 June 2017, released on 7 June 2018</a>	
Organism	Mus musculus	
Samples (3)	<a href="#">Click for detailed sample information and links to data</a>	
Array (1)	<a href="#">A-AFFY-45 - Affymetrix GeneChip Mouse Genome 430 2.0 [Mouse430_2]</a>	
Protocols (5)	<a href="#">Click for detailed protocol information</a>	
Description	Molecular profile of Hesx1-deficient and Hesx1 wild-type ES cells cultured in serum/LIF conditions was evaluated by microarray analysis	
Experiment types	transcription profiling by array, case control design	
Contact	<a href="mailto:sara.pozzi.11@ucl.ac.uk">✉ Sara Pozzi &lt;sara.pozzi.11@ucl.ac.uk&gt;</a>	
MIAME	* — * — *	
	Platforms   Protocols   Variables   Processed   Raw	
Files	<a href="#">Investigation description</a> <a href="#">Sample and data relationship</a> <a href="#">Raw data (1)</a> <a href="#">Array design</a> <a href="#">Click to browse all available files</a>	<a href="#">↓ E-MTAB-5821.idf.txt</a> <a href="#">↓ E-MTAB-5821.sdrf.txt</a> <a href="#">↓ E-MTAB-5821.raw.1.zip</a> <a href="#">↓ A-AFFY-45.adf.txt</a>
Links	<a href="#">ArrayExpress - E-MTAB-5816</a> <a href="#">Send E-MTAB-5821 data to </a>	

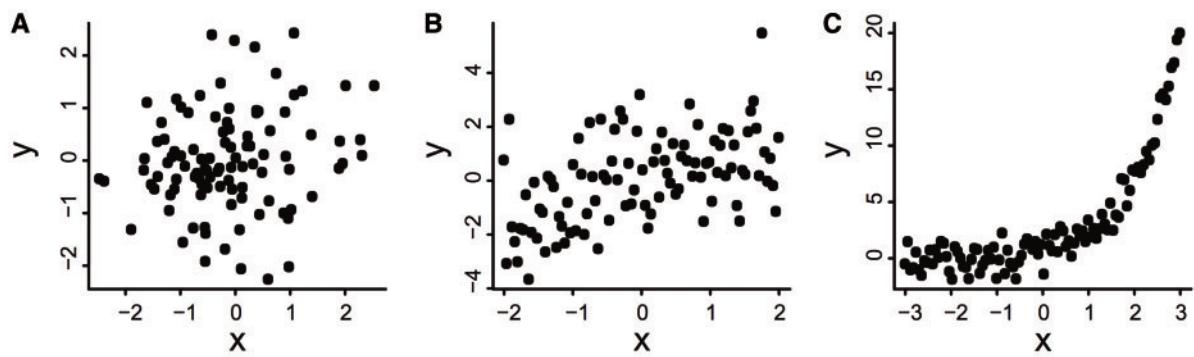
**Figure 10: Portails des bases de données SRA (NCBI) et ArrayExpress (EMBL-EBI).** Exemples de données RNAseq et microarray accessibles. Des informations précises sur chaque échantillon (séquenceur/puce utilisée, protocole expérimental, organisme étudié, laboratoire etc.) ainsi que les liens pour télécharger les données brutes sont disponibles. Liens correspondants : SRA: <https://www.ncbi.nlm.nih.gov/sra/SRX3371179>; ArrayExpress: <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5821>.

(Nazarov et al., 2017). Giorgi et al. arrivent à des conclusions similaires en considérant des réseaux de co-expression d'*Arabidopsis thaliana*, construits à partir de données issus des deux technologies avec des réseaux légèrement plus robustes pour le microarray (Giorgi, Del Fabbro & Licausi, 2013). Un avantage majeur du RNA-seq par rapport au microarray réside néanmoins dans l'aspect qualitatif qui donne accès à la nature des séquences, ce que le microarray n'apporte pas, ainsi qu'une meilleure réutilisation future des données. Dans l'idéal, chaque technique ayant ses propres limites et biais, les deux technologies devraient s'envisager de manière complémentaire pour l'étude de la transcription de gènes. Kogenaru et al. montrent dans une étude comparative que chaque technologie de transcriptomique apporte des informations communes et particulières, suggérant ainsi leur complémentarité dans les études transcriptomiques (Kogenaru et al., 2012). Une grande quantité de données microarray et RNA-seq (de l'ordre de plusieurs milliers d'échantillons pour les espèces modèles) issue d'un large panel d'études scientifiques est disponible librement sur les bases de données Gene Expression Omnibus de NCBI (National Center for Biotechnology Information, [www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/)) et ArrayExpress ([www.ebi.ac.uk/arrayexpress/](http://www.ebi.ac.uk/arrayexpress/)) du European Bioinformatics Institute (EMBL-EBI) (**Figure 10**).

## 2.2. Établir la co-expression

La co-expression est établie par le calcul d'une distance entre deux gènes ou bien entre un gène et un point théorique (k-moyennes) en utilisant leurs niveaux d'expression. Il s'agit d'identifier des ressemblances dans leurs profils d'expression : *i.e.* si ces gènes sont exprimés de manière semblable dans différentes conditions expérimentales (**Figure 11**). La méthode choisie pour établir cette distance entre deux gènes va grandement influencer la topologie du réseau sous-jacent ; il s'agit donc d'une étape cruciale dans l'élaboration de réseaux de co-expression. Elles sont basées soit sur des coefficients de corrélation, soit sur des approches itératives d'apprentissage supervisées ou non.

*Note au lecteur*



**Figure 11: Exemples de différents types de corrélations entre deux variables** (extrait de Siquira Santos et al., 2013) Les axes représentent les valeur d'expression des deux variables X et Y. A. Données indépendantes B. Association linéaire C. Association exponentielle (non linéaire, monotonique)

Il est important de noter, qu'une corrélation entre deux variables ne permet pas de conclusion sur la causalité. De plus, se focalisant au niveau transcriptome, les régulations post-traductionnelles ne sont pas prises en compte dans un réseau de co-expression alors qu'elles peuvent expliquer le fonctionnement de voies cellulaires. Néanmoins, la corrélation entre profils d'expression de gènes fournit une information précieuse sur leurs relations et dans une certaine mesure des causalités peuvent être inférées à partir de ces valeurs (Irvine, 2018).

### 2.2.1. Coefficients de corrélation

La mesure la plus populaire pour déterminer la co-expression entre gènes est le coefficient de corrélation de Pearson (PCC). Il s'agit du coefficient de corrélation obtenu dans le cadre d'une régression linéaire simple entre deux variables (par exemple deux gènes A et B) :

$$PCC_{A,B} = \frac{cov(A, B)}{\sigma_A \sigma_B}$$

Où cov (A ,B) désigne la covariance des variables A et B et  $\sigma$  leurs écarts types.

Ce coefficient sera égal à 1 en cas de corrélation parfaite entre les deux variables, à -1 en cas de corrélation négative (*i.e.* dans le contexte d'expression de gènes, ils vont s'exprimer de manière opposée), et égal à 0 lorsque la corrélation est nulle (absence de dépendance).

Il permet de détecter une corrélation linéaire entre deux variables, corrélation majoritairement attendue entre gènes (Song, Langfelder & Horvath, 2012). Cette mesure possède toutefois l'inconvénient d'être sensible aux valeurs aberrantes : deux gènes fortement corrélés peuvent présenter un PCC affaibli si l'un d'entre eux présente une expression aberrante dans une des conditions. Se pose toutefois la question de la distinction entre une valeur *a priori* aberrante liée à un artefact technique et une qui correspondrait à un véritable signal biologique dans une condition particulière.



Le coefficient de corrélation rang-ordre de Spearman est fondé sur les valeurs classées de PCC pour chaque variable plutôt que sur les données brutes et permet de mettre en évidence des relations monotones. Pour deux gènes, les valeurs d'expression vont être classées du niveau le plus faible jusqu'au niveau le plus fort pour chacun des gènes considérés. Ces valeurs ordonnées vont ensuite être utilisées avec la formule du PCC à la place des mesures d'expression brutes (Usadel et al., 2009). Le SCC est également compris entre -1 et 1. Contrairement au PCC, il est plus robuste face aux valeurs aberrantes et permet de détecter des relations non strictement linéaires.

De nombreux autres coefficients de corrélation utilisables pour établir la co-expression entre deux gènes existent. Le coefficient de Gini (GCC) par exemple, largement utilisé en économie, sociologie, physique et informatique, permet d'établir la corrélation entre deux variables en considérant à la fois la valeur brute et un rang. Il est plus robuste pour des données ne présentant pas une distribution normale et plus stable face aux valeurs aberrantes que les autres coefficients de corrélation, et il permet de mettre en évidence des relations linéaires et non-linéaires (Ma & Wang, 2012).

Ces coefficients basés sur les rangs des valeurs semblent être particulièrement indiqués pour identifier des gènes clés dont l'expression augmente et diminue de manière monotone au cours des conditions étudiées.

Une autre manière d'établir la co-expression entre deux variables repose sur le calcul de l'information mutuelle (*Mutual Information* ou MI). Elle mesure la dépendance statistique entre deux variables au sein du jeu de données. Les données d'expression doivent être représentées par des variables aléatoires discrètes. Pour deux gènes A et B, l'information mutuelle se calcule par :

$$MI_{A,B} = \sum i \sum j p_{ij} \log \frac{p_{ij}}{p_i p_j}$$



La valeur du MI est toujours positive. Elle est égale à zéro si et seulement si A et B sont statistiquement indépendants, signifiant que A ne contient aucune information sur B et *vice versa* (Priness, Maimon & Ben-Gal, 2007). Cette mesure permet de détecter aussi bien les relations linéaires que non linéaires. Pourtant, Steuer et al. montrent une concordance presque parfaite entre les résultats obtenus avec le PCC et la MI sur leur jeu de données (Steuer et al., 2002) et sur des données simulées. Bien qu’étant une mesure de distance largement utilisée pour établir la co-expression, aucune preuve n’a pu être apportée que la MI était plus performante que le PCC pour l’établissement de la co-expression entre gènes (Lindlof & Lubovac, 2005).

Le coefficient de corrélation partielle (*Partial Correlation* ou PC) quant à lui quantifie la corrélation entre deux variables en considérant d’autres. Par exemple entre les gènes A et B, quelle corrélation, conditionnée par une troisième variable, le gène C, existe-t-il entre A et B ? Cette corrélation entre A et B conditionnée par C peut être calculée par une corrélation de premier ordre (corrélation conditionnée par une variable) :

$$PC_{AB,C} = \frac{CC_{AB} - CC_{AC}CC_{BC}}{\sqrt{(1-CC_{AC}^2)(1-CC_{BC}^2)}}$$

Où CC désigne le coefficient de corrélation entre deux gènes. Les PC peuvent être de n’importe quel ordre, mais plus on considère de variables, plus les ressources informatiques sont nécessaires pour pouvoir effectuer le calcul (De La Fuente et al., 2004). D’autres approches computationnelles permettent cependant d’augmenter ce degré, notamment par régression linéaire multiple ajustée par des méthodes de régularisation (lorsque qu’il y a plus de gènes que d’échantillons). L’inversion d’une matrice de covariance donne également accès aux PC (Schäfer & Strimmer, 2005). L’analyse par corrélation partielle de deux variables permet de faire la distinction entre une relation directe ou indirecte entre ces dernières. Cela ne permet pas d’inférer une causalité directe, mais permet d’exclure un certain nombre d’associations potentielles. C’est la base de GGM (Gaussian Graphical Models) (López-Kleine, Leal & López, 2013).



Beaucoup d'autres méthodes non-paramétriques basées sur les valeurs de rangs ou non existent mais sont peu utilisées pour les études de co-expression (rangs pondérés, Coefficient de Hoeffding, Coefficient de Kendall, distance de covariance, Coefficient d'Information Maximale etc.) (Reshef et al., 2011; Kumari et al., 2012; de Siqueira Santos et al., 2013). .

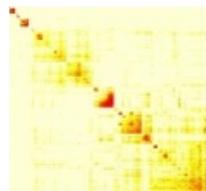
### **2.2.2. Classement de coefficients de corrélation**

Un des inconvénients majeurs des coefficients de corrélation linéaires comme le PCC, est qu'ils fournissent une information ponctuelle sur une similarité d'expression entre deux gènes et uniquement ces derniers. Il apparaît également que les rangs permettent d'améliorer les valeurs brutes d'un coefficient de corrélation, notamment concernant la sensibilité aux valeurs aberrantes (Obayashi & Kinoshita, 2009). Afin d'avoir une vue plus intégrative, à l'échelle du génome, une approche consiste à classer chaque gène par rapport aux autres en fonction du coefficient de corrélation qu'ils partagent (Figure rangs bien expliquer avec exemple). Les rangs attribués avec ce classement ne reflètent ainsi plus simplement une relation binaire, mais la distribution des corrélations du gène considéré au sein du transcriptome en son ensemble.

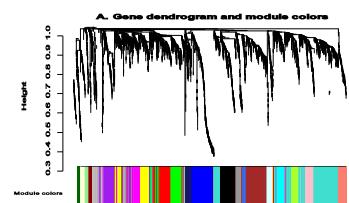
Deux mesures de classement à l'échelle du génome ont été décrites et sont utilisées pour établir la co-expression dans des bases de données bien établies (voir Table). La première méthode a été décrite en 2009 par Obayashi et Kinoshita (Obayashi & Kinoshita, 2009). en utilisant des rangs mutuels (*Mutual Rank* ou MR) basés sur les PCC et les listes de gènes co-exprimés ainsi générées pour l'annotation des gènes. Ces travaux démontrent que la performance des PCC ainsi classés est supérieure à la valeur brute de ces coefficients pour l'annotation fonctionnelle de gènes. Ces rangs mutuels entre deux gènes se calculent en retenant la moyenne arithmétique des rangs qu'occupe chaque gène par rapport à l'autre :

$$MR_{AB} = \frac{(Rang(A \rightarrow B) + Rang(B \rightarrow A))}{2}$$

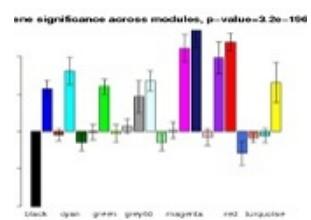
**Construction d'un réseau de co-expression** à partir de données transcriptomiques  
Mise en évidence d'interactions entre gènes (mesures de co-expression)



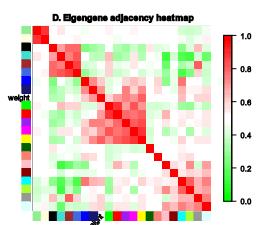
**Identification de modules de gènes** densément connectés (clustering hiérachique, Dynamic Tree Cut)



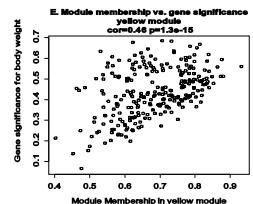
**Identification de modules biologiquement intéressants** par confrontation avec d'autres types de données (SNPs, données cliniques, enrichissements fonctionnels etc.)



**Analyse des relations entre modules** (Eigengene networks)



**Identification de gènes-clés à l'intérieur des modules** afin de mettre en évidence des candidats pertinents pour la validation expérimentale des biomarqueurs etc.  
(connectivité intra-modules, tests de causalité)



**Figure 12 : Principales étapes du workflow du package WGCNA** (d'après Langfelder et al., 2008). Ce paquet R contient de nombreux outils pour l'analyse complète de réseaux de co-expression à partir de larges jeux de données.

Une mesure ressemblante, mais plus stringente a été proposée par en 2010 (Mutwil et al., 2010). Cette mesure retenant le rang maximal entre ceux des deux gènes considérés est nommée le « plus haut rang réciproque » (*Highest Reciprocal Rank* ou HRR).

$$HRR_{AB} = \max(Rang(A \rightarrow B), Rang(B \rightarrow A))$$

Les coefficients de corrélation ainsi que leurs valeurs respectives classées nécessitent généralement de définir un seuil au-delà ou en-dessous duquel les gènes sont considérés comme significativement co-exprimés. Concernant le PCC, un seuil supérieur à environ 0.7 est globalement considéré comme reflétant une forte co-expression entre gènes (Couto, Comin & da Fontoura Costa, 2017). Ce seuil peut être fixé de manière empirique, en se basant sur la distribution des corrélations par exemple ou par des tests inférant une significativité statistique ou biologique aux corrélations calculées (Li, Pearl & Jackson, 2015).

### 2.2.3. Algorithmes complexes et co-expression de gènes

Il existe des algorithmes plus sophistiqués que ceux basés sur les mesures de corrélation pour établir si une dépendance existe entre deux gènes. Tout d'abord, certains programmes proposent des améliorations de Ccs, notamment sur la fixation du seuil de significativité. ARACNE par exemple (Algorithm for the Reconstruction of Accurate Cellular Networks) repose sur une MI à laquelle est assignée une *p-value* par comparaison avec des gènes sélectionnés aléatoirement qui sont donc présumés indépendants statistiquement, permettant de sélectionner par le biais d'un seuil des co-expressions robustes et ainsi la construction précise de réseaux de gènes (Margolin et al., 2006).

Un autre algorithme très largement utilisé est le *Weighted correlation network analysis* (WGCNA) qui met en évidence des clusters de gènes fortement co-exprimés (**Figure 12**). Il repose sur les valeurs absolues de PCC élevés à une puissance donnée pour forcer le réseau résultant à suivre une loi de puissance (mais peut également être utilisé avec d'autres mesure comme les SCC ou la *biweight midcorrelation*, mesure basée sur les médianes). Cette



approche inclut également une analyse des connexions entre gènes par une *Topological Overlap Measure* qui permet de pondérer une distance entre deux gènes en fonction des autres gènes. Elle permet ainsi, d'identifier des modules de gènes fortement connectés au sein du réseau, par le biais d'algorithmes de *clustering* comme le partitionnement hiérarchique. Ces modules peuvent ensuite être confrontés à d'autres données disponibles comme des données de protéomique, génomique, d'ontologie de gènes etc. afin d'identifier des modules ayant une significativité biologique. Ces derniers sont ensuite traités afin d'y mettre en évidence les acteurs clés (Langfelder & Horvath, 2008). Lui et al ont par exemple utilisé cette méthode en comparant 58 échantillons de tissu pulmonaire humain tumoral et sain et ainsi identifier des gènes impliqués dans la régulation de cancers du poumons et constituant de potentiels marqueurs diagnostiques (Liu et al., 2015).

Des modèles graphiques probabilistiques combinent la théorie des graphes et les lois de probabilités pour créer des réseaux de régulation (réseau de co-expression probabilisé et orienté). Les réseaux bayesiens utilisant des données d'expression de gènes par exemple permettent la construction des réseaux de régulation de ces derniers. Ici, un gène A est considéré comme influant sur l'expression d'un gène B si ses niveaux d'expression suivent une distribution gaussienne conditionnée par les niveaux d'expression du gène A. Ces relations étant établies pour tous les gènes, elles nécessitent d'importantes ressources de calcul (Chai et al., 2014). Bien que démontrées comme très efficaces, ces méthodes sont très sensibles au design expérimental et par leur nature de graphes acycliques orientés, ne permettent pas de mettre en évidence des boucles de rétro-contrôle (Friedman, 2004; Li, Pearl & Jackson, 2015).

Une autre approche consiste à considérer directement des groupes de gènes co-exprimés, et donc partageant *a priori* une ou des fonction(s) biologiques, sous forme de partitions ou *clusters* plutôt que des paires de gènes. Un algorithme de partitionnement performant doit minimiser l'inertie intra-groupe pour obtenir des groupes les plus homogènes possibles tout en maximisant l'inertie inter-classe afin d'obtenir des sous-ensembles bien différenciés. Les algorithmes de regroupement (ou *clustering*) peuvent être divisés en deux groupes : le regroupement hiérarchique et non-hiérarchique. La première approche retourne des groupes



imbriqués représentés sous la forme d'un dendrogramme (Larrañaga et al., 2006). Deux approches peuvent être distinguées : soit en partant de partitions contenant des objets uniques qui vont être fusionnés de manière récursive (partitionnement hiérarchique agglomérant), soit en partant d'un groupe qui contient tous les objets et qui va être récursivement divisé en partitions plus petites (partitionnement hiérarchique divergent). Le partitionnement non-hiéarchique regroupe typiquement N objets en K groupes (par exemple N gènes en K groupes correspondants à des gènes fortement co-exprimés entre eux dans les conditions données) lors d'un processus itératif jusqu'à ce qu'un certain critère de qualité soit acquis (D'haeseleer, Liang & Somogyi, 2000). Un exemple de méthode largement utilisée en transcriptomique sont les k-moyennes (ou *k-means*) (MacQueen, 1967). Dans ce cas, N gènes sont partitionnés en K groupes, K étant déterminé préalablement par l'utilisateur. Il n'existe donc pas de groupes déterminés car ils dépendent du nombre de K fixé au préalable. K centroïdes vont être constitués, soit au hasard, soit par l'utilisateur afin de refléter des patterns d'expression représentatifs. Ainsi chaque gène de la matrice d'expression va être attribué au groupe qui contient le centroïde dont il est le plus proche. Le centroïde de chaque groupe va ensuite être calculé de manière itérative jusqu'à l'obtention d'une convergence, c'est à dire une stabilisation des objets dans un groupe donné. Ces méthodes se basant uniquement sur le jeu de données d'entrée pour regrouper les variables peuvent être considérées comme des méthodes non-supervisées. Une des limites du partitionnement est subséquente au fait qu'un gène va être attribué à un cluster unique. Or, ce gène participe potentiellement à des processus biologiques différents et avec des gènes distincts. Pour un jeu de données comportant une large variété de conditions biologiques, un gène devrait donc pouvoir être théoriquement attribué à différents clusters, à la fois reflétant ses différents rôles et partenaires biologiques. De plus, il est toujours difficile de fixer un K optimal.

Des méthodes sophistiquées d'apprentissage automatisé peuvent également être utilisées pour établir des relations transcriptionnelles entre gènes. Elles consistent en l'utilisation de données validées afin d'optimiser la performance d'un modèle prédictif (Larrañaga et al., 2006; López-Kleine, Leal & López, 2013). Le programme va donc se baser sur des connaissances antérieures pour extraire des caractéristiques du jeu de données, et ainsi améliorer par apprentissage sa capacité à reconnaître des schémas ou des associations particulières dans les données qui lui sont fournies. Ces méthodes sont considérées comme supervisées puisque le



jeu de données de départ va être étiqueté avec des connaissances ultérieures (Ma, Zhang & Wang, 2014). Ces méthodes n'utilisent donc pas les connaissances préalables pour confirmer les interactions mises en évidence, mais pour construire ce dernier. Ainsi Ma et al. par exemple, utilisent des données transcriptomiques de conditions de stress abiotiques chez *A. thaliana* pour construire des réseaux de co-expression, ainsi que des caractéristiques d'expression de gènes connues pour être liées aux réponses aux stress abiotiques, afin d'entraîner un algorithme de *Machine Learning*. Cette approche semble plus efficace qu'une analyse classique de gènes différemment exprimés pour identifier des gènes liés à un processus particulier (Ma, Zhang & Wang, 2014).

Ces méthodes complexes ne permettent pas la construction directe d'un réseau qui nécessite généralement d'établir des corrélations par paires afin de représenter les ponts entre nœuds. Il est néanmoins possible d'utiliser ces méthodes pour construire des réseau en petit monde comme le démontrent Liu et collègues avec l'approche des k-moyennes (Li et al., 2006). Elles peuvent également être employées différemment pour donner plus de robustesse aux réseaux de co-expression construits avec les méthodes évoquées ci-dessus comme cela a été démontré pour le WGCNA plus performant après un étape de partitionnement des gènes par k-moyennes (Botía et al., 2017). Par ailleurs, elles se révèlent utiles dans l'analyse de larges réseaux afin de mettre en évidence des gènes densément connectés (modules) (Mutwil et al., 2010).

Ces différentes approches évoquées ci-dessus sont loin de constituer un catalogue exhaustif et représentent un échantillon des possibilités pour établir la co-expression entre gènes, accessibles à ce jour. Un grand nombre d'études ont comparé différentes distances, mais bien souvent sur des petits jeux de données ou des données simulées (Liesecke et al., 2018). Face à ce large choix de méthodes disponibles se posent plusieurs questions. Quelle mesure de co-expression devons-nous choisir en fonction de l'objectif de l'étude et donc de la question biologique à laquelle nous essayons de répondre ? Kumari et al. démontrent en effet que différentes mesures de corrélation ont des performances différentes en fonction de la catégorie fonctionnelle des gènes considérés (décrire un peu plus?) (Kumari et al., 2012). De plus, la nécessité d'avoir une vision globale du génome va impliquer de prendre en considération le



temps de calcul pour l'établissement de la corrélation. Mais également, s'il existe une mesure plus adaptée qu'une autre au jeu de données de départ (nature et pré-traitement des données, nombre d'échantillons etc.) ? Enfin, y-a-t-il une mesure particulièrement robuste quelque soit le jeu de données utilisé (« *one size fits all* »).

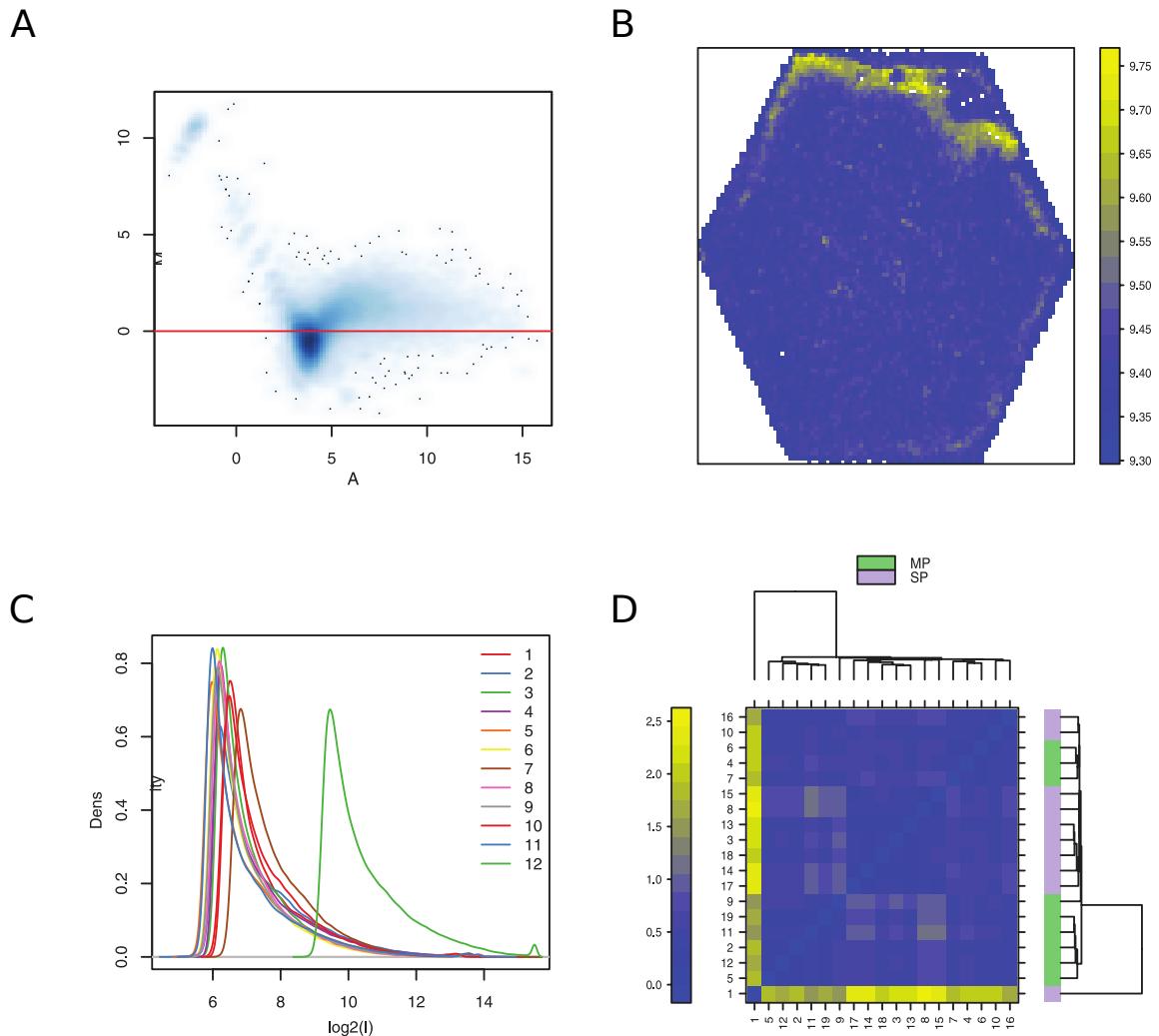
## 2.3. Caractéristiques du jeu de données initial

### 2.3.1. Technologie utilisée pour la génération des données : RNA-seq et microarray

Comme nous l'avons abordé précédemment, la transcriptomique à haut débit réalisée par microarray ou RNA-seq génère des données de nature différente selon la technique utilisée: le premier repose sur les propriétés d'hybridation des acides nucléiques et utilise des intensités de fluorescence alors que le deuxième mesure des comptages de lecture issus d'un séquençage par synthèse afin d'estimer l'abondance de l'expression de gènes.

### 2.3.2. Pré-traitement des données

La première étape dans l'utilisation de données de quelque nature qu'elles soient, est de s'assurer de leur qualité et d'en écarter celles pouvant être considérées comme aberrantes (*outliers*) afin de ne pas fausser les analyses subséquentes. Chaque étape expérimentale du processus de génération de données est susceptible de générer des biais (génération des banques, hybridation des sondes, fragmentation, concentration d'ADNc dans les puits, lecture de fluorescence *etc.*) qui induisent de la variabilité technique au sein des échantillons et pouvant altérer la mesure de la variabilité biologique que l'on cherche à discerner. Un biais particulier pouvant conduire à des erreurs d'interprétation, pas seulement pour les technologies à haut débit, mais pour toute expérience biologique, et notamment observé lors de la combinaison d'échantillons issus de différents études est le *batch effect*. Cet effet est dû au fait que les mesures sont toujours affectées par les conditions du laboratoire, les lots de

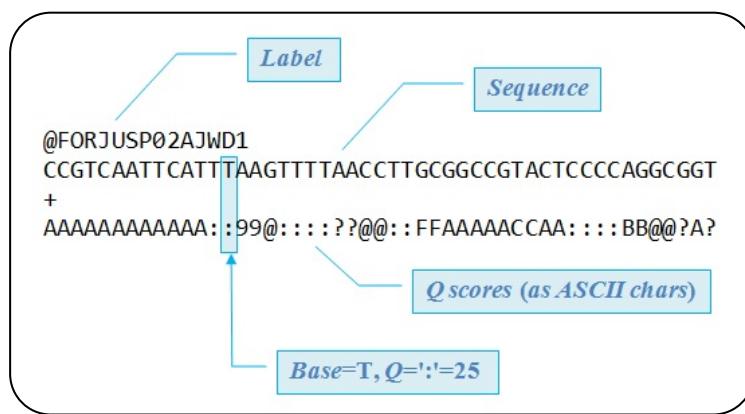


**Figure 13: Contrôle qualité de données microarray avec le package R 'arrayQualityMetrics'** (extrait de Kauffmann et al., 2008). A. MA-plot d'une puce Agilent ( $M=\log$  des intensités entre 2 conditions;  $A=\log$  moyenne de leurs logs) permettant d'évaluer la dépendance entre niveaux d'intensités et distribution des ratios. On s'attend à ce que la majorité des gènes ne présente pas de changement d'expression entre 2 conditions différentes; ici les valeurs de  $M$  ne sont pas centrées autour de 0, indiquant qu'une normalisation des données est nécessaire. B. Distribution spatiale du bruit de fond des canaux verts d'une puce Illumina. On peut constater une distribution anormale de fortes intensités (en jaune) pouvant correspondre à des problèmes techniques au niveau de la lame (bulles d'air, défaut d'impression par exemple). C. Représentation de la densité des valeurs logs d'intensité de fluorescence entre différentes lames (variabilité inter-lames) Affymetrix issues d'un même projet. On peut constater que la lame n°12 (en vert clair) présente un décalage dans la distribution. D. Carte de densité des distances entre lames, calculées comme la différence des moyennes absolues des valeurs de  $M$ . La lame 1 est considérée comme un outlier.

réactifs ainsi que les personnes effectuant les expériences et la préparation des échantillons. Le batch effect reflète donc une certaine manière de réaliser les expériences et peut rendre difficile la comparaison de résultats. Il est également totalement lié à la variabilité biologique des organismes vivants. Il se traduit par le fait qu'un résultat obtenu dans des conditions données et à un temps  $t$  pourrait varier de manière importante s'il est obtenu à un temps  $t+1$ . Cet effet peut être quantifié en utilisant une Analyse en Composantes principales (ACP) ou des méthodes de partitionnement hiérarchique incluant des variables comme le protocole utilisé ou la date de l'expérience (Leek et al., 2010). Le principe de base implique que beaucoup de gènes ont une expression basale peu affectée par la nature de l'échantillon. Ainsi, il est attendu que des lames de microarrays ou des runs de RNA-seq soient assez comparables dans la manière dont les données sont distribuées. Partant de ce postulat, plusieurs méthodes ont été développées pour contrôler la nature ces données.

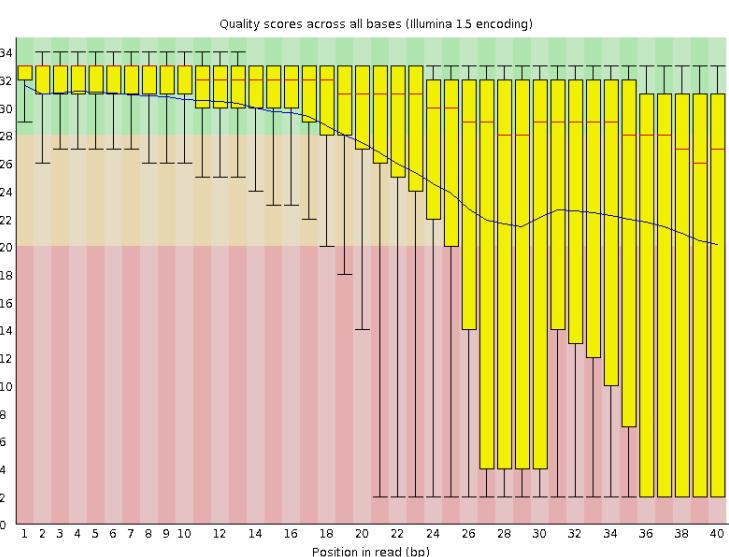
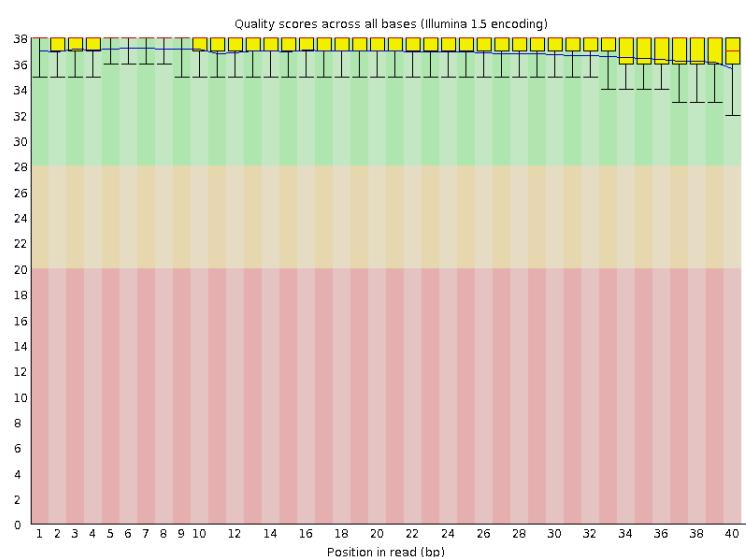
Divers paquets R sont disponibles pour s'assurer d'une qualité suffisante des échantillons traités. Concernant les données Microarray, les paquets peuvent être spécifique à la plateforme Illumina ou Affymetrix (Gautier et al., 2004; Ritchie et al., 2011) ou aux arrays à deux couleurs par exemple (Buness et al., 2005). En 2009, Kauffmann et al. introduisent le paquet R « arrayQualityMetrics » permettant l'évaluation de la qualité quelque soit la plateforme ou la lame utilisée (Kauffmann, Gentleman & Huber, 2009) (**Figure 13**). Cette dernière est évaluée au niveau d'une même lame par la dépendance des niveaux d'intensité et la distribution des ratios. Dans ce paquet, plusieurs évaluations sont proposées telles que la répartition des quantiles, les ACP ou bien la comparaison à une fonction de distribution cumulative empirique par un test de Kolmogorov-Smirnov. A chaque fois, l'objectif est d'observer si certains échantillons présentent une distribution des signaux différents des autres et qui pourrait révéler un problème technique. L'homogénéité peut également être mesurée entre différentes lames permettant de s'assurer de la comparabilité des intensités inter-lames.

Avec l'essor des technologies de séquençage, des méthodes d'évaluation spécifiques au RNA-seq ont également été développées. Différents programmes contrôlant la qualité des lectures sont disponibles : RSeQC, (un programme écrit en C et en Python) ou RNA-SeQC (une

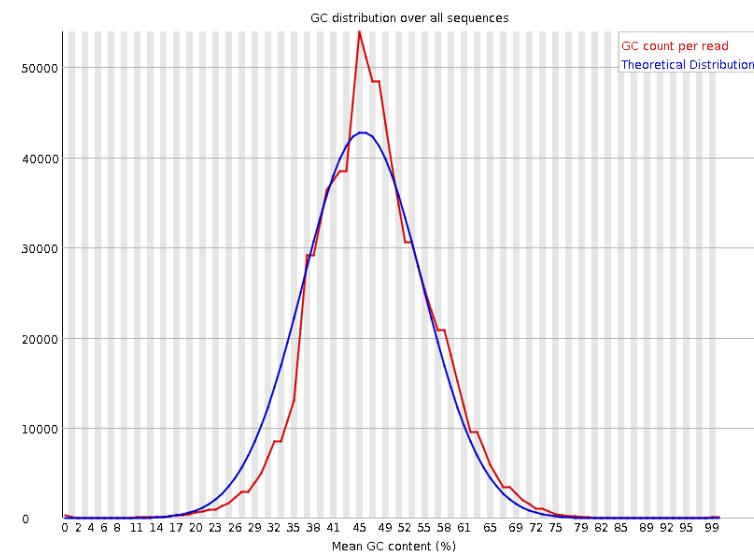
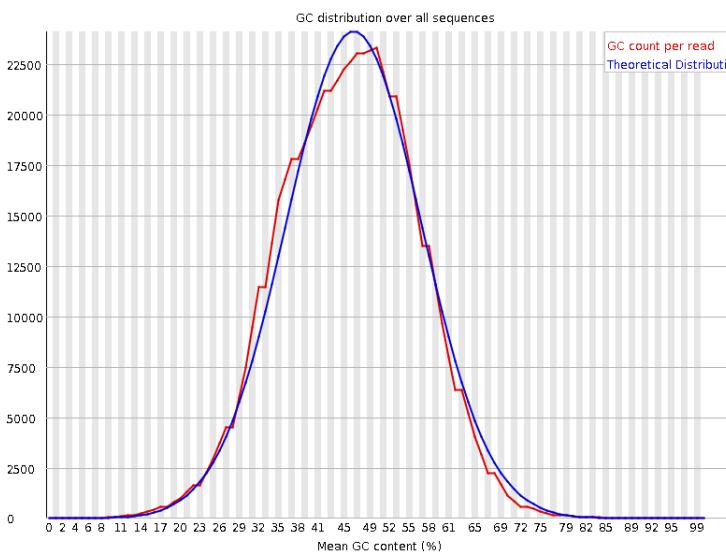


**Figure 14: Extrait de fichier de sortie RNA-seq au format FastQ.** La première ligne contient un '@' suivi du nom du read et éventuellement des spécifications sur le séquençage, suivi de la séquence brute accompagnée d'un score de qualité (Q-score) pour chaque base séquencée et codé en ASCII. (d'après <http://drdk.me/ngs.html>).

A



B



**Figure 15: Contrôle qualité de données RNA-seq.** A. Extrait d'un rapport d'analyse qualité de deux runs Illumina effectué avec le logiciel FASTQC (Andrews, 2010). A. Distribution des Q-scores pour un read de 40 pb. On peut constater que la séquence de gauche présente un excellent score sur toute la longueur de la lecture contrairement à celle de droite. B. Analyse du contenu en GC sur les mêmes runs, on peut constater que le contenu en GC correspond à la distribution théorique contrairement à celle de droite, ce qui peut indiquer des erreurs dans le séquençage. Pour les deux échantillons, les rapports plus détaillés considérant d'autres critères évaluables sont disponibles : [www.bioinformatics.babraham.ac.uk/projects/fastqc/](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

plateforme indépendante en Java) permettent d'évaluer par exemple des statistiques de l'alignement avec la référence comme la couverture de chaque position nucléique, l'évaluation des intervalles non séquencés, la détection de jonctions d'épissage par comparaison avec un modèle de référence ainsi que des biais sur le séquençage générés par le contenu en GC etc. (DeLuca et al., 2012; Wang, Wang & Li, 2012) (**Figures 14 et 15**). La mesure de l'expression de certains gènes de référence permet globalement une bonne évaluation de la pertinence d'un *run* de RNA-seq (Levin et al., 2010).

Ce contrôle qualité des données acquises peut être fait avant ou après la normalisation des données. Dans le second cas, elle permet de s'assurer que la normalisation a permis une réduction du bruit de fond.

### **2.3.3. Normalisation des données**

L'objectif de la normalisation est double : elle permet de réduire le bruit dû aux biais techniques comme vu précédemment, mais également d'assurer une meilleure comparabilité entre échantillons.

#### **2.3.3.1. Normalisation des données Microarray**

Compte tenu de la relative ancienneté de cette technologie, une grande quantité de méthodes de normalisation est disponible pour ce type de données. Elle peut s'effectuer au sein d'une même lame ; dans ce cas, l'enjeu principal est de réduire les biais induits par l'efficacité non constante du marquage au fluorophore et de l'hybridation, ainsi que du scan conduisant à des variations non biologiques de l'intensité des spots selon leur emplacement sur la lame (Smyth & Speed, 2003). Lorsque l'étude combine différentes lames, une normalisation supplémentaire entre ces dernières est nécessaire afin de réduire l'effet batch. Pour faciliter ces étapes de pré-traitement et la correction du bruit de fond, certaines lames incorporent des témoins positifs (sondes comportant les séquences exoniques de gènes de ménage ayant une



expression constitutive et utilisable comme référence ou correspondant à des séquences absentes de l'organisme d'intérêt, lesquelles étant ajoutées en quantité connues dans l'échantillon, c'est le spike-in) ainsi que des témoins négatifs (spot « vide », sondes comportants les séquences introniques de gènes de ménage). Dans le cas des puces Affymetrix, des sondes complémentaires des cibles à un nucléotide près sont également présentes pour mesurer un bruit de fond de manière très sensible pour chaque hybridation possible, tout en s'assurant de leur spécificité.

Le paquet ‘affy’ pour R est largement utilisé pour l’analyse de microarrays. La fonction ‘justRMA’ permet de transformer les fichiers de données brutes au format .CEL en valeurs d’expression corrigées par rapport au bruit de fond et normalisées. La méthode de normalisation utilisée repose sur la normalisation par quantiles (Bolstad et al., 2003) qui permet aux intensités d’avoir une distribution empirique semblable entre lames et canaux (normalisation entre lames). Le paquet R ‘limma’ permet une normalisation intra-lame (fonction ‘normalizeWithinArrays’) en utilisant différentes méthodes, celle par défaut étant la méthode de « print-tip loess » (Ritchie et al., 2011). Elle s’appuie sur la construction des microarrays de type oligonucléotide qui se fait par zones sur la lame et applique une régression de type Loess par zone pour corriger des variations spatiales dans l’hybridation. La fonction ‘normalizeBetweenArrays’ permet, comme son nom l’indique, de normaliser les données de différentes lames en utilisant également la méthode des quantiles par défaut. Les données qui pourront ensuite être traitées dans l’analyse sont généralement transformées en valeurs logarithmiques, permettant une stabilisation de la variance et une distribution des valeurs vers une loi normale facilitant les analyses statistiques.

### ***2.3.3.2. Normalisation des données RNA-seq***

Les données RNA-seq brutes comportent également des artefacts techniques (Hitzemann et al., 2013). La plupart des comptages sont produits par une faible quantité (< 10% des comptages totaux) de gènes très fortement exprimés, alors que beaucoup de gènes d’intérêt peuvent être exprimés de manière beaucoup plus faible, ce qui peut rendre leur détection



difficile lors d'une étude de gènes différentiellement exprimés. La longueur des gènes peut également apporter un biais car les transcrits longs sont susceptibles de générer plus de fragments de par leur longueur, et être par conséquent artificiellement plus représentés au niveau du comptage de lectures. De plus, une lecture peut être alignée à différentes localisations de la référence (répétitions de séquences, gènes paralogues présentant une forte similarité de séquence). Des algorithmes récemment développés sont supposés limiter ce risque en maximisant la probabilité d'une association correcte d'une lecture avec son fragment d'origine. Ce genre d'approche a été développé notamment en utilisant une boucle d'espérance-maximisation (Li et al., 2010). Pour effectuer le comptage des lectures, il faut également tenir compte de paramètres utilisés lors du séquençage comme la profondeur de ce dernier.

Il convient donc a priori de ne pas utiliser les comptages de lectures bruts pour l'estimation de l'abondance des transcrits. Ces comptages peuvent être exprimés en RPKM (*Reads per Kilobase Million*) pour les *runs* en *single-end* ou en FPKM (*Fragments per Kilobase Million*) pour les runs en *paired-end* prenant en compte que deux lectures peuvent être alignées à un même fragment et sont comptabilisées ensemble (Mortazavi et al., 2008; Wang, Gerstein & Snyder, 2009).

- **RPKM (Reads Per Kilobase Million)**

$$\text{Facteur de scaling} = \frac{\text{Nombre de lectures}}{1\ 000\ 000}$$

$$\text{Lectures par million (RPM)} = \frac{\text{Comptage de lectures}}{\text{Facteur de scaling}}$$

Normalise pour la profondeur de séquençage

$$\text{Lectures par kilobase par million (RPKM)} = \frac{\text{RPM}}{\text{Longueur de du gène en kilobase}}$$

Une alternative sont les TPM (Transcripts per million) (Patro et al., 2017). La somme des TPM dans chaque échantillon étant constante, cette mesure facilite la comparaison entre ces derniers (Wagner, Kin & Lynch, 2012).



- **TPM (Transcripts Per Million)**

$$RPK(\text{lectures par kilobase}) = \frac{\text{Comptage de lectures}}{\text{Longueur de chaque gène en kilobases}}$$

$$\text{Facteur de scaling} = \frac{\text{Total des RPK}}{1\ 000\ 000}$$

$$\text{Transcrits par million (TPM)} = \frac{RPK}{\text{Facteur de scaling}}$$

Ces mesures d'expression améliorent la comparabilité entre gènes et entre échantillons (Mortazavi et al., 2008). Afin d'accroître cette comparabilité, les valeurs d'expression peuvent ensuite être normalisées par une transformation stabilisant la variance (Variance-Stabilizing Transformation ou VST) (Anders & Huber, 2010; Giorgi, Del Fabbro & Licausi, 2013). Toutes ces méthodes de normalisation peuvent être appliquées avec le paquet 'DESeq2' pour R (Love, Huber & Anders, 2014).

En plus de ces normalisations plateforme-spécifiques, des normalisations plus traditionnelles peuvent aussi être appliquées. Il s'agit de la transformation en log2 et du centrage-réduction (*scaling*). La première donne un poids moins important aux valeurs très fortes, mais n'impacte pas les corrélations. Pour le *scaling*, les valeurs d'un gène pour un échantillon sont soustraîtes à la moyenne des expressions de tous les gènes dans cet échantillon, et cette valeur centrée est réduite en la divisant par l'écart-type des expressions. L'intérêt de cette transformation est de donner un poids équivalent à tous les échantillons, au risque de masquer des différences d'expression absolues qui sont biologiquement significatives. La normalisation peut aussi s'appliquer sur les gènes pour mettre en évidence des profils d'expression.

Quelque soit le type de données, la normalisation est souvent une étape nécessaire et va avoir un impact sur les analyses subséquentes. Il a pu être démontré, en comparant diverses méthodes de normalisation appliquées à des données Microarray, qu'elles présentent des



performances différentes pour déterminer la corrélation entre gènes (Lim et al., 2007; Usadel et al., 2009). Cependant, Obayashi & Kinoshita (Obayashi & Kinoshita, 2009), met en évidence qu'utiliser des valeurs de PCC classés ou les rangs mutuels (voir Partie 1) pour établir la corrélation entre paires de gènes permet de réduire la sensibilité à la méthode de normalisation utilisée. Par ailleurs, les données RNA-seq transformées par la VST les rendent plus comparable aux données microarray (Giorgi, Del Fabbro & Licausi, 2013). Comme mentionné dans la partie précédente, il existe de nombreuses manières de normaliser un jeu de données et le choix de la méthode de normalisation est donc un paramètre important à prendre en compte en fonction de l'étude envisagée.

### 2.3.4. Taille du jeu de données

L'impact de la taille du jeu de données sur la robustesse des réseaux de co-expression construits à partir de ce dernier a été relativement peu évalué comparé à d'autres paramètres, comme la mesure de corrélation (D'haeseleer, Liang & Somogyi, 2000; Priness, Maimon & Ben-Gal, 2007) ou le pré-traitement du jeu de données (Giorgi, Del Fabbro & Licausi, 2013). Empiriquement, nous pouvons imaginer qu'un grand nombre de conditions permettant de couvrir des conditions biologiques larges permettent de construire des réseaux de co-expression plus robustes, bien que peut-être au détriment de la détection de relations spécifiques et transitoires.

Les corrélations établies avec un plus grand nombre de conditions expérimentales sont plus susceptibles d'être statistiquement significatives (Reverter & Chan, 2008). Concernant la distribution des coefficients de corrélation, Reverter et al montrent à partir de données simulées pour 500 gènes qu'avec un nombre croissant de conditions, les valeurs de coefficients de corrélation ont tendance à se centrer autour de 0 nécessitant une adaptation du seuil de significativité en fonction de la taille du jeu de données. A partir de jeux de données réels, combinant différents projets de recherche, et de données simulées, Altay et al (2012) évaluent l'influence du nombre de conditions microarray (entre 4 et 2000 conditions) sur la construction de réseaux de co-expression sur l'information mutuelle à grande échelle chez



*E.coli* (Altay, 2012). Ils démontrent que pour ce type de réseau, un nombre d'échantillons compris entre environ 64 et 256 semble suffire pour construire un réseau robuste, et qu'accroître le nombre d'échantillons au-delà n'améliore pas la performance de manière notable. Une autre étude, menée également sur des données microarray de *E.coli* et incluant 725 conditions, évalue également l'impact de ce paramètre sur les réseaux de co-expression basés (Cosgrove, Gardner & Kolaczyk, 2010). Les chercheurs arrivent à la conclusion qu'un grand nombre d'échantillons n'accroît pas nécessairement la qualité du réseau sous-jacent. Ce résultat serait lié à la dépendance qui existe entre les différents échantillons (redondances d'expériences, de tissus et de thématiques). Gibson et al évaluent des réseaux de co-expression basés sur des données microarray humaines (2000 conditions), de riz (1360 conditions) ainsi que de *S.cerevisiae* (1701 conditions) et utilisant les PCC comme mesure de corrélation et associés à algorithme de *Random Matrix Theory* afin de fixer le seuil de significativité (Gibson et al., 2013). Les réseaux construits à partir de matrices d'expression contenant une sélection aléatoire de 25 %, 50 % et 75 % des jeux de données initiaux ont été comparés aux réseau élaborés à partir de la totalité des échantillons en fonction des ponts entre gènes partagés. Les auteurs montrent qu'un nombre d'échantillons compris entre 300 et 500 semble suffisant pour construire des réseaux robustes. Au contraire, Ballouz et al (2015) démontrent que les réseaux de co-expression qu'ils soient construits à partir de données microarray ou RNA-seq gagnent en robustesse en augmentant le nombre d'échantillons considérés (Ballouz, Verleyen & Gillis, 2015). Ces derniers utilisent cependant une technique d'agrégation de réseaux projet-spécifique plutôt que des distances calculées sur l'ensemble des conditions combinées.

### 2.3.5. Autres paramètres

De nombreuses autres caractéristiques du jeu de données peuvent influencer la topologie du réseau de co-expression final. Inéluctablement, les conditions expérimentales des échantillons utilisés gouvernent qualitativement et quantitativement les associations potentiellement détectables entre gènes. Comme évoqué précédemment, un grand nombre de conditions expérimentales est susceptible de couvrir une large gamme de conditions biologiques, donnant ainsi un aperçu global des relations entre les gènes étudiés. Un jeu de données



construit sans *a priori* avec les données d'expression issues de différents études est cependant susceptible de contenir des redondances et des sur-représentations de conditions expérimentales au niveau d'un tissu spécifique ou d'un stress par exemple. De plus, le grand nombre de témoins ainsi intégrés à l'étude pourrait masquer des interactions occurrant dans des conditions plus spécifiques. Feltus et al rapportent que la qualité de leurs réseaux de co-expression d'*Arabidopsis thaliana*, basées sur des données microarray et des corrélations établies par le coefficient de corrélation de Pearson, peut être augmentée en pré-clusterisant les échantillons de la table d'expression par k-moyennes, permettant notamment une réduction du bruit de fond (Feltus et al., 2013). Le choix des conditions considérées est inéluctablement lié à la nature de la question biologique posée. En effet, dans le cadre d'une question portant sur un phénomène physiologique particulier comme un stress ou une pathologie, il pourrait convenir d'utiliser des données thématiques permettant de mettre en évidence les relations occurring dans cette situation précise. Des réseaux dédiés à l'étude de l'immunité chez les plantes ont ainsi été construits à plusieurs reprises (Leal, López & López-Kleine, 2014; Dong et al., 2015) ou bien sur les stress abiotiques (Priest et al., 2014; Des Marais et al., 2017) ou encore purement développemental (Ruiz-Sola et al., 2016).

Il est également possible de filtrer des variables (gènes) du jeu de données qui pourraient impacter de manière négative les analyses subséquentes sans apporter d'information valable. Ceci est notamment le cas pour les gènes présentant des niveaux d'expression très faibles ou nuls. Ils sont en effet susceptibles, dans le cas d'une analyse de gènes différentiellement exprimés par exemple, de biaiser les tests statistiques conduisant une augmentation de faux positifs et/ou de faux négatifs. En ce qui concerne la construction de réseaux, des gènes peu exprimés peuvent tout de même présenter des dépendances significatives avec d'autres plus exprimés et donc être retenus dans le réseau final. Ils peuvent potentiellement être considérés comme du bruit de fond dans la construction du réseau, comme il l'a été montré sur des réseaux utilisant des données obtenues sur l'Homme (Ballouz, Verleyen & Gillis, 2015).

Une diminution du nombre de gènes peut également être envisagée par exemple en se restreignant à des gènes différentiellement exprimés afin de réduire la taille physique du jeu de données et par conséquent les temps de calculs *in silico*. En effet, dans le cadre d'études ciblées, écarter dès le début de l'analyse les gènes ne montrant aucune corrélation avec les



**Figure 16: Spécificité et sensibilité d'un test statistique.** La sensibilité mesure la capacité d'un test à donner un résultat positif lorsque l'hypothèse initiale est vérifiée ("vrai positif"), alors que la spécificité mesure sa capacité à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée ("vrai négatif"). Ces caractéristiques permettent de conclure sur les capacités prédictives d'une évaluation. Dans le cadre d'un réseau de co-expression, le vrai positif correspond à des liens entre gènes correspondants à des associations biologiquement validées, un faux positif à une association non vérifiée (car non existante ou non décrite présentement). Au contraire, un vrai négatif correspond à une association non prédictive et effectivement non décrite alors qu'un faux négatif représente une association effective non détectée dans le réseau.

gènes candidats, ne devrait pas influencer les paysages transcriptionnels de ces derniers ou alors réduire un certain bruit de fond comme décrit ci-dessus. La taille du dataset peut aussi être simplement réduite aux transcrits présentant une expression minimale fixée empiriquement (par exemple sur la somme des échantillons).

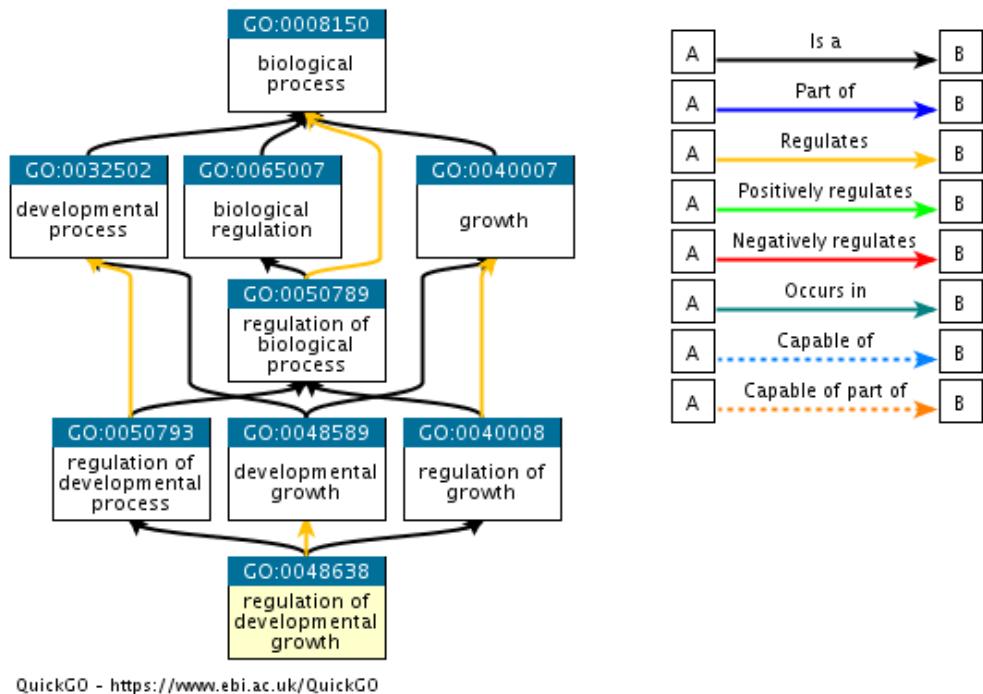
Il apparaît ainsi que la topologie du réseau de co-expression, et par extension sa qualité et l'information biologique qui peut en être extraite, est fortement influencée par différents paramètres et leurs dépendances : choix du type de données (conditions, technologie, qualité, normalisation), taille du jeu de données, méthode utilisée pour déterminer la co-expression ainsi que les seuils de significativité fixés etc. Il est donc primordial de déterminer à chaque étape de la construction de réseau, les paramètres optimaux permettant d'obtenir un réseau le plus robuste possible.

### 3. Validation et analyse des réseaux

#### 3.1. Validation

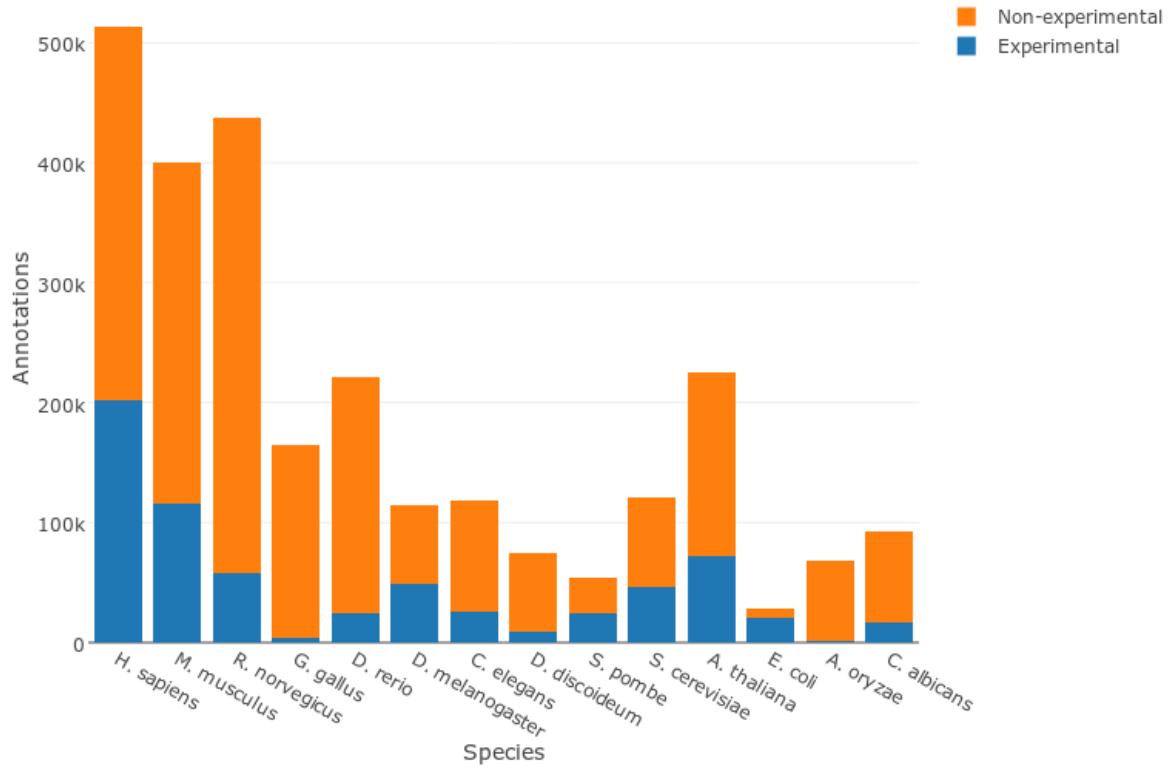
Nous avons pu aborder précédemment le fait que la construction *in silico*, d'un réseau de co-expression est un procédé complexe mettant en jeu un grand nombre de paramètres influençant la topologie du réseau final. L'objectif principal est d'optimiser la combinaison de tous ces paramètres afin d'obtenir un réseau robuste, *i.e.* qui contient un maximum d'associations significatives correspondant à des réalités biologiques (vrais positifs), tout en minimisant celles qui correspondent à des faux positifs et des faux négatifs (**Figure 16**). Différentes approches confrontant les réseaux à des connaissances préalables acquises ont été développées. Cette confrontation permet d'évaluer, par l'intermédiaire de la sensibilité et de la spécificité des approches utilisées, la prédictivité du réseau construit.

A



B

Experimental annotations by species



**Figure 17: Annotation "Gene Ontology" (GO).** A. Diagramme ancestral pour le terme GO:0048638 "Régulation de la croissance développementale". Les termes les plus généraux décrivant le processus se trouvent en amont du graphique et se précisent progressivement à chaque niveau descendant. La couleur des flèches indique le type d'interaction entre différents termes GO. B. Nombre d'annotations disponibles pour différentes espèces modèles. Source: <http://www.geneontology.org/>

### 3.1.1. Données disponibles

En 1998, de la volonté de créer une représentation homogénéisée de la connaissance grandissante expliquant comment les gènes codent des fonctions biologiques au niveau moléculaire, cellulaire et tissulaire au sein de systèmes vivants, naît le Consortium international du *Gene Ontology Project* (<http://www.geneontology.org>). L'objectif de ce projet bio-informatique a été de créer une nomenclature globale, c'est à dire commune à toutes les espèces, décrivant l'annotation fonctionnelle de gènes. Il a débuté sur les bases de données de trois espèces modèles, la drosophile, la levure et la souris, et couvre à ce jour l'essentiel des organismes modèles microbiaux, végétaux et animaux, tout en étant applicable à n'importe quel organisme. Les termes de Gene Ontology ou termes GO fournissent une ontologie de termes définis décrivant les propriétés de gènes dans son contexte cellulaire. Ces annotations sont classées en trois catégories :

- les composants cellulaires décrivant la localisation subcellulaire du produit du gène
- les fonctions moléculaires comme les réactions catalysées par exemple
- les processus biologiques impliquant le produit du gène

Ces termes GO peuvent être très généraux, comme « activité catalytique » ou très précis, décrivant une activité catalytique spécifique, par exemple. Les termes présentent donc des relations étroites et hiérarchiques permettant leur représentation sous la forme d'un graphe orienté acyclique (**Figure 17A**). Les annotations de gènes peuvent être inférées expérimentalement, phylogénétiquement ou par des études *in silico*. En 2014, la base de données regroupe un total de 41 775 termes GO décrivant 53 042 843 produits de gènes chez 461 573 différentes espèces (The Gene Ontology Consortium, 2015), (**Figure 17B**), voir [http://amigo.geneontology.org/amigo/base\\_statistics](http://amigo.geneontology.org/amigo/base_statistics) pour une description détaillée du contenu de la base de données).

Beaucoup d'autres annotations fonctionnelles sont librement disponibles comme les termes d'ontologies MAPMAN, l'annotation par domaines CDD, PFAM, SMART, PROSITE, ou InterPro (Finn et al., 2017). Ces domaines représentent des signatures qui peuvent renseigner



sur la fonction potentielle d'une protéine donnée à partir de sa séquence peptidique. Il existe certaines bases de données regroupant l'ensemble de ces annotations fonctionnelles, tel qu'Uniprot qui bénéficie d'une curation manuelle très poussée.

Différentes bases de données décrivent des voies biologiques et les interactions moléculaires au sein de celles-ci pour un large éventail d'espèces. Le KEGG (Kyoto Encyclopedia of Genes and Genomes, <http://www.kegg.jp>) (Kanehisa et al., 2016) a été créée en 1995 afin de faciliter l'interprétation biologique à partir de données de séquences génomiques, et est aujourd'hui une des bases de données biologiques les plus utilisées. Le but principal a été de fournir des collections de gènes regroupés sous un même terme d'orthologie (accessions KO) afin d'établir des relations entre ces termes et ainsi de leurs fonctionnalités. En 2015, plus de 4000 génomes complètement séquencés ont été annotés avec la nomenclature du KEGG. Un des inconvénients de cette base de donnée est qu'elle utilise souvent sa propre nomenclature pour les identifiants de gènes et qu'il peut être délicat de faire correspondre des locus tags utilisés pour l'annotation des transcrits et des gènes. Cependant, KEGG offre aussi un pipeline d'annotation en ligne (KAAS). Une autre base de données créée en 2005 et recensant des informations sur les voies de signalisation et métaboliques d'une vingtaine d'espèces est la base Reactome ([www.reactome.org](http://www.reactome.org)), fonctionnant sur un principe de peer review assurant que seules des données décrites dans la littérature soient intégrées à la base de données (Joshi-Tope et al., 2005; Fabregat et al., 2018).

Toutes ces bases de données fournissent aussi bien une interface en ligne afin d'investiguer les relations ou fonctions de gènes candidats, qu'une possibilité de téléchargement en vue d'une intégration dans les analyses bio-informatiques.

Il est également possible de s'appuyer sur des données d'interactions protéine-protéine afin de valider les relations mises en évidence dans le réseau. Différentes bases de données sont disponibles, comme la base STRING (<http://string-db.org>) comportant 9 643 763 de protéines de 2 031 organismes et basée sur des interactions directes (physiques) et indirectes (fonctionnelles) connues à partir de données à haut débit, des études de co-expression, du *textmining* automatisé et des connaissances d'autres bases de données ou prédites à partir du contexte génomique (Szklarczyk et al., 2017). En fonction des indices regroupés sur une



interaction potentielle, un score est attribué pour en quantifier la robustesse, ce qui permet de filtrer facilement les meilleures associations de cette base d’interactions.

Une autre manière de confirmer la véracité des interactions mises en évidence dans un réseau sont bien sûr la bibliographie (*textmining*), et à un autre niveau la validation fonctionnelle de la co-expression des gènes identifiés par des techniques de laboratoire comme la RT-qPCR, l’utilisation de mutants KO, ou lorsqu’une interaction physique est suspectée des techniques comme le BIFC ou le double hybride en levure.

### **3.1.2. Validation des relations mises en évidence dans le réseau**

Une première approche intuitive, ne nécessitant pas l’utilisation de statistiques poussées, est de tirer profit de ces annotations disponibles et d’évaluer les termes GO ou les interactions protéine-protéine connues entre gènes reliés correctement mises en évidence au sein du réseau de co-expression (Obayashi & Kinoshita, 2009; Giorgi, Del Fabbro & Licausi, 2013). Suivant le principe de culpabilité par association, ces gènes sont susceptibles d’intervenir dans des processus communs et donc de porter une annotation fonctionnelle commune. La fiabilité de cette évaluation sera fortement dépendante de la qualité de l’annotation (Lee et al., 2004), ce qui peut poser un problème sur l’évaluation de réseaux de co-expression pour des espèces non-modèles. Notamment pour les interactions protéine-protéine, il convient de noter premièrement qu’une interaction indirecte peut exister au niveau de la régulation de l’expression des gènes qui n’est pas nécessairement reflétée par une interaction physique entre leurs produits. De plus, ces interactions sont souvent mises en évidence expérimentalement par des techniques de double hybride en levure qui peuvent présenter un nombre non négligeable d’artefacts car, même si ces interactions sont possibles physiquement, elles ne révèlent pas inéluctablement une réalité biologique, par exemple en raison d’une répartition tissu-spécifique des protéines.

De nombreuses méthodes statistiques sont disponibles pour évaluer la qualité de réseaux. Des modules ou communautés regroupant des gènes densément connectés peuvent généralement être mis en évidence. Cette forte interconnexion devrait *a priori* exprimer le partage de



fonctions communes entre ces gènes. Il est donc possible d'analyser si des fonctions sont significativement enrichies dans ces modules, ce qui est aisément réalisable avec un test hypergéométrique (Horan et al., 2008). Une autre stratégie de confirmation des associations mises en évidence largement utilisée repose sur la fonction d'efficacité du récepteur ou plus communément « courbe ROC » (Reciever Operating Characteristic), qui permet de mesurer la sensibilité en fonction de la spécificité d'un test. Cette courbe représente ainsi le nombre de vrais positifs par rapport au nombre de faux positifs prédicts par le test. L'efficacité peut ainsi être mesurée par l'aire sous la courbe (Area Under the Curve ou AUC) ; plus cette aire est grande, plus la probabilité que les interactions identifiées dans le réseau correspondent à une réalité biologique est grande. Comme pour les tests hypergéométriques, les annotations disponibles sont utilisées ici pour estimer le taux de résultats positifs et négatifs (Obayashi et al., 2014; Ballouz, Verleyen & Gillis, 2015).

Des algorithmes plus complexes de machine learning comme le Support Vector Machine (SVM) ou les *random forest* peuvent également être utilisés à cette étape afin de mesurer la performance du réseau, neighbour voting, régression logistique (Ballouz, Verleyen & Gillis, 2015; Verleyen, Ballouz & Gillis, 2015).

## 3.2. Analyse

### 3.2.1. Réseaux globaux et réseaux ciblés

Un réseau de co-expression global (ou *genome-wide*) repose sur la représentation de l'ensemble des transcrits d'un système vivant et de leurs interconnexions. Ces réseaux très denses nécessitent cependant d'importantes ressources informatiques pour leur construction, à partir de matrices d'expression larges et leur analyse, mais permettent de représenter sans *a priori* les liens transcriptionnels mis en évidence dans les conditions étudiées au sein d'un organisme. En raison de leur taille importante, leur visualisation directe n'est généralement pas interprétable. Une manière d'en réduire la complexité est de conduire une approche

$n \times p$  expression matrix

$n$  genes

$p$  samples

$n \times n$  correlation or rank matrix

set of  $g$  **guide genes**

A C

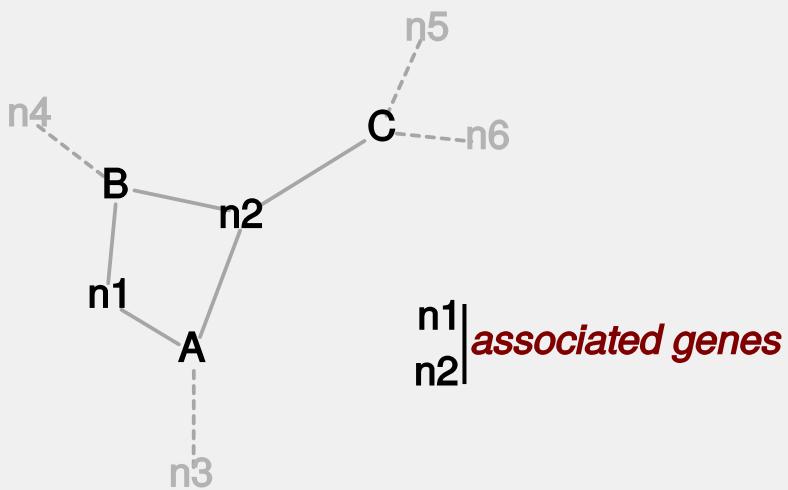
B

extract a  $n \times g$  guide matrix from correlation or rank matrix

define best coexpressed genes  
( $\text{PCC} > x$  or  $\text{rank} < y$ ) with each gene in  $g$

A	B	C
n1	n1	n2
n2	n2	n5
n3	n4	n6

group genes in  $g$  according to overlaps between best coexpressed genes



**Figure 18: Méthodologie pour la construction de réseaux ciblés de type Pathway Level Co-expression (PLC).** A partir d'une matrice d'expression, les corrélations sont établies deux à deux entre l'ensemble des gènes de la matrice initiale afin de construire une matrice de corrélation. Une sous-matrice correspondant aux corrélations de gènes guides avec le reste du génome est extraite. Dans cette matrice, pour chaque gène guide, vont être retenus les gènes présentant des corrélations correspondant à un seuil fixé au préalable permettant ainsi d'établir des listes de gènes fortement co-exprimés pour chaque gène guide. Ces listes vont être croisées et les gènes co-exprimés avec plusieurs gènes guides, formant les gènes associés, permettent de grouper les gènes guides.

ciblée, en interrogeant le voisinage d'une liste de gènes d'intérêt, ou alors d'annotation fonctionnelle de nouveaux gènes en se basant sur les gènes voisins regroupés au sein de modules et dont la fonction a été décrite ultérieurement (Aoki et al., 2007; Freeman et al., 2007; Mao et al., 2009).

La construction de réseaux ciblés vise à extraire une information spécifique du réseau en utilisant comme « appâts » les gènes connus pour être liés au processus étudié. Cette méthode se révèle très utile pour compléter les éléments manquants d'un processus particulier déjà relativement bien décrit, comme une voie moléculaire présentant des étapes non élucidées par exemple. Le paysage transcriptionnel des gènes décrits antérieurement comme intervenant dans le processus va être capturé et suivant le principe de « culpabilité par association », les gènes connectés à ces derniers sont fortement susceptibles d'englober les gènes manquants au sein de la voie. Une méthode intuitive consiste à récupérer les listes de gènes fortement exprimés avec les gènes d'intérêt (appelées « gènes guides »). Ces listes sont ensuite comparées afin de mettre en évidence des gènes co-exprimés partagés entre différents gènes guides : ces « gènes marqueurs » vont constituer les liens entre gènes guides au sein du réseau (**Figure 18**). Cette approche nommée *Pathway Level Coexpression* (PLC) par Wei et al. a été utilisée pour élucider différentes voies métaboliques chez les végétaux (Wei et al., 2006). Ainsi, Lissso et al utilisent la PLC sur des données microarray d'*Arabidopsis thaliana* afin de construire un réseau ciblé basé sur les coefficients de Spearman, et utilisant des gènes connus pour participer dans le métabolisme des brassinostéroïdes comme guides. Plus récemment, cette approche a permis l'étude de différentes voies métaboliques. Ruiz-Sola et al ont pu mettre en évidence une geranylgeranyl diphosphate synthase impliquée dans la synthèse d'isoprenoïdes chez *A. thaliana* en utilisant les gènes de la voie du mévalonate et du méthylérythritol phosphate, ainsi que des gènes codant des enzymes intervenant en aval de ces voies de biosynthèse comme gènes guides pour la construction d'un réseau de co-expression basée sur des données microarray et un calcul de corrélations par PCC (Ruiz-Sola et al., 2016). Guérin et al ont utilisé la PLC sur un jeu de données RNA-seq et en évaluant encore une fois les distances avec le PCC, afin d'élucider la synthèse d'acides gras chez le palmier à huile (*Elaeis guineensis* et *Elaeis oleifera*) (Guerin et al., 2016). Ces deux derniers exemples illustrent bien la puissance et l'applicabilité à des espèces non modèles de ce type d'analyse.



### **3.2.2. Du réseau à l'information biologique : détection de communautés**

La représentation sous forme de graphe facilitant la lecture des interactions, un réseau de petite taille peut être analysé visuellement pour définir la topologie des relations entre les gènes le composant. Or les réseaux biologiques, représentant un système complexe, sont souvent composés d'un grand nombre de nœuds rendant nécessaire de dégager les modules ou communautés formés par les gènes plus densément connectés entre eux qu'au reste du réseau, et correspondant *a priori* à des associations biologiques préférentielles entre ces derniers. Cette détection de communauté est un processus qui peut être complexe dans le cas de larges réseaux et peut faire appel à plusieurs critères basés sur la topologie ou bien la classification fonctionnelle pour estimer ou évaluer la qualité du partitionnement (Horvath & Dong, 2008; Langfelder et al., 2011; Li, Pearl & Jackson, 2015). Lors d'une analyse par des tests d'enrichissement en termes Gene Ontology par exemple, il est possible de relier ces groupes de gènes fortement associés à des processus physiologiques voire de leur attribuer une fonction précise potentielle. La manière dont les sous-réseaux sont constitués est donc critique.

Le paquet R ‘WGCNA’ contient diverses fonctions permettant la détection de modules au sein des réseaux établis. Le partitionnement hiérarchique non supervisé (sans set de gènes de référence) est utilisé par défaut , les branches du dendrogramme correspondant aux modules de gènes qui peuvent être identifiés par des algorithmes comme le *Dynamic Tree Cut* utilisant un processus adaptatif et itératif de décomposition et combinaison de clusters (il commence avec un faible nombre de grands clusters analysés afin d'identifier des motifs de sous-clusters dans quel cas le cluster va être rediscuté de manière récursive) (Langfelder & Horvath, 2008). Le paquet offre d'autres possibilités pour l'identification de communautés, comme le *Automatic block-wise module detection* particulièrement adapté aux données impliquant un grand nombre de gènes car permettant d'économiser des ressources informatiques. Cet algorithme définit en premier lieu des « pré-clusters » appelés blocs en utilisant une variante des k-moyennes. Ces blocs vont être soumis à un partitionnement hiérarchique, les modules représentant les branches du dendrogramme. Afin de synthétiser la détection de ces modules au sein des blocs, une étape automatisée va fusionner les clusters dont les « eigengenes »



(gène représentatif qui permet de résumer les profils d'expression du module (Langfelder & Horvath, 2008) sont fortement corrélés. Enfin, il est également possible de comparer des modules de différents réseaux avec le *Consensus module detection* afin de construire un réseau consensus. Il s'agit de l'un des programmes les plus utilisés en biologie pour construire, analyser et visualiser des réseaux de co-expression (le paquet est cité 2917 fois le 02/08/2018).

Un autre programme disponible en tant que paquet pour R et Python ainsi que sous forme de librairie C, très performant pour la visualisation ainsi que l'analyse de la structure de réseaux, est le paquet ‘igraph’ (Csardi & Nepusz, 2006). Il permet de visualiser un réseau à partir d'une liste de liens entre variables ou d'une matrice d'adjacence et nécessite donc une détermination préalable et indépendante des corrélations entre gènes. Il incorpore différentes méthodes pour déterminer les structures modulaires au sein du réseau. Une fonction efficace est basée sur un score de *edge betweenness* de liens, c'est à dire le nombre de chemins le plus courts entre nœuds passant par ce lien. En effet, des liens connectant des communautés séparées sont susceptibles de présenter des scores élevés puisque tous les chemins les plus courts passant d'une communauté à l'autre doivent passer par ce lien. Si on retire graduellement les liens présentant les scores les plus élevés, on peut mettre en évidence les modules séparés et ainsi révéler la structure sous-jacente du graphe (Girvan & Newman, 2002). Une méthode plus rapide et donc plus adaptée aux réseaux denses est basée sur un algorithme glouton (*greedy*) qui va rechercher nœud par nœud à identifier les communautés, afin d'optimiser la modularité du réseau (Clauset, Newman & Moore, 2004).

Ces modules ou communautés peuvent ensuite être analysés par des tests d'enrichissement en termes Gene Ontology par exemple permettant de relier ces groupes de gènes fortement associés à des processus physiologiques voire de leur attribuer une fonction précise potentielle.

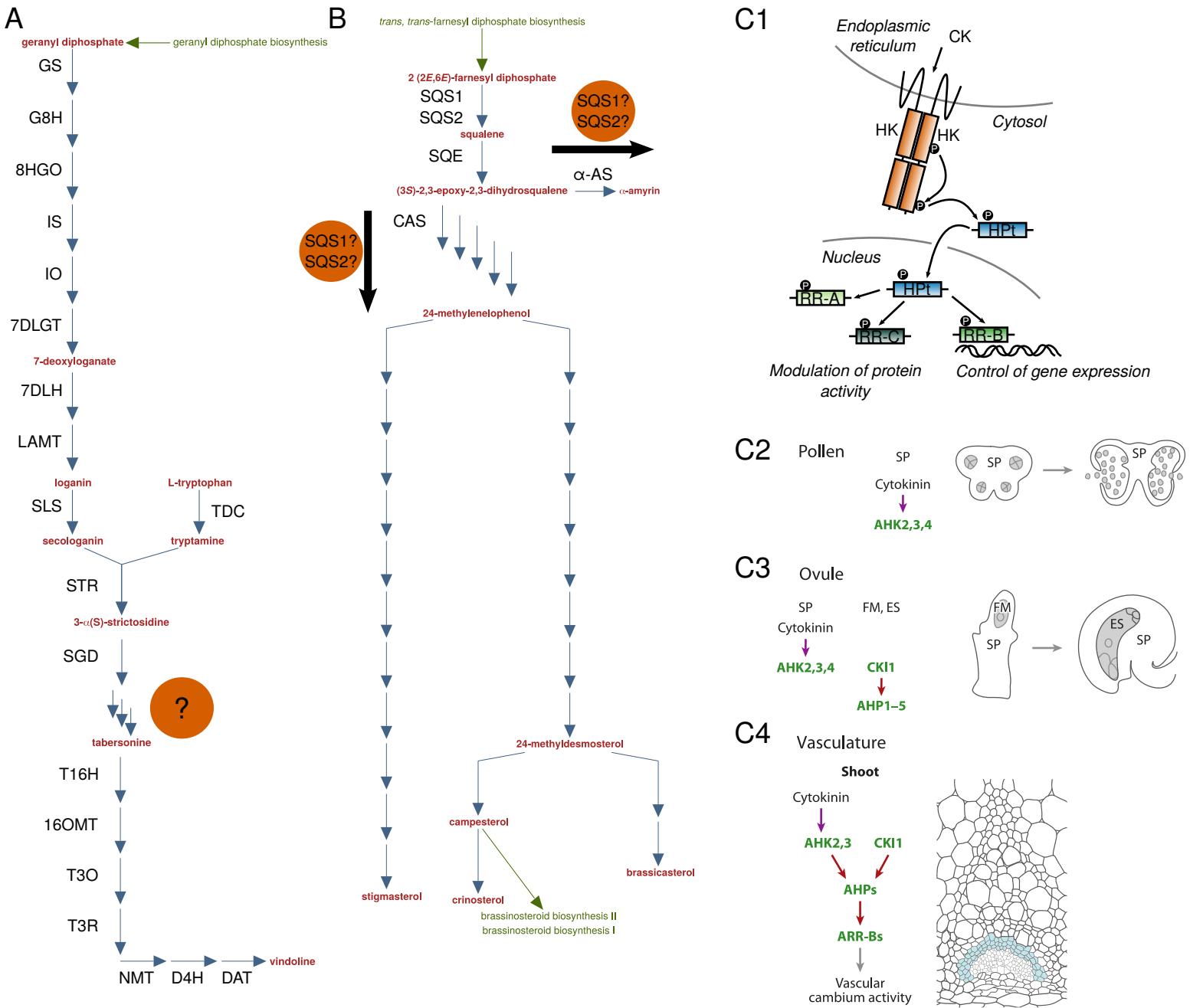
La mise en évidence de ces modules permet d'utiliser la topologie intrinsèque du réseau par un large panel d'algorithmes afin de mettre en exergue des groupes de gènes fortement reliés entre-eux et partageant donc, d'après le principe de « culpabilité par association » une



éventuelle fonction biologique. Il apparaît qu'au delà d'une représentation graphique des relations entre gènes dans un ou des contextes données, le réseau de co-expression permet, par son analyse, d'extraire *in silico* de données biologiques à haut débit des informations biologiques valables. Compte tenu du nombre de paramètres inter-dépendants et des nombreux choix méthodologiques auxquels est confronté le biologiste exploitant ces données, ainsi que l'inexistence d'une méthodologie universelle, tout le défi réside dans l'optimisation de chaque étape du processus de construction de réseau de co-expression en ne perdant pas de vue la question ou l'hypothèse biologique de départ.



# Contexte, méthodologie générale et organisation de la thèse



**Figure 19: Problématiques pour lesquelles une analyse de co-expression pourrait être envisagée.** A, voie de biosynthèse des Alcaloïdes Indoles Monoterpéniques (AIM) chez *Catharanthus roseus*. GS, Geraniol Synthase; G8H, Geraniol 8-Hydroxylase; 8HGO, 8-HydroxyGeraniol Oxidase; IS, Iridoid Synthase; IO, Iridoid Oxidase; 7DLGT, 7-Deoxyloganetic acid; 7DLH, 7-DeoxyLoganic acid-7-Hydroxylase; LAMT, Loganic Acid Methyl Transferase; SLS, Secologanin Synthase; TDC, Tryptophan Decarboxylase; STR, Strictosidine Synthase; SGD; Strictosidine Glucosidase; T16H, Tabersonine 16-Hydroxylase; 16OMT, 16-hydroxytabersonine O-Methyl Transferase; T3O, Tabersonine 3-Oxidase; T3R, Tabersonine 3-Reductase; NMT, N-Methyl Transferase; D4H, Deacetoxyvindoline 4-Hydroxylase; DAT, Deacetylvinodoline O-AcetylTransferase. La conversion de l'aglycone de strictosidine en tabersonine nécessite des étapes non connues avant 2018. B, voie de biosynthèse des phytostérols et des triterpènes chez les plantes. SQS, Squalène Synthase; SQE, Squalène Epoxidase; CAS, Cyclo-Artenol Synthase; α-AS, α-Amyrin Synthase. Le flux de squalène peut être dirigé vers la production de phytostérol (voie de la CAS) ou bien vers les triterpènes (voie de l' α-AS). L'existence de deux isoformes de SQS suggère une implication spécifique pour ces deux voies. C, signalisation cytokinique (CK) chez les plantes (C1). La signalisation implique un récepteur de type Histidine Kinase (HK), une phospho-transférase (HPt) et des régulateurs de réponse (RR), de type A, B ou C. Chaque niveau est représenté par plusieurs protéines (par exemple AHK2, 3 et 4 chez *Arabidopsis*). Les panneaux C2, C3 et C4 montrent des implications différentes dans certains processus biologiques. SP, CK provenant du Sporophyte; ES, sac embryonnaire; FM, mégasporangium avant maturation. Une approche de co-expression viserait à montrer de nouvelles associations entre les différents éléments de signalisation.

La compréhension de voies de signalisation hormonales ainsi que des voies de biosynthèse de métabolites spécialisés des plantes est étudié depuis de nombreuses années au laboratoire EA2106. Les études menées sur ces problématiques incluent classiquement des approches de biologie moléculaire : caractérisation enzymatique, invalidation fonctionnelle de gènes et études de localisation pour la validation de gènes candidats. Comme montré en introduction, une quantité massive et croissante de données transcriptomiques est disponible dans les bases de données publiques entraînant aussi le développement important des bases de données de co-expression. Ainsi, il a été entrepris d'initier, parallèlement aux approches expérimentales, des approches bio-informatiques axées sur le transcriptome pour contribuer à la résolution des problématiques abordées à l'EA2106. Parmi ces thématiques, le métabolisme des alcaloïdes indoles monoterpéniques (AIM) chez *C. roseus*, le métabolisme des triterpènes chez *M. domestica* et la voie de signalisation répondant au cytokinines (CK) chez les plantes (**Figure 19**) pourraient bénéficier de la mise en place de réseaux de co-expression ciblés comme nouvel outil analytique.

Concernant les voies de biosynthèse du métabolisme spécialisé, les objectifs de la co-expression visent à (i) reconstruire des blocs de voie par identification d'étapes enzymatiques encore non connues et (ii) décrire leur paysage transcriptionnel pour comprendre leur intégration dans le métabolisme cellulaire, notamment par l'identification de transporteurs et de facteurs de transcription associés (**Figure 19A et B**). Ces blocs de voie sont supposés fonctionner de manière coordonnée, et il est attendu que les gènes codant les enzymes impliquées dans ce bloc montrent une co-expression forte. Dans le cas de voies de signalisation, l'approche de co-expression pourrait permettre de mieux appréhender leur fonctionnement global et leur intégration dans le métabolisme cellulaire.

**L'ensemble du travail de thèse présenté ici s'oriente vers le développement et l'optimisation d'une méthodologie visant à construire des réseaux de co-expression ciblés pour remplir les objectifs décrits ci-dessus.**



## 1. Mise en place initiale

La voie de signalisation répondant au CK est complexe dans son fonctionnement, et correspond à une voie de type Multi-Step Phosphorelay (MSP). Elle fait appel à 3 types d'acteurs différents, des récepteurs (de type Histidine Kinase, HK, ou CHK pour le cas des HK fixant les CK), des phosphorelais (HPt) et des régulateurs de réponse (RR), chaque niveau impliquant généralement plusieurs gènes (**Figure 19C**). Par exemple, chez *A. thaliana*, il existe 3 récepteurs aux CK, 6 HPt et plus d'une 20aine de RR. Connaissant la diversité des processus biologiques/physiologiques régulés par les CK, il est attendu que des associations préférentielles entre certains acteurs permettent de réguler des processus spécifiques. Les études de co-expression sur cette voie devraient idéalement permettre de conforter cette hypothèse. Cependant, il est important de prendre en considération le fait que cette voie fait appel de manière importante à des processus de régulation post-traductionnelle, ce qui pourrait rendre l'analyse de co-expression moins performante. D'un point de vue théorique et de par sa complexité, cette voie pourrait donc être idéale pour trouver une méthodologie optimale permettant de construire une co-expression robuste dessus.

Au sein l'EA2106, l'un des objectifs du travail de thèse de Dimitri Daudu, initié en 2014 et encadré par Joël Crèche, Gaëlle Glévarec et Sébastien Besseau a porté sur la caractérisation des récepteurs aux CHK chez le pommier, incluant une étude de leur spécificité d'association avec les autres membres de la voie CK. De plus, un travail collaboratif mené à cette période avec l'Institut de Recherche sur la Biologie de l'Insecte et plus précisément l'équipe de David Giron a porté sur l'étude de la mineuse du pommier, un ravageur intra-foliaire dont l'installation entre les cellules de la plante nécessite l'activation de la voie CK (Zhang et al., 2016). Cette activation semble être liée au maintien en vie du tissu colonisé, formant des « îles vertes » caractéristiques sur des feuilles pourtant sénescentes (Kaiser et al., 2010). Ainsi, la voie des CK pourrait également entrer en jeu dans des processus de réponses aux agents biotiques, et une meilleure connaissance de son fonctionnement ne pourrait qu'être favorable à la compréhension globale de ce processus biotique.

Une étude de co-expression a donc été initiée en premier lieu sur la voie de signalisation répondant aux CK pour mettre en évidence des associations préférentielles qui pourraient correspondre à des processus physiologiques particuliers. Des données de coexpression



disponibles dans les bases de données ATTED-II et PlaNet en 2014 ont alors été utilisées en utilisant les gènes d'*A. thaliana* connus pour participer au fonctionnement de la voie. Les travaux du laboratoire étant réalisés sur les CK du Pommier, il paraissait impossible de mener l'analyse de co-expression sur cette espèce, étant donné la faible quantité de données transcriptomiques disponibles, en comparaison avec la plante modèle *A. thaliana*. Les deux bases de données nommées ci-dessus ne permettaient pas de récupérer les gènes communément co-exprimés avec un ensemble de gènes guides, rendant difficile une approche de type PLC. De plus, sur la database PlaNet par exemple, seuls des échantillons correspondant à des stades développementaux étaient intégrés. Or dans notre objectif de comprendre le fonctionnement de la voie CK, il paraissait important d'inclure également des données transcriptomiques obtenues lors de situations physiologiques particulières (stress biotiques par exemples). Dans le cas des données disponibles sur ATTED-II, nous avions rencontré des difficultés quant à leur utilisation notamment car certains gènes ne pouvaient être identifiés correctement par leurs numéros d'accession. Malgré ces quelques inconvénients, une première ébauche de réseau de co-expression a pu être faite sur la voie de signalisation des CK, basée uniquement sur des données microarrays.

## 2. Méthodologie générale

### 2.1. Identification et préparation des données d'expression

Dynamisés par ces premiers résultats dont certains étaient bien confirmés par la littérature antérieure, nous avons voulu contrôler le processus de création du réseau dans son intégralité, en travaillant à la fois sur des données microarrays et RNA-seq. Après identification des données dans les bases spécialisées (**Table III**), un grand nombre de données expérimentales obtenues sur la puce Affymetrix ATH-1 (GPL198) a pu être retraité à partir de la base ArrayExpress hébergée par l'EBI. Le re-traitement des données RNA-seq est plus long en raison de la taille des fichiers traités, mais en adaptant la procédure de quantification, une matrice large a pu être générée. Étant donné le nombre important de fichier dans les deux jeux de données, des procédures automatisées ont été mises en œuvre par l'utilisation de scripts bash (**Table IV**).



**Table III: Organisation des données transcriptomiques déposées dans les bases de données publiques.**

Microarray	RNA-seq
<i>Accession de projet :</i> GSE	<i>Accession de projet :</i> ERP/DRP/SRP
<i>Accession de plateforme :</i> GPL	<i>Accession d'échantillon :</i> ERS/DRS/SRS
	<i>Accession d'expérience :</i> ERX/DRX/SRX
<i>Accession de lame:</i> GSM	<i>Accession de run :</i> ERR/DRR/SRR

La table présente les différents préfixes accompagnant les numéros d'acquisition.

**Table IV: Étapes pour la récupération des données publiées à partir de bases publiques.**

	Microarray	RNA-seq
1. identification	ArrayExpress	NCBI SRA, run info
2. téléchargement	R avec le paquet « ArrayExpress », multicoeur	EBI ENA, transfert ftp prozilla (multicoeur) + array job
3. préparation des données	R avec le paquet affy, normalisation par quantile pour chaque étude (acquisitions GSE)	Quasi-alignement sur transcriptome de référence avec Salmon
4. sortie	Fichiers .CEL normalisés pour chaque lame	Fichiers quant.sf pour chaque run

Chargement des paquets 'igraph' et 'data.table':

```
library(igraph)
library(data.table)
```

Lecture de la matrice d'expression, donner les dimensions et afficher les 5 premières lignes et colonnes

```
a<-fread("160422_ath_gpl198_AE_norm_table_100x20")
a<-data.frame(a, row.names=1)
dim(a)
```

```
## [1] 99 19
a[1:5,1:5]
```

```
##          cell12_ATH1.CEL cell13_ATH1.CEL root1_ATH1.CEL root2_ATH1.CEL
## 244901_at      6.237768    6.073733    6.449131    6.001477
## 244902_at      6.816436    6.605402    7.280846    6.237227
## 244903_at     10.687331   10.106031   10.524236   7.540518
## 244904_at      8.768581    8.274659    8.057254    5.592247
## 244905_at      4.052098    3.713479    3.416662    3.296971
##
##          root3_ATH1.CEL
## 244901_at      6.156032
## 244902_at      6.276389
## 244903_at      7.710363
## 244904_at      6.104748
## 244905_at      3.099650
```

Calcul de toutes les corrélations, afficher les dimensions de la matrice de distance

```
cor.a<-cor(t(a), method="pearson")
dim(cor.a)
```

```
## [1] 99 99
```

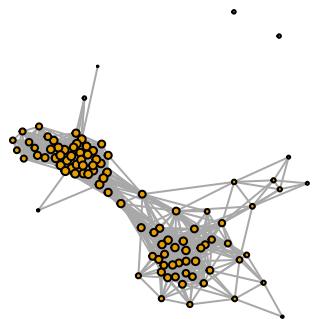
Retenir les associations pour laquelle  $r > 0.9$ . Attribuer 1 si significatif, 0 si non significatif et visualiser les 5 premières lignes et colonnes de la matrice de proximité

```
cor.a.09<-cor.a
cor.a.09[cor.a.09<0.9]<-0
cor.a.09[cor.a.09>=0.9]<-1
cor.a.09[1:5, 1:5]
```

```
##          244901_at 244902_at 244903_at 244904_at 244905_at
## 244901_at      1       1       1       1       0
## 244902_at      1       1       1       1       0
## 244903_at      1       1       1       1       0
## 244904_at      1       1       1       1       0
## 244905_at      0       0       0       0       1
```

Convertir la matrice de proximité en graphique et visualiser

```
cor.a.09.net<-simplify(graph_from_adjacency_matrix(cor.a.09, mode="undirected"))
plot(cor.a.09.net, vertex.label="", vertex.size=log2(degree(cor.a.09.net)+1))
```



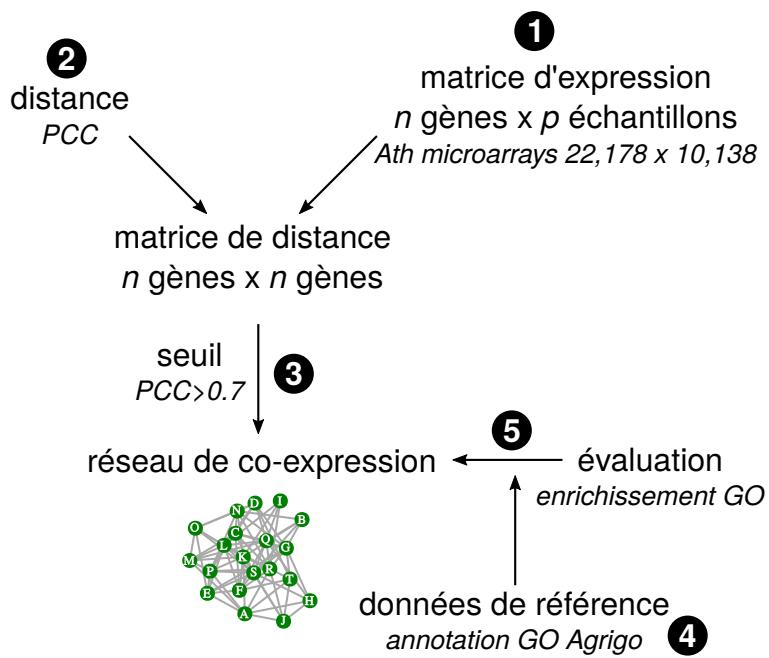
**Figure 20: exemple de code R pour la création d'un réseau de co-expression.** L'exemple utilise un extrait de la matrice d'expression d'*Arabidopsis thaliana*. Les zones de code R sont représentées par des blocs grisés.

## 2.2. Ressources informatiques

La réalisation de ce travail bioinformatique n'aurait pas été possible sans la mise à disposition de ressources de calcul. Deux ressources principales ont été utilisées, la machine Neptune hébergée par l'Université de Tours et le cluster Artemis hébergé à l'Université d'Orléans et géré par la fédération CaSciMoDot (Calcul Scientifique et Modélisation Orléans-Tours). L'accès illimité à ces machines financées par la Région Centre et le CNRS a été d'une aide précieuse pour la collecte et le traitement des données. Les opérations réalisées sont consommatrices de mémoire et de ressource CPU. Pour diminuer les temps de calcul, les opérations ont été conduites en parallèle dans la mesure du possible, c'est-à-dire avec une répartition des tâches sur plusieurs CPUs. Les machines d'Artemis étant organisée par SLURM, des tâches parallèles, soit sous forme de job array soit via GNU parallel, ont pu être lancées assez facilement. Un résumé des ressources utilisées dans Artemis est présenté dans la

### Table V.

Le langage de base utilisé pour la construction et l'analyse des réseaux est R (**Figure 20**). Des boucles répartissant des tâches sur plusieurs CPUs peuvent être créées pour augmenter la rapidité du calcul. Cependant, certaines étapes tel que l'évaluation par *neighbour voting* par l'intermédiaire du paquet EGAD (Ballouz et al., 2016) mettent en mémoire beaucoup de données, en particulier sur des réseaux larges et avec beaucoup de termes GO, et peuvent prendre beaucoup de temps.



**Figure 21: Protocole de construction d'un réseau de co-expression.** Les termes en italique représentent des valeurs variables, des données variables ou des distances variables et constituent donc des exemples dans le schéma ci-dessus. Les numéros correspondent aux points critiques de la construction et de l'évaluation du réseau comme décrit dans le contexte.

**Table V: résumé des consommations informatiques sur les ressources Artemis et Neptune.**

	Partie 1 + 2	Matrice RNA-seq Arabidopsis (4,308)	Toutes données RNA-seq	Toutes données microarrays
<b>Temps total (jours)</b>	160	102	166	0,3
<b>Temps CPU (jours)</b>	3 280	848	1 844	4
<b>Nombre de travaux</b>	989	16	701	52
<b>Nombre de tâches</b>	227 432	16	4 666	52
<b>Nombre de CPUs moyen</b>	15	7	10	10
<b>Ecriture disque totale (To)</b>	19,4	16,6	65,3	0,3
<b>Nombre de lignes de codes bash</b>	Partie 1 : 1 511 Partie 2 : 2 148	104	104	87
<b>Nombre de lignes de codes R</b>	Partie 1 : 2 748 Partie 2 : 2 517	74	74	289

Le nombre de ligne de codes n'inclue pas les codes utilisés en local pour l'analyse des résultats de ces travaux.

### 2.3. Construction des réseaux de co-expression

Une fois ces données récupérées, la construction des réseaux a été engagée. Si l'objectif premier était l'étude de voies biologiques pour fournir des éléments (gènes, associations,...) à valider fonctionnellement, la manipulation des données et l'inférence des réseaux s'est révélée cacher plusieurs points critiques. Ce processus engage différents choix comme indiqués dans la **Figure 21** : la distance, la matrice d'expression initiale, le seuil, le set de validation et la technique de confrontation (réseau vs set de validation).

Les différents points expliqués ci-dessous sont représentés par leurs numéros respectifs dans la **Figure 21**.



### **2.3.1. Choix de la matrice initiale**

Comme vu dans l'introduction, les trois paramètres qui influenceront la construction seront le nombre de gènes, le nombre d'échantillons et le type de normalisation. La variation du nombre de gènes jouera sur la topologie finale, d'autant plus si certains hubs sont manquants. En fonction du type de calcul choisi, la distance entre deux gènes pourra être modifiée par la présence ou l'absence d'autres (cas des rangs et des corrélations partielles par exemple). Assez peu de littérature est disponible à ce sujet. La variation du nombre d'échantillons est aussi connue pour influencer la pertinence des réseaux, mais encore une fois assez peu de données ont été publiées sur ce point. La nature des échantillons va impacter le calcul des distances entre gènes, et il semble que l'information capturée sature à partir d'un certain nombre d'échantillons lors de l'utilisation de coefficients de corrélation. Enfin, le type de normalisation affecte la distributions des valeurs d'expressions. Celles-ci vont donc différer en fonction de la normalisation appliquée, conduisant des différences d'expression entre deux gènes qui ne seront pas les mêmes. Il est à noter que certaines distances (notamment les coefficients de corrélation ordonnés) sont moins sensibles à l'impact de la normalisation.

### **2.3.2. Choix de la distance**

Comme vu précédemment, il existe de nombreuses manières pour établir une dépendance entre deux gènes. Le choix de la méthode sera guidée à la fois par sa performance et par son applicabilité. La performance se mesure par la pertinence des associations retenues dans le réseau final. Elle est peut être évaluée de différentes manières (cf paragraphe suivant). De nombreuses études ont comparé différentes méthodes mais il en ressort globalement que des coefficients de corrélation classiques (Pearson ou Spearman) apportent des résultats tout à fait satisfaisants, tout en tenant compte de leur applicabilité. En effet, cette notion d'applicabilité est également critique car le calcul d'une distance entre deux gènes peut être lourd, notamment lorsqu'il intègre les valeurs d'expression d'un ou plusieurs autres gènes. Certaines méthodes, en particulier celles basées sur des inférences probabilistes, semblent très complexes à mettre en place (voire inapplicables) sur des données larges. D'autres sont plus



facilement parallélisables en terme informatique et par conséquent plus aptes à décrire l'ensemble des combinaisons de gènes possibles. Ceci est d'autant plus crucial si de nombreux réseaux doivent être construits (approches multi-espèce ou comparaison de différentes matrices de données). La littérature (par exemple (Ballouz, Verleyen & Gillis, 2015)) s'accorde sur le fait que les coefficients de corrélation sont les plus simples à calculer car ils nécessitent globalement peu de ressources informatiques.

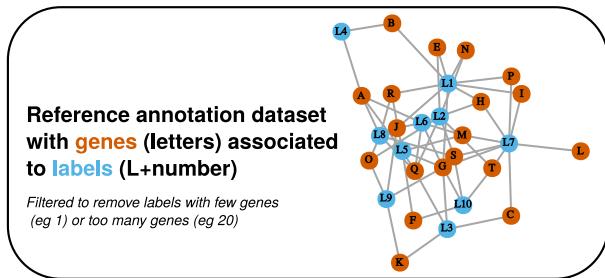
### **2.3.3. Choix du seuil**

Le seuil fixe le nombre d'associations retenues pour construire le réseau. Les effets de ce seuil sur la topologie finale sont bien connus (Couto, Comin & da Fontoura Costa, 2017). Comme mentionné en introduction, il existe des manières de fixer des seuils optimisant l'inférence du réseau. Une manière plus empirique consiste à regarder des paramètres topologiques tels que la densité ou la transitivité pour différentes valeurs de seuils. Des topologies de type invariant d'échelle sont bien plus faciles à analyser notamment en terme de détection de communautés. A l'inverse, des réseaux très denses contenant à la fois beaucoup de connexions et de gènes seront difficilement interprétables d'un point de vue biologique. Pour chaque seuil testé, un réseau est généré et doit être confronté au set de validation.

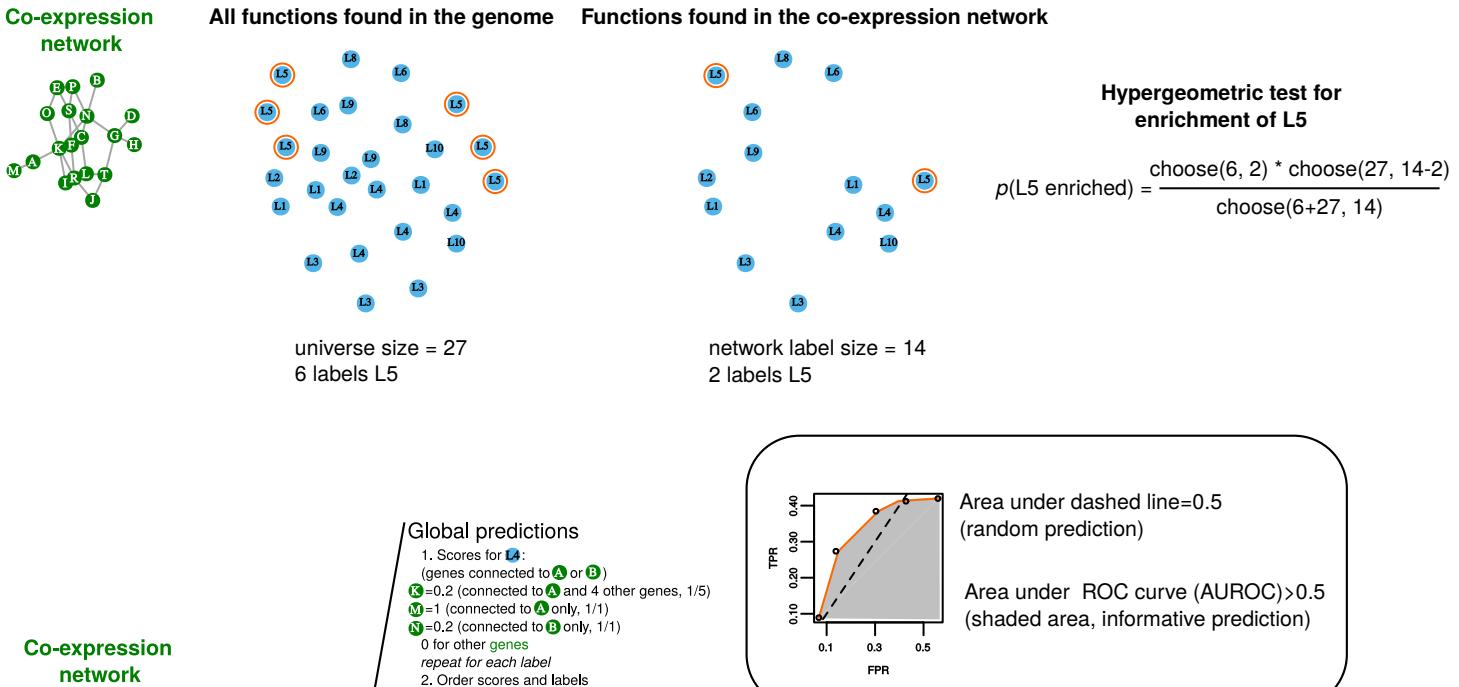
### **2.3.4. Le set de validation**

L'analyse du réseau décrite dans l'introduction implique nécessairement une étape d'évaluation qui confronte le réseau obtenu à des données validées expérimentalement ou transférées *in silico* d'espèces modèles. Il peut s'agir d'une liste de gènes annotés par catégories fonctionnelles, tel qu'avec la Gene Ontology (disponibles sur des databases précises (par exemple Agrigo) ou généralistes comme ou Uniprot) ou bien de groupes de gènes dont les produits agissent ensemble dans une même voie (principe de la PLC exposée en introduction). Le choix de la distance et de la matrice initiale devra être effectué de manière à ce que le réseau final ait la meilleure performance par rapport à ce set. Ce point est

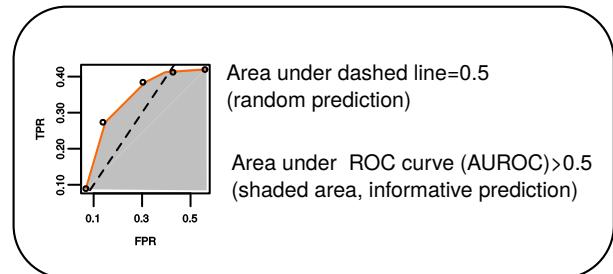
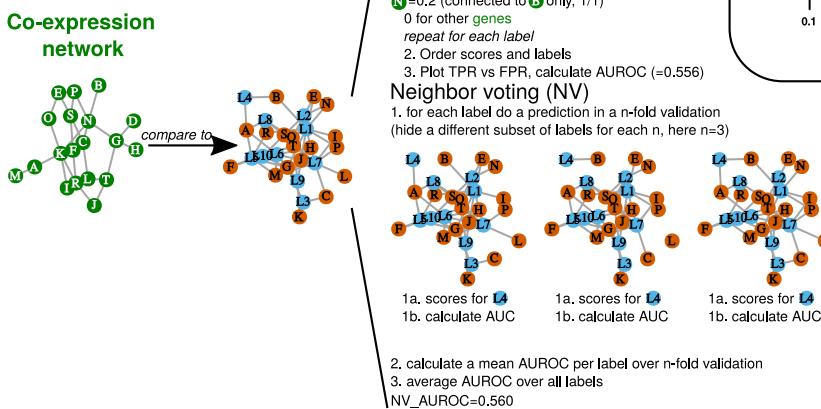
A



B



C



**Figure 22: Confrontation d'un réseau de co-expression à une annotation de référence.** A, réseau montrant l'annotation fonctionnelle d'une partie des gènes d'un organisme. Une annotation "label" (par exemple un terme GO ou domaine Pfam) peut être partagée par plusieurs gènes, et un gène peut avoir plusieurs annotations. B, calcul de l'enrichissement d'annotation par test hypergéométrique. Ce genre de test vise à déterminer si une annotation est particulièrement sur-représentée par des gènes connectés dans le réseau (ou dans une communauté d'un réseau). Il s'agit de calculer la probabilité d'enrichissement (par exemple pour l'annotation L5,  $p(\text{L5 enriched})$ ) qui tient compte du nombre d'occurrences de cette annotation dans le génome et dans le réseau ainsi que du nombre total d'annotations dans le génome et dans le réseau. Le calcul inclue des combinaisons au sens mathématique (choose) de n éléments parmi k (par exemple  $\text{choose}(6,2)=6!/(6-2)! \times 2!$ ). C, évaluation de la prédictivité par calcul des taux de vrais positifs (TPR) et de faux positifs (FPR). Un graphique représentant les valeurs des TPR en fonction des FPR s'appelle une courbe ROC (Receiver Operating Characteristic), dont l'aire informe sur la prédictivité du réseau. Si  $\text{TPR} = f(\text{FPR})$  est de type  $x=y$ , alors l'information donnée par le réseau n'est liée qu'au hasard. Dans ce cas, l'aire sous la courbe  $\text{TPR}=f(\text{FPR})$  autrement appelée Area Under ROC curve ou AUROC est égale à 0.5. Si cette aire est >0.5 (cas de la partie grisée) alors le réseau présente une certaine prédictivité, qui sera parfaite si AUROC=1. Le calcul de cette prédictivité se fait grâce au paquet R 'EGAD' (Ballouz et al 2016), soit d'un point de vue global en considérant chaque terme, soit en calculant une moyenne des AUROC pour chaque terme GO. Dans ce cas, les AUROC de chaque terme sont calculées par un algorithme de type neighbor voting en appliquant une validation croisée multiple (ici 3) qui consiste à mesurer la prédictivité lorsqu'une partie de l'annotation est cachée.

critique car il permet de renforcer la validité d'associations présentes dans le réseau mais non décrites par ailleurs dans la littérature.

### 2.3.5. La technique de confrontation

Il s'agit de la manière de quantifier comment le réseau s'aligne avec le set de validation (**Figure 22**). Une manière directe est de compter les associations correspondant à des relations connues (par exemple deux gènes annotés avec un même terme GO) et d'en mesurer leur enrichissement, c'est-à-dire déterminer si leur présence quantitative dans le réseau est liée au hasard ou statistiquement significative (**Figure 22B**). Le calcul de probabilité par rapport à une distribution hypergéométrique est tout à fait appropriée dans ce cas. Ce genre de distribution permet de calculer la probabilité d'obtenir  $x$  observations valides parmi  $k$  tirées d'un univers de  $n$  observations possibles contenant  $m$  observations valides ( $x < n$ ,  $k < n$ ,  $x < m$ ). Une  $p$ -value est obtenue pour chaque terme GO et est ajustée en fonction du nombre de tests réalisées, donnant la  $q$ -value.

$$p(x) = \text{choose}(m, x) \text{ choose}(n, k-x) / \text{choose}(m+n, k)$$

La pertinence d'un réseau estimée par ce genre de test se voit par le nombre de termes significativement enrichis et les valeurs de  $x$  et de  $p(x)$ .

D'autres méthodes moins directes visent à calculer une performance absolue, c'est-à-dire une valeur unique qui permet d'évaluer directement la pertinence du réseau. Dans ce cas, l'algorithme, généralement construit sur des méthodes de machine learning, calcule la prédictivité du réseau par rapport au set de validation. Une manière assez simple et performante est d'appliquer une méthode de *neighbor voting* reposant sur le principe de culpabilité par association ((Ballouz, Verleyen & Gillis, 2015), **Figure 22C**). En cachant plusieurs fois une partie des données du set de validation, et pour chaque étiquette (terme GO par exemple) et en attribuant les étiquettes restantes aux gènes connectés dans le réseau, elle calcule le taux de faux positifs et de vrais positifs sur le réseau. Dans ce cas, les vrais positifs



correspondent à des gènes connectés dans le réseau et possédant une même étiquette. Dans le cas de ces méthodes, il est possible de calculer un taux de faux positifs (FPR, false positive rate) et un taux de vrais positifs (TPR, true positive rate) et donc une aire sous la courbe de  $\text{TPR} = f(\text{FPR})$  autrement appelée Receiver Operating Characteristic (ROC). L'aire sous la courbe ROC (AUROC, Area Under ROC curve) peut être utilisée comme un indicateur de la puissance prédictive du réseau. Elle est généralement comprise entre 0.5 et 1, 0.5 révélant une prédiction non différente du hasard et 1 une prédictivité parfaite. Un paquet R appelé EGAD a été récemment développé et offre une véritable opportunité pour l'évaluation de réseaux de co-expression larges par le calcul d'AUROCs (Ballouz et al., 2016).

La validation du réseau peut aussi se faire par l'étude de communautés. Comme vu précédemment, ces communautés regroupent des paquets de gènes plus densément connectés, voire éventuellement des cliques (ensemble de gènes dont aucun n'est relié à d'autres ailleurs dans le réseau). L'étude de voies métaboliques par PLC, et de la manière dont les gènes impliqués dans une même voie se retrouvent répartis dans une même communauté, peut être un bon indice de la qualité des associations récupérées dans le réseau. Nous avons développé à ce titre une mesure de la qualité de répartition de gènes dans ces mesurées basée sur une table de contingence et un test du Chi<sup>2</sup>.

### **3. Organisation de la partie Résultats**

Compte tenu de la complexité inhérente à la construction d'un réseau de co-expression, nous avons étudié un certain nombre de points pour optimiser cette construction et trouver une feuille de route robuste.

**La première partie, intitulée « Le classement des mesures de corrélation à l'échelle du génome améliore les réseaux de co-expression globaux et ciblés construits à partir de données microarray et RNA-seq »,** porte sur l'optimisation du calcul de distance entre gènes mais aussi sur la normalisation des données RNA-seq et la méthode d'évaluation des réseaux. Il s'agit donc des points 1, 2 et 5 détaillés ci-dessus. Cette première partie nous a permis de fixer une préparation des données ainsi qu'une distance basée sur des PCC réciproquement



ordonnés (PCC-HRR) pour construire des PLC robustes, tout en montrant la pertinence d'une évaluation du réseau par les AUROC mais aussi d'autres indices comme celui basé sur un test du Chi<sup>2</sup>. Pour cette mise en place, nous avons focalisé sur des données transcriptomiques (microarrays et RNA-seq) d'*Arabidopsis thaliana*.

**Pour la seconde partie, intitulée « Une affaire de taille : gestion d'échantillons de larges jeux de données d'expression dans la construction de réseaux de co-expression robustes »,** nous avons voulu étudier l'impact de la taille des matrices d'expression en termes de nombres d'échantillons sur la qualité du réseau inféré. Ce point a été assez peu étudié dans la littérature et a donc nécessité une évaluation exhaustive. Utilisant les PCC-HRR comme distance de base, un grand nombre de matrices de tailles différentes obtenues aléatoirement ou selon un groupage thématique (par étude ou par *k*-moyennes) a été évalué sur trois espèces, *A. thaliana*, *Solanum lycopersicum* et *Zea mays*. Pour chaque espèces, des jeux de données obtenus par microarrays et RNA-seq ont été testés. Il ressort de cette étude que des réseaux construits sur des datasets larges sont pertinents, mais qu'il est possible d'obtenir une performance accrue en agrégeant plusieurs réseaux construits à partir de jeux de données préalablement partitionnés.

**Les deux phases d'optimisation décrites dans ces deux parties ont pour but de fournir des outils pertinents pour l'étude de voie biologiques.**

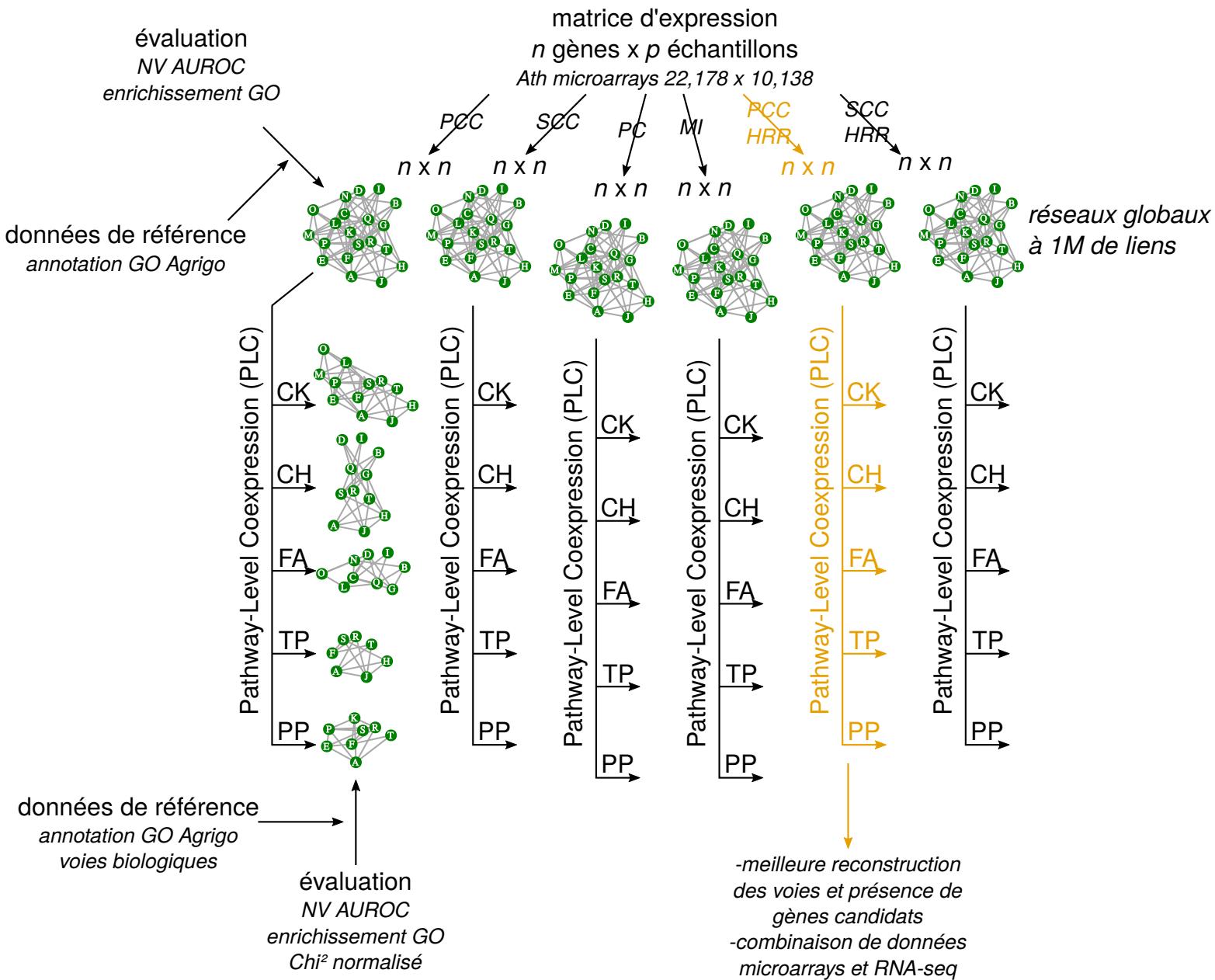
**Dans la troisième partie, intitulée « Associations préférentielles au sein des voies de signalisation de type phospho-relais à étapes multiples : étude comparative de réseaux de co-expression ciblés chez 15 espèces végétales »,** la voie de signalisation CK a été étudiée par une approche comparative de réseaux de co-expression. Tirant parti de l'optimisation de la distance et de la performance affichée de jeux de données larges, des réseaux de co-expression ont été construits sur une quinzaine d'espèces de végétaux, allant de l'algue verte unicellulaire *Chlamydomonas* aux Angiospermes. L'objectif de cette étude a été de mettre en évidence des associations transcriptionnelles conservées entre espèces taxonomiquement différentes au niveau de cette voie de transduction.

**L'application des réseaux de co-expression aux voies des AIM et des triterpènes sera présentée en discussion.**



# Partie I

**Le classement des mesures de corrélation à l'échelle du génome améliore les réseaux de co-expression globaux et ciblés construits à partir de données microarray et RNA-seq**



**Figure 23: Impact de la distance sur la performance de réseaux de co-expression.** 7 matrices d'expression ont été initialement testées: microarrays (indiqué dans le schéma ci-dessus) et 6 matrices RNA-seq normalisées avec différentes approches. Pour chaque matrice, 6 distances ont été testées, générant à chaque fois un réseau global (42 au total) qui a été interrogé par une série de 6 sets de gènes guides (CK, signalisation cytokinine, CH, carbohydrates, FA, facides gras, TP, terpènes et PP, polyphenol). A chaque réseau ou sous réseau, une évaluation de la qualité est faite. La méthodologie indiquée en orange correspond à la plus appropriée pour l'analyse de ces voies.

Notre ambition est de mettre au point une méthodologie simple pour la construction de réseaux ciblés de co-expression pour une analyse robuste de voies cellulaires. La construction d'un réseau de co-expression nécessite en tout premier lieu de choisir un jeu de données et une distance qui permettra d'identifier des associations potentielles entre transcrits.

Nous avons choisi l'espèce *Arabidopsis thaliana* comme modèle, étant donné l'important savoir biologique accumulé pour cette plante depuis des décennies. Ces connaissances sont cruciales pour considérer un jeu de données de référence qui servira à évaluer les réseaux qui seront construits.

Concernant le choix de la distance, plusieurs études préliminaires suggèrent (i) que les méthodes non supervisées telles que les coefficients de corrélation peuvent fournir de bons résultats et (ii) que les méthodes supervisées complexes peuvent apporter plus de précisions mais sont difficilement applicables à des matrices d'expression larges. Dans notre optique de comprendre le fonctionnement de voies cellulaires par l'étude de leur paysage transcriptionnel, il est pourtant capital de considérer un maximum de gènes du génome. Par conséquent, nous avons voulu tester l'applicabilité de méthodes non supervisées sur l'inférence de réseaux de co-expression à partir d'une matrice large. Au total, six méthodes ont été testées : corrélations partielles (PC), information mutuelle (MI), PCC, SCC, PCC-HRR et SCC-HRR. Les versions classées réciproquement des PCC et SCC par HRR ont été introduites précédemment par Mutwil et al 2012 et aucune étude n'a réalisé de comparaison de leur performance vis-à-vis d'autres distances non supervisées.

La stratégie adoptée est présentée en **Figure 23**, avec une évaluation utilisant en entrée des données microarray ou RNA-seq. Pour les données RNA-seq, six différents types de normalisation ont été testées : Transcrits Par Million (TPM), TPM normalisés par stabilisation de variance (VST), log2 TPM, comptes bruts, comptes bruts normalisés par VST et log2 comptes brutes. Au total, 7 jeux de données ont été utilisés et pour chaque, 6 distances ont été testées pour inférer des réseaux de co-expression. L'évaluation de ces réseaux a été faite sur leur globalité (sur 1 million de liens), puis après capture de voies spécifiques selon le principe de PLC (carbohydrates, acides gras, phénylpropanoïdes, terpènes et cytokinine) (**Figure 18**). La qualité des réseaux est estimée par le nombre de termes GO significativement enrichis, les valeurs d'AUROC total et NV\_AUROC (voir la partie Méthodologie générale, **Figure 21**). Pour les PLC, la performance de capture des termes GO est estimée similairement mais nous avons aussi suivi la manière dont les gènes guides sont organisés en communautés. Plus



particulièrement, nous avons déterminé si les gènes guides d'un même bloc de voie étaient correctement regroupés au sein d'une même communauté. Cette répartition est évaluée en calculant une valeur de Chi<sup>2</sup> qui est normalisée par rapport à la valeur maximale qu'aurait le Chi<sup>2</sup> dans le cas d'une partition parfaite de ces gènes guides.

Cette étude a clairement montré la supériorité des PCC-HRR à la fois sur les données microarrays et RNA-seq. De plus, cette distance permet également de créer un réseau de co-occurrence regroupant les liens conservés entre les réseaux issus de ces deux technologies. Une telle co-occurrence est potentiellement synonyme de robustesse car elle implique l'existence d'associations transcriptionnelles révélées simultanément par les deux technologies. Des gènes marqueurs clés non utilisés en gènes guides ont été retrouvés dans les PLC sur les 5 voies cellulaires en utilisant cette distance. Ces résultats montrent la puissance des PLC pour étudier et mieux comprendre le fonctionnement de ces voies.

**Article publié dans *Scientific Reports***



# SCIENTIFIC REPORTS



OPEN

## Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks

Received: 23 March 2018

Accepted: 27 June 2018

Published online: 18 July 2018

Franziska Liesecke, Dimitri Daudu, Rodolphe Dugé de Bernonville, Sébastien Besseau, Marc Clastre, Vincent Courdavault , Johan-Owen de Craene , Joel Crèche, Nathalie Giglioli-Guivarc'h, Gaëlle Glévarec, Olivier Pichon & Thomas Dugé de Bernonville 

Co-expression networks are essential tools to infer biological associations between gene products and predict gene annotation. Global networks can be analyzed at the transcriptome-wide scale or after querying them with a set of guide genes to capture the transcriptional landscape of a given pathway in a process named Pathway Level Coexpression (PLC). A critical step in network construction remains the definition of gene co-expression. In the present work, we compared how Pearson Correlation Coefficient (PCC), Spearman Correlation Coefficient (SCC), their respective ranked values (Highest Reciprocal Rank (HRR)), Mutual Information (MI) and Partial Correlations (PC) performed on global networks and PLCs. This evaluation was conducted on the model plant *Arabidopsis thaliana* using microarray and differently pre-processed RNA-seq datasets. We particularly evaluated how dataset  $\times$  distance measurement combinations performed in 5 PLCs corresponding to 4 well described plant metabolic pathways (phenylpropanoid, carbohydrate, fatty acid and terpene metabolisms) and the cytokinin signaling pathway. Our present work highlights how PCC ranked with HRR is better suited for global network construction and PLC with microarray and RNA-seq data than other distance methods, especially to cluster genes in partitions similar to biological subpathways.

Constructing global gene co-expression networks is a popular approach to highlight transcriptional relationships (edges) between genes (vertices). The ‘Guilt-by-Association’ (GBA) principle supposes that genes sharing similar functions are preferentially connected and aims at predicting new functions for proteins by determining how their respective encoding genes are co-expressed with others using a reference dataset containing known gene functions such as the Gene Ontology (GO)<sup>1</sup>. Defining edges connecting genes remains a critical step in global co-expression network construction. Expression data (microarray or RNA-seq) are used to construct expression matrices (genes  $\times$  samples) and to calculate a distance or a similarity for each possible gene pair. The resulting pairwise distance matrix is then thresholded to obtain an adjacency matrix that discriminates relevant edges. Only edges with a distance below (or a similarity above) the set threshold are considered significant and retained for network construction. The procedure is expected to remove non biologically relevant gene associations while retaining the relevant ones and can be assessed with any reference dataset. Alternatively, guide gene sets may be used to extract more human-readable information from large networks in a process named Pathway-Level Coexpression (PLC)<sup>2–7</sup>. This approach aims at capturing the best transcriptional associations of a gene set and at highlighting functional gene groups such as known subpathways in this set. There are two types of approaches to determine transcriptional associations of genes: those that are supervised and those that are unsupervised. Supervised approaches such as regression and machine learning based methods require a prior knowledge which is used as a training dataset to recover biologically relevant gene associations and are used to infer regulatory networks, *i.e.* to uncover preferential and sequential interactions of a gene over the others. The superiority of

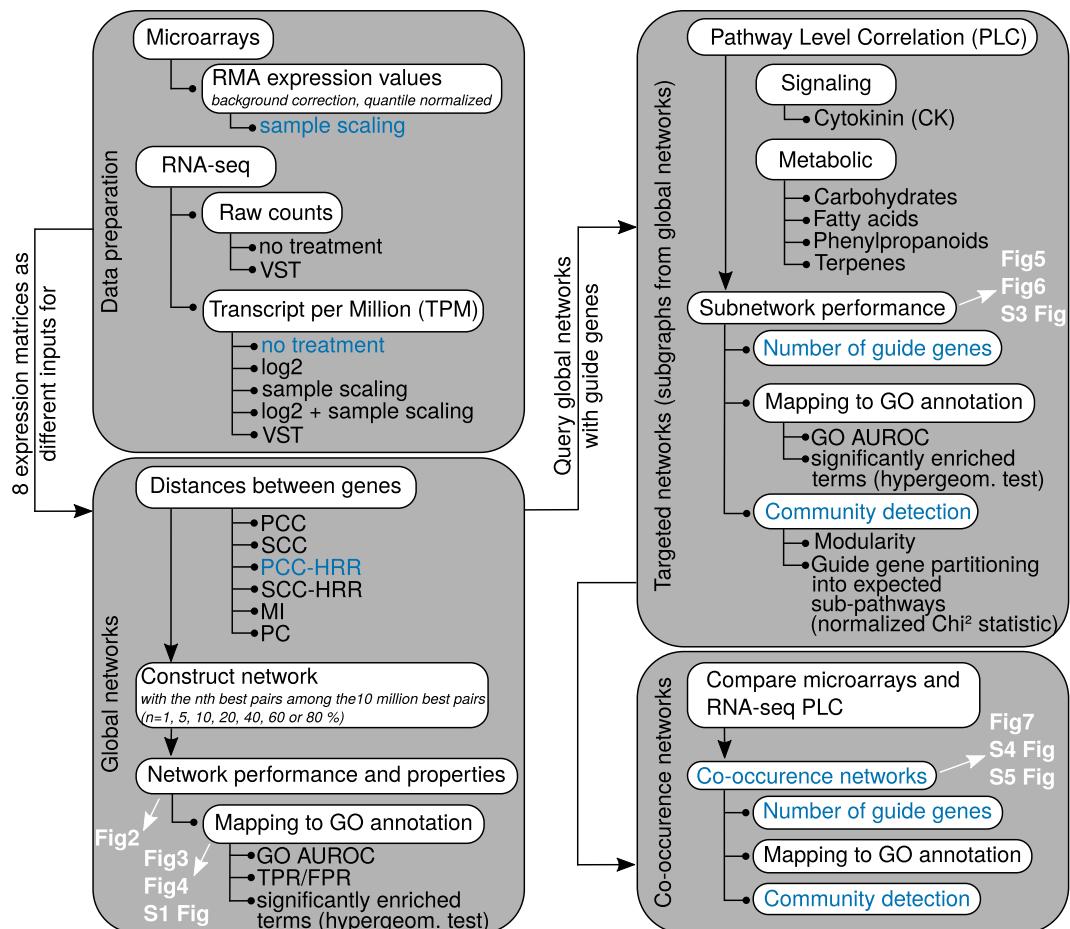
Université de Tours, EA2106 Biomolécules et Biotechnologies végétales, Tours, 37200, France. Correspondence and requests for materials should be addressed to T.D.d.B. (email: [thomas.duge@univ-tours.fr](mailto:thomas.duge@univ-tours.fr))



supervised methods in extracting potential physical regulatory interactions between genes has been demonstrated using simulated and real *E. coli* and *S. cerevisiae* subnetworks<sup>8</sup>. This study has revealed that prediction accuracy is higher with smaller networks and concluded that inferring genome-scale networks remains elusive unless performing a feature selection step to reduce inference problem size (because of the under determined nature of current expression datasets). Contrastingly, unsupervised approaches are used to capture transcriptional associations in co-expression networks. Although these associations may not be causative, a higher propensity of strongly correlated genes is a valuable information which may have a biological meaning in a cellular or physiological context. Among the unsupervised methods, four are commonly used and have been thoroughly tested. The first approach is Mutual Information (MI) which measures a statistical dependence between two variables<sup>8</sup>. It is based on density function estimates and has been shown to perform well with non linear relationships<sup>9</sup>. The second approach which relies on integrating multiple transcriptional associations is Partial Correlation (PC). First-order PCs may be simply calculated using the following equation  $R_{xy.z} = \frac{R_{yx} - (R_{yz})(R_{xz})}{\sqrt{(1 - (R_{yz})^2)(1 - (R_{xz})^2)}}$  with  $R_{xy}$

$R_{yz}$  and  $R_{xz}$  the simple correlation coefficients between genes x and y, y and z, x and z respectively<sup>10</sup>, but such computations may be very time consuming for large datasets. In this case, PCs should rather be calculated by multiple linear regression including a feature selection step<sup>11</sup>. In this way, PCs aim at explaining a gene's expression profile by a small number of strongly correlated genes after eliminating those less correlated that do not significantly explain this gene's expression profile. The two last methods are Correlation Coefficients (CCs), either Pearson CC (PCC) or Spearman CC (SCC), which are the classical estimators of linear transcriptional relationship among genes<sup>11,12</sup>. CCs are 2-dimensional distance measurements because a CC between two genes does not take into account the expression of the remaining transcripts in the whole transcriptome. To improve CCs, it has been proposed to recursively correct expression values according to a neural network algorithm with a weighted CC matrix until an equilibrium is reached between input and output expression<sup>13</sup>. Used on a small dataset, this method was useful to uncover interaction networks even for genes included in background cellular activities. Neural networks have however been reported to be computationally expensive for larger datasets<sup>14</sup>. Another improvement was to use ranked CCs instead of raw values. Ranking CC implies that for every gene, all CCs are calculated with the N-1 remaining genes (where N is the number of genes) and are ranked from 1 to N. Within a pair of genes A and B, rank(A to B) differs from rank(B to A) because the two genes display different expression profiles and different relationships with the remaining transcripts in the transcriptome. Two related ranking methods have been developed. One is mutual ranking (MR, geometric mean of the two ranks) which has been shown to improve GO term recovery with PCC using large microarray data from Arabidopsis, Human, mouse and rat<sup>15</sup>. MR has been successfully used in multispecies analysis of co-expression modules<sup>16</sup>. Another is Highest Reciprocal Ranking (HRR, maximum value of the two ranks)<sup>17</sup>. MR and HRR are thought to be more integrative than unranked CCs because they depend on other CC values around that of a gene pair. It might also be expected that ranking CCs partially correct for the “range restriction effect” observed for CCs leading to robust correlations for high variance genes only<sup>18</sup>. Although not as robust as supervised methods, unsupervised methods can efficiently capture relevant gene associations as previously shown<sup>9</sup>. These authors have shown that non parametric CC and MI calculations were more efficient than PCC on a small dataset. Among other unsupervised methods, SCC calculations have been similarly shown to outperform other distance measurements in Human expression data<sup>19</sup>. In this case, SCC were calculated from RNA-seq or microarray data in order to construct several smaller networks subsequently aggregated to yield the final network. We firmly believe that genome-scale networks inferred with CCs, especially when combined with a ranking procedure, are helpful to find new associations between genes. Although CCs are not efficient in detecting non linear associations<sup>9</sup>, gene-to-gene relationships have been predicted to be essentially linear<sup>20</sup> suggesting that CCs are valuable distance measurements. To date, there is no clear evaluation of how ranked CCs affect genome-scale network reconstruction with RNA-seq data in comparison with other unsupervised methods. We evaluated ranked CC, raw CC, MI, and PC performance in global and targeted network construction using Arabidopsis microarray and differentially processed RNA-seq expression data (Fig. 1). By using unsupervised methods, our aim was to highlight transcriptional associations between genes from an expression dataset. Querying genome-scale datasets is a useful preliminary step in identifying genes involved in biological pathways. Because correlation does not imply causality, the study of regulatory interactions or sequential orders through supervised methods is beyond the scope of our work. Although specific dedicated conditions are useful to find groups of intercorrelated genes<sup>21</sup>, large scale expression matrices were used here in order to increase the number of different experimental conditions and obtain robust averaged correlations. This approach should be useful for non-model species which global transcriptional landscapes have to be studied. Performance was measured as network ability to capture biologically relevant gene associations found in a Gene Ontology (GO) annotation reference set but also to correctly cluster guide genes in PLC. Global network quality was first evaluated according to the different dataset  $\times$  distance measurement combinations. The resulting global networks were next interrogated in PLC analyses with five different guide gene sets corresponding to four different metabolic pathways and one signaling pathway. Whereas metabolic pathways have relatively clearly defined and partially linear partitions, signaling pathways usually involve post transcriptional regulations and a more intricate organization, which might render gene transcriptional associations less evident. We looked at the dataset  $\times$  distance measurement combinations optimizing pathway reconstruction and maximizing co-occurrence quality between microarray and RNA-seq networks. Our results show that, of the six methods evaluated, PCC ranked with HRR generated the best biologically relevant networks according to initial guide gene representation and clustering in distinct modules. In addition, it offers the possibility to merge subgraphs obtained by microarrays and RNA-seq to generate high confidence networks.



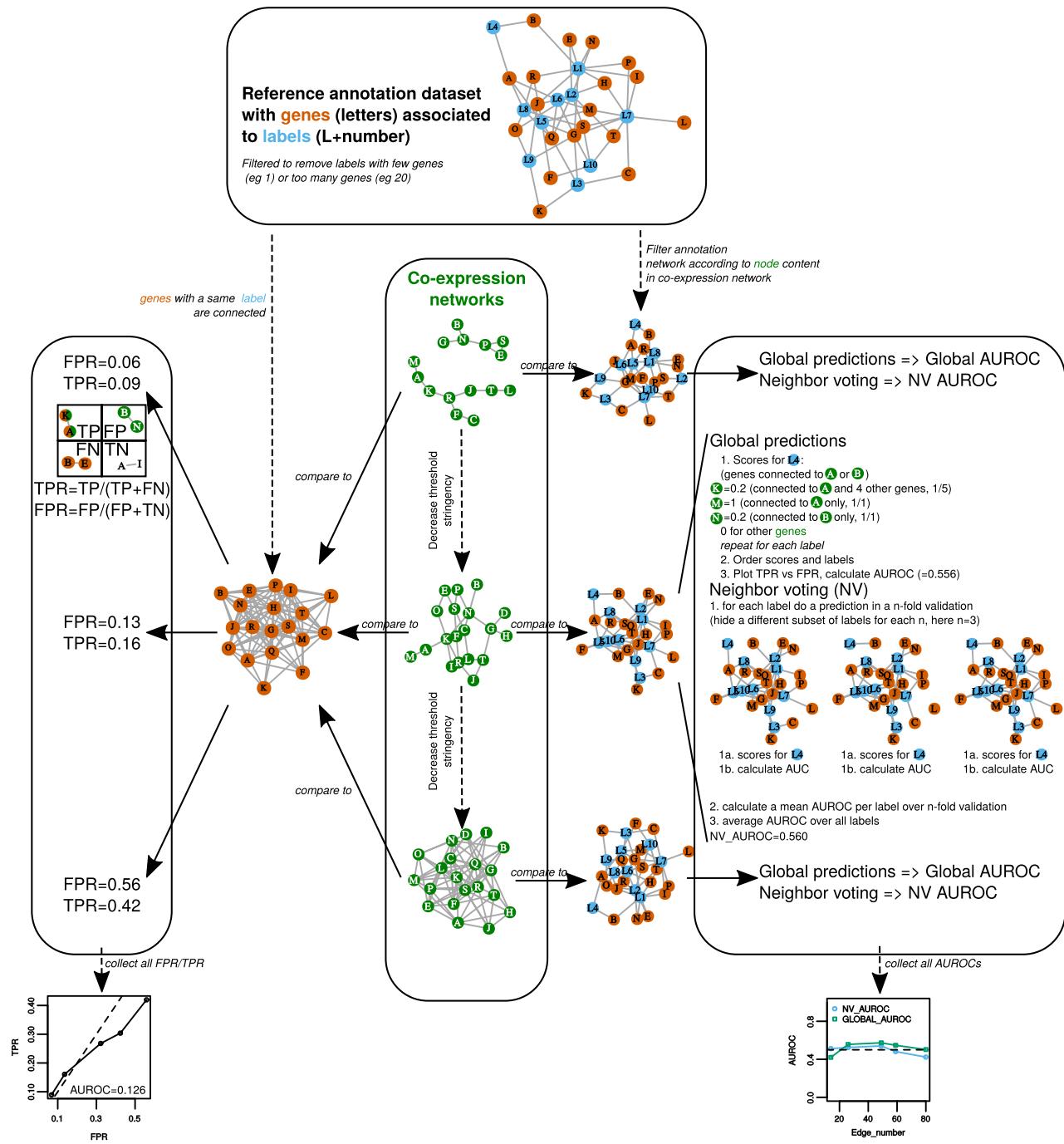


**Figure 1.** Workflow for global and targeted network analyses. One microarray dataset and a RNA-seq dataset prepared according to 7 normalization procedures were used to generate eight expression matrices analyzed with six different distance measurements (Pearson's or Spearman's Correlation Coefficient, unranked or ranked with HRR, Mutual Information (MI) or Partial Correlations (PC)) to obtain 48 distance matrices. Each of these matrices was thresholded to obtain global networks at different confidence thresholds. Global networks were evaluated and also queried with specific guide gene sets reflecting 5 different pathways in a process named Pathway Level Correlation (PLC). The resulting subnetworks were evaluated and used to construct co-occurrence networks between microarray and RNA-seq datasets. In white are indicated the figures corresponding to the different steps analyzed. Dataset × distance combinations are indicated in blue and characteristics that are improved by these combinations.

## Results

**Inferring global co-expression networks and comparing correlation measurements.** Large co-expression networks were obtained by varying the confidence threshold (correlation value above or rank below) within lists containing the 10 million best gene pairs from eight different datasets and six data measurement combinations (Fig. 1). Each of the 10 million best pair lists was filtered at different confidence thresholds (1, 5, 10, 20, 40, 60 or 80% best pairs from these lists) to evaluate the effect of network size on performance. Expression datasets included a microarray-based expression matrix and seven RNA-seq based expression matrices normalized with different methods to evaluate their effect on network inference: transcript per Million (TPM), log2 TPM, sample scaled (ss) TPM, ss log2 TPM, raw counts, variance stabilized transformed (VST) raw counts and VST-TPM. The six distance measurements were: raw PCC, raw SCC, PCC-HRR, SCC-HRR, PC and MI. Each network performance was considered as a network ability to capture edges corresponding to functional associations found in the GO reference dataset and was evaluated in 4 different ways (Fig. 2): GO term enrichment (GO terms that are significantly enriched with gene pairs from the co-expression network), a ROC curve constructed with TPR and FPR calculated for each confidence threshold and two ROC analyses based on the GBA concept, an average 3-fold cross validated neighbor voting (NV) AUROC and a global AUROC. AUROCs correspond to Area Under Receiver Operating Characteristic curves calculated for every network either from each GO (with three test sets obtained after hiding part of the gene labels, NV AUROC corresponding to the average of AUROCs for all GO terms) or the whole annotation dataset (global AUROC). AUROCs are used as global indicators of a dataset performance, a value of 0.5 indicating a random attribution of labels in the network and a value of 1 indicating a perfect match with the reference dataset. AUROC > 0.6 may be considered as moderate<sup>19</sup>. In global

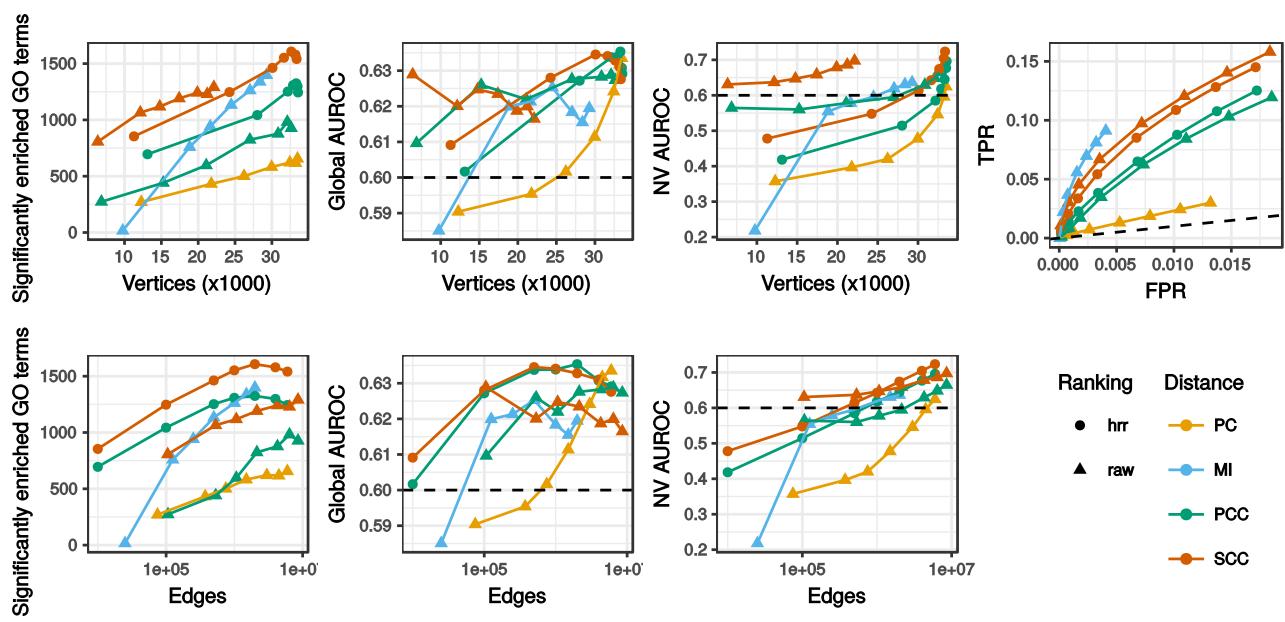




**Figure 2.** Network performance. This small example describe strategies to evaluate networks according to a reference functional annotation. Co-expression networks were obtained for each dataset  $\times$  distance measurement combination (Fig. 1) at different confidence thresholds, resulting in networks increasing in size with lower stringency. A total evaluation was made with True Positive Rate (TPR) vs False Positive Rate (FPR) analysis (left panel) by classifying edges as True positives (TP), False Positives (FP), False Negatives (FN) or True Negatives (TN). Single network evaluation was performed by calculating AUROCs with the EGAD R package, either as a global prediction or using a neighbor voting (NV) algorithm with a 3-fold cross validation (right panel). All indicated values are in accordance with the small networks in this example. In addition to these 3 evaluations (FPR vs TPR, global AUROC and NV AUROC), GO term significant enrichment was statistically tested with a hypergeometric distribution (not shown in this example).

TPR vs FPR curves, the line extending from (0,0) to (1,1) has an AUROC = 0.5 and points above this line indicate more predictive networks than a random selection (Fig. 2). The GO annotation table was filtered to perform these analyses by removing weakly represented or non-specific GO terms ( $> 5$  or  $< 100$  genes).



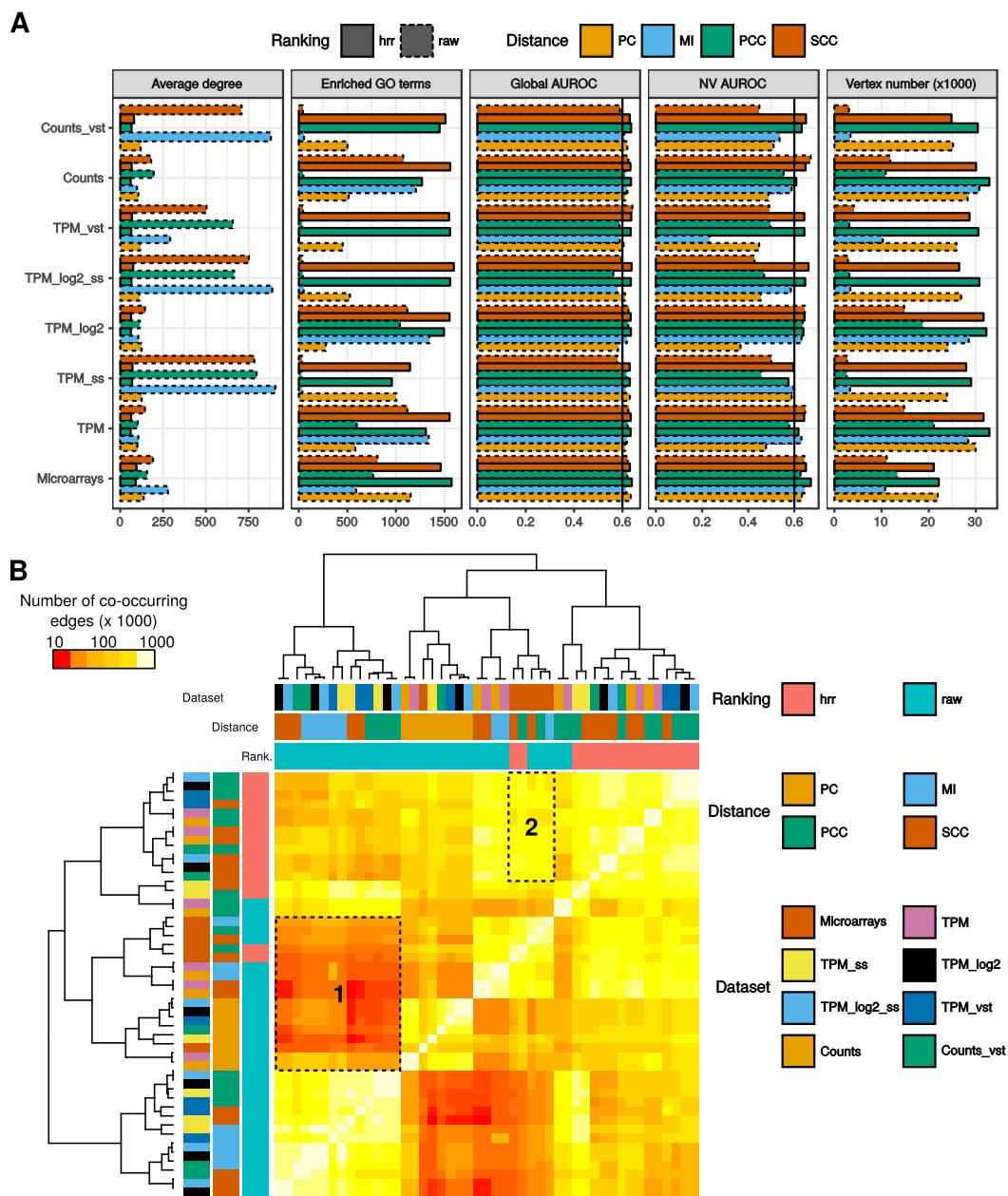


**Figure 3.** Global network characteristics. Only results for the RNA-seq TPM dataset without further normalization are shown. The horizontal dashed line indicates a 0.6 AUROC value taken as a threshold separating good and poor network predictability. In the TPR = f(FPR) panel, the dashed line corresponds to a random selection (with AUROC < 0.5). This panel is partial and the highest FPRs correspond to 10 million gene pairs.

Figure 3 displays TPM network evaluation at different confidence thresholds and Fig. 4 shows networks having 1 million of edges across all dataset  $\times$  distance combinations. Metrics for all other dataset  $\times$  distance measurement combinations are presented in Supplementary Fig. 1. All networks combined, pairwise correlations between enriched GO counts, global and NV AUROC performance metrics were moderate (Spearman's  $\rho > 0.4$ ) but significant ( $p < 0.001$ ) indicating these three performance metrics evaluated networks in different ways. The highest correlation was observed between NV AUROC and enriched GO counts ( $\rho = 0.70, p < 0.001$ ) showing their consistency. The NV AUROC was the most positively correlated with edge number ( $\rho = 0.55, p < 0.001$ ) suggesting that decreasing the confidence threshold and adding more edges in networks did not result in a significant increase in false positives. This was confirmed by the partial ROC curves (obtained for a maximum FPR at 10 million edges) drawn from the TPR and FPR (Fig. 3, Supplementary Fig. 1), where up to 10 million best pairs, TPR increased faster than FPR. Although counts of significantly enriched GO terms were positively correlated to NV AUROC, we observed a slight decline in the largest networks which might reveal a saturation in these enriched GO terms. It is possible that with the hypergeometric testing, some GO classes are fully enriched in smaller networks leading to a decrease in their significance as network size increases. The global AUROC displayed a very low variation (min = 0.55, average = 0.61, max = 0.68) and was significantly correlated to vertex number ( $\rho = 0.43, p < 0.001$ ) only. This observation suggests that the global AUROC is not an appropriated measure in our case.

At equivalent edge numbers, different distance measurements generated networks varying considerably in vertex number (Fig. 4A, Supplementary Fig. 1). Considering all datasets and distance measurements, raw PCC, raw SCC and MI resulted on average in fewer vertices and higher node degree (vertex number/node degree: 13,164/511, 9,986/465 and 14,074/468 respectively) than PCC-HRR, SCC-HRR or PC (26,645/116, 24,731/124 and 23,966/166 respectively). This trend was clearly observed when setting an edge number to 1 million (Fig. 4A). Expression networks constructed from microarrays, TPM, TPM log2, and counts displayed very similar ROC curves: PC based networks followed random predictions (NV AUROC = 0.5) and the other distance measurements were above the random prediction with similar AUC (Supplementary Fig. 1). This was confirmed for PC by NV AUROC and enriched GO term counts. Performance of the other distance measurements in the global TPR/FPR curves did not exactly match that measured with AUROCs. Taking the TPM dataset as an illustration (Fig. 3), the MI ROC curve was above the others while NV AUROC for similar edge numbers was slightly below that measured for SCC. This was probably due to differences in network topologies (see above) and the procedures underpinning the two evaluations. The global TPR/FPR curve does not measure a network predictability *per se* as NV AUROC does and considering any gene pair sharing a same GO term as valid could have overestimated TP (Fig. 2). As a general trend, raw PCC and raw SCC generated smaller networks than PCC-HRR and SCC-HRR but displayed similar TPR/FPR curves, *i.e.* for a similar performance, HRR-ranked CC networks had more vertices and fewer edges than raw CC based networks (Fig. 3). CC ranked with HRR always generated relevant networks for TPM ss, TPM log2 ss, TPM VST and counts VST, which was not the case for raw CC (Supplementary Fig. 1). These normalizations induced strong biases in CC distribution as revealed by thresholds used to obtain the 10 million best pairs (Supplementary Table 1) but these biases were compensated by HRR. For these four normalization methods, thresholds used to get the best co-expressed gene lists were  $> 0.9$ . Contrastingly, thresholds





**Figure 4.** Comparison of dataset  $\times$  distance measurement combinations for networks with a million gene pairs. Network topology and performance in GO recovery were analyzed (A). Vertical lines at 0.6 indicate AUROCs above which network predictability can be considered as moderate. Co-occurring edges were also counted in every possible comparison between 2 networks (B). Area 1 corresponds to RNA-seq networks having few genes in common with PC networks and microarrays networks and area 2 to combinations maximizing edge co-occurrence between microarray and RNA-seq.

were  $<0.7$  for the other methods (without VST or scaling) and it indicated a bias in CC distribution towards 1. However using ranked CCs with VST or scaling normalizations resulted in HRR thresholds in range with that obtained for the other normalization methods (around 1300, at the exception of TPM ss at 2200). In addition to this bias in CC values, VST or sample scaling normalization procedures were associated with higher node degree (average node degree of 712 vs 114 for other normalization procedures; Welch Two Sample t-test  $p < 5 \cdot 10^{-7}$ ) and lower performance (both in the number of significantly enriched GO terms, average 36 vs 1,145,  $p < 9 \cdot 10^{-12}$ , and NV AUROC, average of 0.47 vs 0.62,  $p < 0.001$ ) (Fig. 4A and Supplementary Fig. 1). These results clearly shows that network topologies were strongly impacted by VST and sample scaling procedures which resulted in less vertices that were more tightly connected together with a loss of GO term capture. Taken together, these results revealed that HRR CCs are able to generate complete genome-wide networks with good performances similar to other classical measures such a MI and PC. Node degree AUROC measures whether genes are more likely associated according to their number of connections rather than to their function. A positive correlation was found

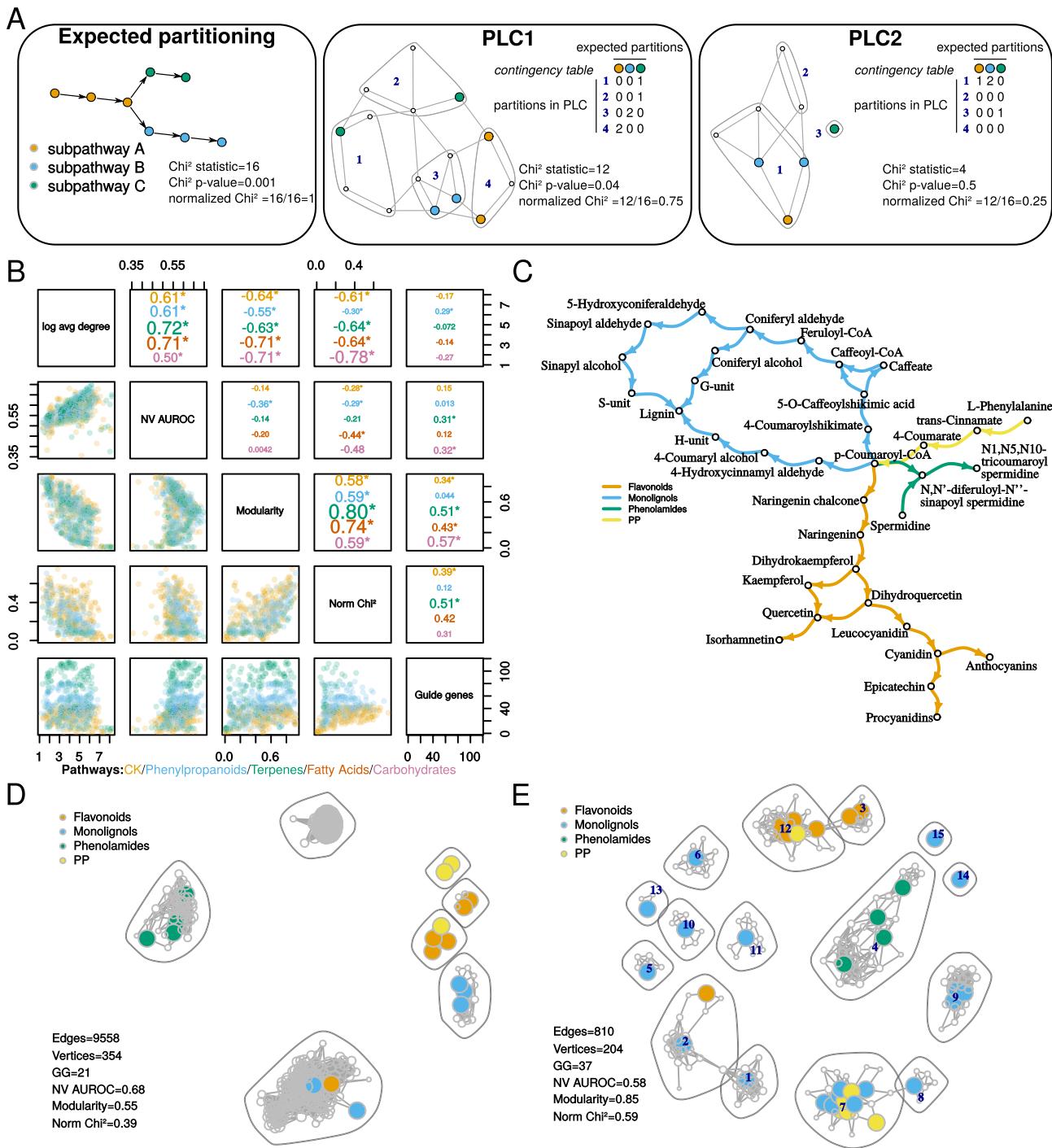


between NV AUROC and degree AUROC ( $\rho = 0.47, p < 2e-16$ ) indicating that highly predictive networks (NV AUROC  $> 0.7$ ) also had a higher node degree AUROC. Node degree AUROC was generally under 0.55. We therefore considered that in our conditions, this bias was only limited. Concerning edge co-occurrence between the different dataset  $\times$  distance combinations, the lowest conservation was observed with raw (MI, PCC and SCC) RNA-seq datasets and PC networks and microarrays networks (Fig. 4B, area 1). Interestingly, many edges were conserved between VST or sample scaled normalized datasets when correlations were calculated with raw PCC (4B; below block 1). This was probably due to their particular topologies (see above) having few vertices very tightly connected. More co-occurring edges were found when microarray networks were compared to RNA-seq networks obtained with CC-HRR (mean of 97,646 vs 25,277; Fig. 4B, area2). This indicated that microarrays and RNA-seq networks were more comparable when obtained with HRR, reinforcing their validity. The previous section focused on global network properties. Community detection procedures can be applied to such global networks to cluster tightly connected genes into modules. In our case, we rather used a knowledge-driven approach known as Pathway-Level Coexpression (PLC) to extract gene pairs associated within a given pathway (Supplementary Fig. 2). PLC are particularly interesting in plants for example to decipher incomplete specialized metabolic pathways. It aims at capturing a transcriptional landscape for genes known to be involved in a given pathway, in order to highlight their organization as well as finding new genes (transporters, transcription factors,...) associated with the process. In the next part, we evaluated the ability of all previous networks to capture relevant information associated with four metabolic and one signaling pathways. We selected two primary metabolic pathways (carbohydrate and fatty acid metabolisms), two specialized (secondary) pathways (phenylpropanoid and terpenoid metabolisms) and the cytokinin signaling pathway.

**Assessing PLC quality: trade-off between GO term representation and guide genes.** The PLC procedure is expected to cluster together guide genes with many co-expressed genes ('associated genes') and to reflect the subpathway organization (Fig. 5A). For PLC, we systematically removed all genes showing a degree value of 1 (*i.e.*, those connected to only one guide gene). However we included edges between associated genes if they were found among edges retained at the selected threshold. Using five pathways (Table 1, Fig. 5B, Supplementary Table 2 and Supplementary Fig. 3), we extracted five PLC from the global networks generated above to determine the best suitable dataset  $\times$  distance measurement combinations. All pathways have modular structures with gene sets forming specific sub-pathways (also called partitions or modules). We expected that PLC would be able to reconstruct such a partitioning, by connecting guide genes with associated genes. Pathways investigated here are typically reaction chains involving several enzymes working in coordination and having no direct effect on the transcription of their gene mates. As a consequence, specific sub-pathways are expected to form coordinated and specific correlation units highlighted within specific experimental conditions<sup>21</sup>. Concerning the signaling pathway, some of the query gene products could directly bind the promoter of other guide genes to control their expression<sup>22</sup>. Its study may clearly require a regulatory network approach but we investigated how co-expression network performed with such a signaling pathway. The phenylpropanoid pathway contains a core module composed of 3 genes leading to a precursor used by 3 other distinct subpathways<sup>23–26</sup> (Fig. 5B, Table 1, Supplementary Table 2). The three other metabolic pathways, carbohydrates, fatty acids and terpenoids, were structured in modules as described on the KEGG database<sup>27</sup> (Table 1, Supplementary Table 2). The fatty acid pathway contains 97 genes divided into 6 modules. The central carbohydrate metabolism contains 202 genes partitioned in 8 modules. Finally, the terpene pathway has 64 genes partitioned into 6 modules. Pathway organizations were used as indicated in the KEGG database (apart from phenylpropanoid pathway which was manually curated from our previous work) and compared to PLC subnetworks. The plant cytokinin (CK) pathway is known to regulate many processes in plant physiology and is hierarchically organized in three levels: a histidine kinase receptor, a transducer (histidine phosphotransfer proteins) and a response regulator (type A/B/C) which may act as a transcription factor<sup>28</sup> (Table 1, Supplementary Table 2). Although CK pathway members are relatively well known, each level is represented by several members which may have specific roles and it is still unclear how they biologically interact with each other to drive a specific physiological response. We expected that PLC would group some of these actors according to specific physiological responses. CK pathway includes both transcription activating and repressing activities (via response regulators) and post-transcriptional (phosphorylations) and would therefore be an excellent test of PLC applicability on associations expected to be more complex than in metabolic pathways. In addition, we included other histidine kinases integrating other signals and known to crosstalk with the CK pathway<sup>29</sup>. We therefore included 2 ethylene receptors, ETR1 and ERS1 to determine whether they could be clustered with CK histidine kinase. The initial pathway was not partitioned into sub-pathways but rather into 5 levels (receptor, transducer, type A/B/C response regulator) because interactions between specific actors of each level are not completely understood.

Subgraphs of global networks were constructed for each pathway by retrieving edges involving at least one guide gene and were partitioned into communities with a fast greedy algorithm designed to maximize network modularity and which has been shown to extract relevant communities from large networks<sup>30</sup>. We compared guide gene distribution in these communities to target subpathways using a normalized Chi<sup>2</sup> test which values range from 0 to 1, 1 being the expected partition and 0 a random partition of guide genes or very few guide genes (Fig. 5A). All networks having a Chi<sup>2</sup>  $p$ -value  $> 0.05$  were considered to have a Chi<sup>2</sup> statistic equal to 0. PLC performance in recovering GO terms was evaluated by counting significantly enriched GO terms and by calculating a NV AUROC for each network. A good PLC was expected to contain a large number of guide genes and to have both a good score in grouping them into expected partitions (high normalized Chi<sup>2</sup> value) and a good score in overall biologically relevant edge recovery (NV AUROC  $> 0.6$ ). We first analyzed correlations between all these metrics (NV AUROC, number of guide genes and Chi<sup>2</sup> statistic) together with two topological metrics (mean node degree and modularity), for each pathway separately (Fig. 5B). Strongest correlations were observed between NV AUROC and mean node degree ( $\rho > 0.5, p < 0.001$ ) and between modularity and





**Figure 5.** Trade-off in PLC subnetworks between performance in GO term recovery and partitioning guide genes into expected communities. (A) Example showing normalized Chi<sup>2</sup> statistic and *p*-value calculations comparing guide gene distribution into PLC communities (numbers in deep blue within polygons) to the expect partitioning (left; 3 subpathways). Two PLCs (one with a good partitioning (center); one with a weak partitioning (right)) are shown here but the contingency matrix used in Chi<sup>2</sup> calculations is described for only one of them (center). (B) Pair plot showing correlations (Spearman's rho, asterisks show significance  $p < 0.001$ , upper panel) and scatterplots (lower panel) between average network node degree, NV AUROC, normalized Chi<sup>2</sup>, modularity and the number of guide genes in the network. Each point in the lower panels (scatterplots) represent one network for which 2 characteristics (eg NV AUROC and modularity) are compared. Data are presented for each pathway separately with a specific color. (C) The expected partitioning of phenylpropanoid related guide genes was compared to two PLC: (D) higher predictability and lower modularity (microarrays raw PCC) and (E) lower predictability and higher modularity (microarrays PCC-HRR). In (D and E), colored vertices correspond to genes encoding enzymes catalyzing steps of similar color in (C). Community (surrounded by grey polygons) numbers in (E) are indicated in deep blue and can be used to access Supplementary Table 3.



Pathway	Genes	Number of subpathways	Subpathway names (KEGG module accession)
Phenylpropanoids	43	4	core phenylpropanoid (PP), flavonoids, monolignols, phenolamides
Fatty acid	97	6	fatty acid biosynthesis (initiation (M00082), elongation (M00083), its ER-localized part (M00415)), jasmonic acid phytihormone biosynthesis (M00113) and $\beta$ -oxidation (M00086 and M00087)
Carbohydrate	202	8	glycolysis (Embden-Meyerhof pathway (M00001) and the core module involving three-carbon compounds (M00002)), neoglucogenesis (M00003), pyruvate oxidation (M00307), citrate cycle (M00010), pentose phosphate pathway (M00004, M00006 and M00007)
Terpenes	64	6	mevalonate (M00095), methylerythritol (M00096), C10-C20 isoprenoid (M00366), beta-carotene (M00097), abscisic acid hormone (M00372) and phytosterol (M00371) biosynthetic blocks
Cytokinin signaling	37	?	?

**Table 1.** Pathway description. ?Indicates that partition in sub-pathway is not known.

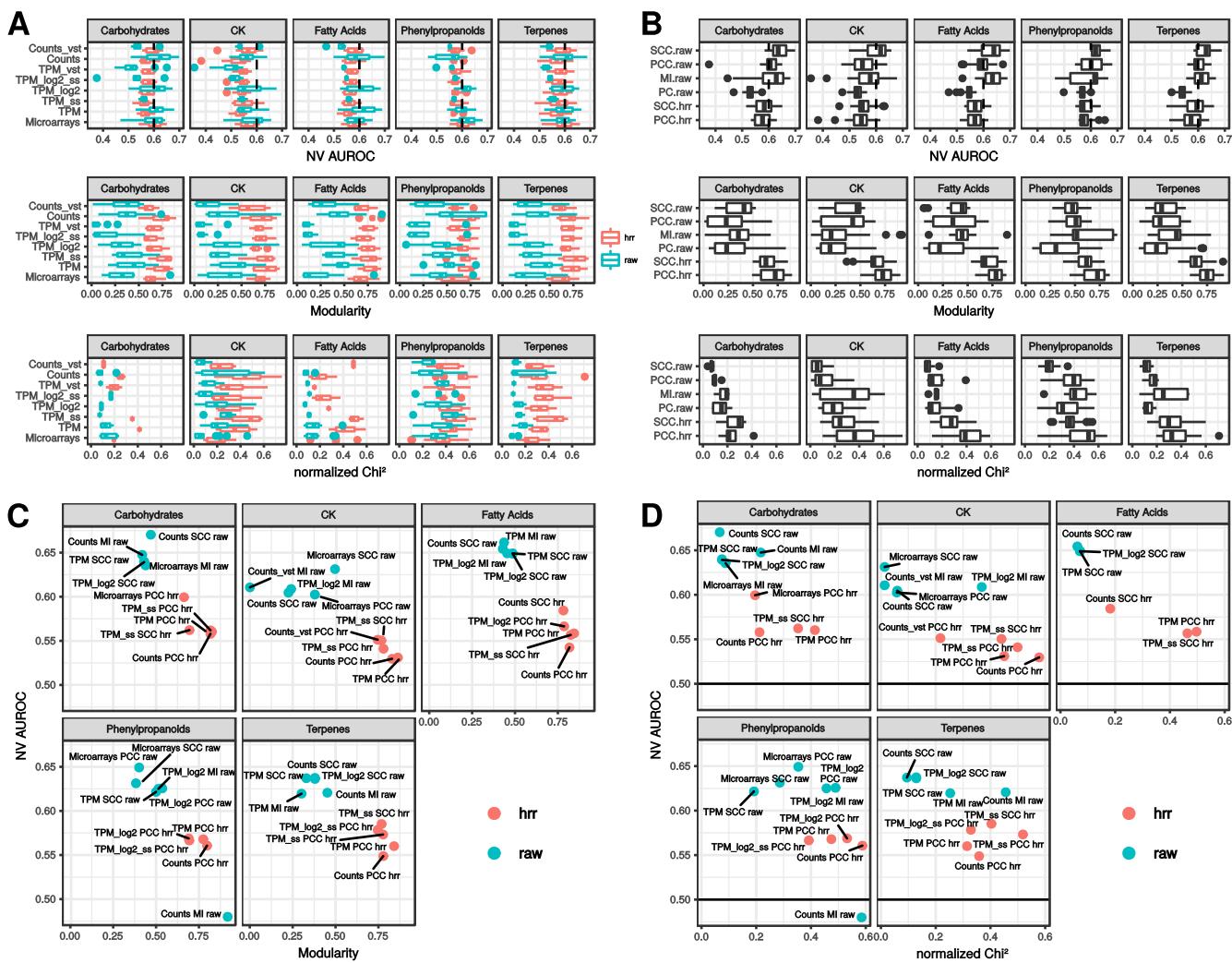
normalized Chi<sup>2</sup> ( $\rho > 0.59$ ,  $p < 0.001$ ). We found that PLC performance (NV AUROC) was almost negatively correlated with normalized Chi<sup>2</sup> ( $\rho < -0.2$ ) indicating that guide genes were clustered correctly at the expense of capturing GO associated gene pairs. Given the CK pathway structure, partitioning based on protein functions (receptor, transducer or response regulator) did not resulted in high Chi<sup>2</sup> values, suggesting that partitions in the co-expression networks contained guide genes from different levels, reinforcing the existence of specific sub-pathways. These results indicated a trade-off in PLC between edge quality and guide gene partitioning. A visual examination of PLC with either lower modularity and higher NV AUROC (Fig. 5D) or higher modularity and lower NV AUROC (Fig. 5E) revealed that PLC with higher modularity as well as higher Chi<sup>2</sup> values displayed a biologically relevant organization. Such subgraphs had generally a lower average node degree and a higher representation of guide genes rendering their analysis more convenient. Taking the phenylpropanoid pathway as an example, the PCC-HRR based TPM network (Fig. 5E, with a higher modularity) correctly clustered genes from the core phenylpropanoid (PP) and the flavonoid modules while the raw PCC network did not (Fig. 5D, with a higher NV AUROC). Similar results were observed with the four other pathways with either microarray or RNA-seq datasets (Supplementary Fig. 3). Modularity and normalized Chi<sup>2</sup> could therefore be considered as consistent quality metrics for PLC. NV AUROC should also be considered to ensure that subgraphs had a minimum predictability ( $> 0.55$ ).

**HRR-CCs optimize recovery and clustering of guide genes in PLC.** The best performing dataset  $\times$  distance measurement combinations were searched by analyzing NV AUROC, modularity and normalized Chi<sup>2</sup> among networks with a Chi<sup>2</sup>  $p < 0.05$ . Statistical effects of dataset, distance, ranking and their interactions on subgraph characteristics were analyzed by ANOVA for each pathway. Ranking and distance measurements had generally the strongest effects on modularity and normalized Chi<sup>2</sup> ( $p < 2e-5$ ) (Fig. 6A,B). Ranking had a significant effect on NV AUROC ( $p < 0.01$ ) but was weaker than distance measurement ( $p < 1e-4$ ). Datasets only had a significant effect on modularity ( $p < 0.002$ ). Significant interactions were rarely observed between these three factors (*i.e.* in few pathways and with a weak effect). This revealed that the different RNA-seq normalizations had only minor effects on these PLCs. Taken as a whole, networks obtained with raw datasets had a significant higher NV AUROC (t-test, mean in raw = 0.58, mean in HRR = 0.57,  $p < 0.01$ ) but significant lower modularity (mean in raw = 0.35, mean in HRR = 0.68,  $p < 2.2e-16$ ) and lower normalized Chi<sup>2</sup> value (mean in raw = 0.20, mean in HRR = 0.35,  $p < 2.2e-16$ ) (Fig. 6A). It therefore appeared that clustering guide genes correctly was improved with CC ranked with HRR at the expense of performance. NV AUROCs in HRR-based networks were generally higher than 0.55, indicating an average low performance in GO capture (Fig. 6A). In non-ranked distances, PC resulted in the weakest NV AUROC, while MI and raw SCC based networks displayed the highest NV AUROC (Fig. 6B). This weakness in PC based networks was compensated neither by a higher modularity nor by a higher normalized Chi<sup>2</sup> statistic.

A more detailed examination of best PLC subgraphs maximizing either modularity or NV AUROC, revealed that each of the five pathways involved specific dataset  $\times$  distance measurement combinations. PCC-HRR based networks were always found to maximize modularity (Fig. 6C) and normalized Chi<sup>2</sup> (Fig. 6D) with almost all datasets. Raw distance based PLCs had a higher NV AUROC and some of them also had a good modularity but they also had a lower normalized Chi<sup>2</sup> statistic indicating they contained fewer guide genes (*e.g.* raw RNA-seq counts with raw SCC in the terpene PLC). The results suggest that PCC-HRR could be used as a reliable distance measurement whatever the dataset. Careful analysis of PLC obtained from PCC-HRR revealed the presence of relevant associations in each PLC (Supplementary Fig. 3 and Table 3). For example, community 12 from the phenylpropanoid PLC obtained with microarray data processes with PCC-HRR (Fig. 5D) contained AT1G06000 encoding a Flavonol 7-O-rhamnosyltransferase and was clearly associated with other genes from the flavonoid sub-pathway. This gene was not detected in the raw PCC PLC (Fig. 5C). Other examples are highlighted in yellow in Supplementary Table 3.

**Vertex and edge co-occurrence in microarray and RNA-seq based PLC subgraphs.** Edge co-occurrence in networks constructed from expression datasets obtained by different technologies may be considered as a further validation. Quantifying gene expression with microarrays relies on probe hybridization by sequence complementary while with RNA-seq, short reads are mapped back *in silico* to the reference transcriptome. The two main differences





**Figure 6.** PLC subnetwork performance. Performance in capturing GO terms (NV AUROC), modularity and normalized Chi<sup>2</sup> value distribution in interactions between datasets and ranking methods (**A**) and between distance measurement and ranking methods (**B**) showing the dominant effect of the ranking procedure (raw vs HRR) on these metrics. (**C**) Modularity and NV AUROC of the five top NV AUROC networks and 5 top modularity networks. (**D**) Normalized Chi<sup>2</sup> statistic and NV AUROC for the same networks.

between these technologies are (i) the number of quantified transcripts (due to the completion of genome annotation) and (ii) the dynamic range (fluorescent probe intensities for microarrays, *in silico* read counts for RNA-seq). Because microarrays and RNA-seq technologies differ, edges co-occurring in networks obtained from these two technologies are probably more relevant. In Fig. 4C, we analyzed co-occurrence in global networks and found that HRR ranked CCs apparently increased the number of co-occurring edges between microarrays and RNA-seq. To get more insights into co-occurrence in PLCs, common edges and vertices were counted in pairwise intersections of networks (RNA-seq vs microarrays) obtained with the six distance measurements and set at a 1,000 vertices. The resulting intersection networks were further characterized by the number of represented guide genes, their normalized Chi<sup>2</sup> statistic, modularity and NV AUROC. This evaluation was performed with the RNA-seq dataset expressed as TPM only because we showed in the previous section that normalization methods had a minor impact on PLC. In addition, TPM networks with raw distance methods had enough vertices to correctly extract PLC (it was not the case with raw distances, e.g. for TPM normalized with VST as revealed by their very low normalized Chi<sup>2</sup> statistics; Fig. 6A).

Many more co-occurring edges were generally recovered when raw CC and MI networks were compared (e.g. 18,334 averaged over the five pathways with MI networks vs 550 with PCC-HRR networks; Supplementary Fig. 4). At a 1,000 vertices, all raw networks but PC contained more edges (221,297 and 85,059 in average for microarrays and TPM) than HRR-CCs networks (12,431 and 12,877). This might have resulted in more co-occurrences between MI networks. PC networks had the lowest number of co-occurring vertices (94 in average) but intersections from MI and/or raw CC had comparable vertex number (268) to intersection networks from CC-HRR (267 in average) (Supplementary Fig. 4). These results suggest that HRR-based networks have strong overlaps. Intersections of PCC-HRR subgraphs were able to maximize the % of guide genes (mean of 75% over the 5 PLC),



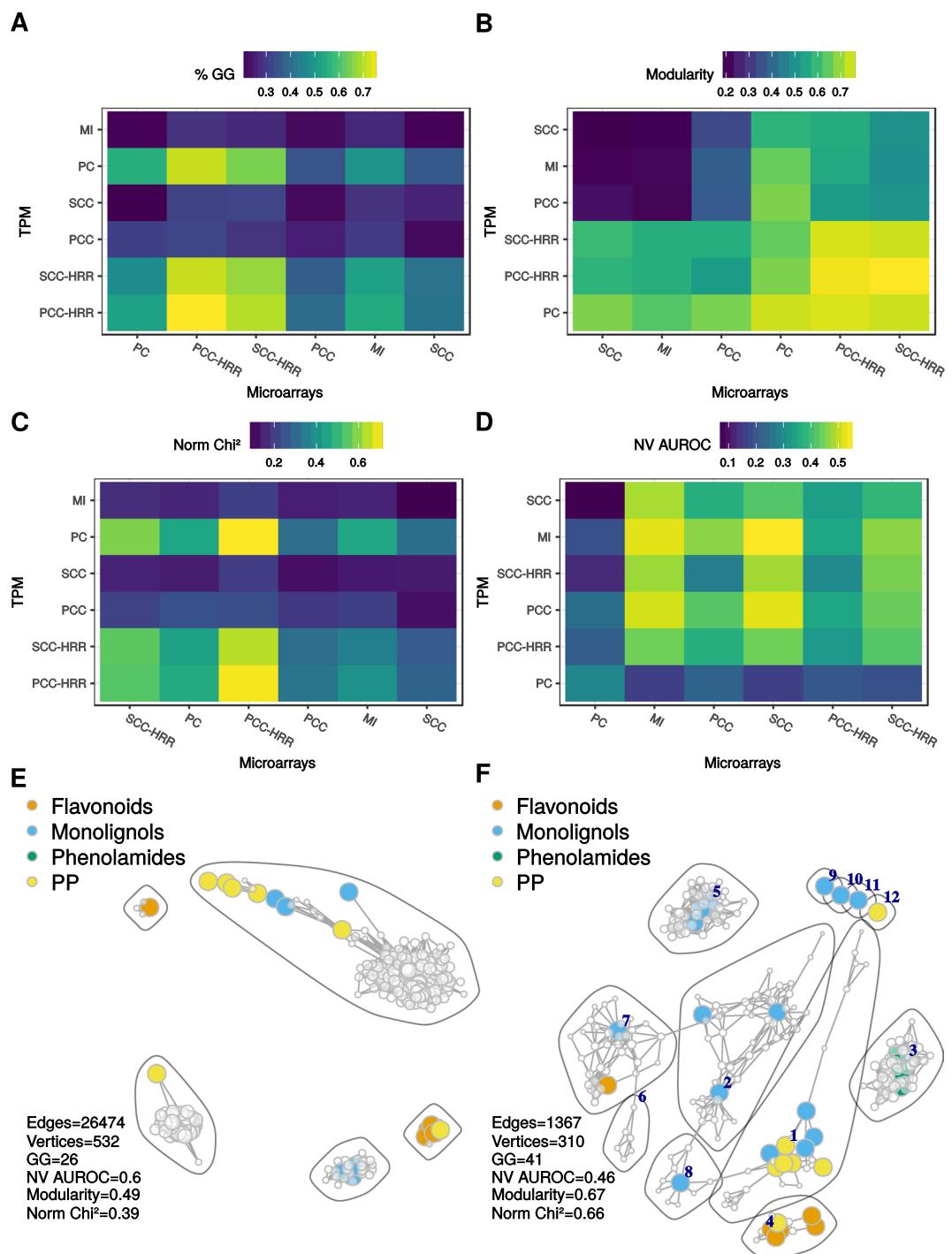
modularity (0.78) and normalized Chi<sup>2</sup> statistic (0.70) (Fig. 7). Detailed characteristics for each PLC are presented in Supplementary Fig. 4. Modularity was generally high in the intersection between CC-HRR networks ( $>0.70$ ) but intersections with SCC-HRR displayed lower normalized Chi<sup>2</sup> values ( $<0.6$ ). Intersection network performance in recovering GO terms was globally low (Fig. 7D). The highest NV AUROCs were observed in intersections between MI networks (0.52), MI (microarrays)–raw SCC (TPM)(0.54) and raw PCC (microarrays)–raw SCC (TPM)(0.52) (Fig. 7D). Intersection networks and their contents are available in Supplementary Fig. 5 and Supplementary Table 4. Again, we found candidate genes not included in the guide gene sets that were correctly associated with other guide genes (highlighted in yellow in Supplementary Table 4). Taking the phenylpropanoid pathway as an example, Fig. 7E shows edge and vertex co-occurrence between MI networks and Fig. 7F between PCC-HRR networks. The co-occurrence network obtained from MI contained fewer guide genes (26 vs 41) and displayed lower modularity (0.49 vs 0.67) and normalized Chi<sup>2</sup> statistic (0.39 vs 0.66). Although it had a higher NV AUROC (0.6 vs 0.46), its structure did not reflect that of the expected pathway (Fig. 5C). For example, phenolamide related genes were not represented. Average guide gene degree (33) was below the average degree of the remaining nodes (100) indicating that guide genes were only slightly connected to other genes in this co-occurrence network from MI PLC. By contrast, guide gene degree (11.4) was very similar to the other node degree (11.1) revealing an uniform integration of guide genes with other genes in the co-occurrence network of PCC-HRR PLCs. As observed in co-occurrence in large networks (Fig. 4B), RNA-seq TPM normalized with VST had slightly more edges in common with microarray networks. We therefore compared PCC-HRR PLC between microarrays and RNA-seq TPM normalized with VST. Intersection networks had very similar characteristics to that observed between microarrays and RNA-seq TPM. Although it contained slightly more co-occurring vertices and edges in average (360 and 1,252 respectively with TPM VST vs 240 and 550 with TPM), it displayed fewer guide genes (54 vs 57). TPM normalized with VST could therefore be an interesting alternative to TPM. PLC intersection networks and their description are available in Supplementary Fig. 6 and Table 5.

## Discussion

Pathway Level-Correlation (PLC) is an interesting approach to capture biologically relevant transcriptional relationships using guide genes (e.g. genes involved in a same metabolic pathway) from transcriptome-wide co-expression networks. Our present work highlights that distances between genes calculated with highest reciprocally ranked PCC (PCC-HRR) improve PLC. The main improvement was guide gene representation. PCC-HRR based PLCs contained more guide genes than observed with other distances and they were generally more correctly partitioned into expected sub-pathways in the co-expression network. This was associated with a lower mean node degree and a higher modularity but also with a slightly weaker performance in GO term recovery. Our results propose that modularity and normalized Chi<sup>2</sup> values could be used as reliable indicators of PLC quality. We also observed that edge and vertex co-occurrences in PLCs obtained with PCC-HRR and microarray and RNA-seq TPM data can be used to construct relevant networks. A surprising observation was that in our conditions, for most combinations tested, true positive rates remained higher than false positive rates in spite of increasing network sizes. A similar trend using small *E. coli* and *S. cerevisiae* networks ( $<110$  nodes) has been previously observed with CCs<sup>8</sup>. However increasing network sizes inherently increases the number of false positive associations and it is clear that TPR is likely to increase at a slower rate than FPR above a given threshold. Thresholding is a very complex procedure which deeply impact network topology (Couto *et al.* 2017). The user should probably focus on a threshold that maximizes GO term capture while keeping the total number of edges as low as possible to avoid too strong an increase in FPR. The resulting FP which may have different origins: (i) true FP corresponding to apparent direct associations between two genes because they are both more correlated to a third gene (as exemplified in De la Fuente *et al.* 2004) or (ii) yet unknown true associations. In all cases, functional experimentation is required to validate or reject an association. This suggests that co-expression studies should test different confidence thresholds to efficiently capture gene associations. Evaluating network quality was done in respect of the Arabidopsis reference GO annotation set. We found that the NV AUROC<sup>19</sup> evaluates networks efficiently and was generally in accordance with significantly enriched GO term counts and TPR vs FPR curves. NV AUROC has the advantage of being a more global measure of predictability (values above 0.6 can be considered as moderate). Different distance measurements displayed different efficiencies according to the dataset but as a general trend, performance of the different combinations were similar (e.g. between microarrays and RNA-seq TPM in Fig. 3B). The same performance was obtained for different topologies: high node degree (more edges and fewer vertices) for MI and raw CC networks vs lower node degree (fewer edges and more vertices) for CC-HRR networks. PC networks displayed a high performance with microarray data only, suggesting that PCs calculated with ‘corpcor’ R package may not be recommended for RNA-seq data. A recent study has focused on metabolic pathways in plants using mutual ranks, another CC ranking method<sup>16</sup>. Complementary to this previous work, we found that ranking CCs increases vertex number without penalizing absolute network performance. Contrastingly, an opposite trend was observed in another study<sup>31</sup>, where larger networks displayed a lower Matthew Coefficient when compared to protein-protein interactions or regulatory networks. This indicates that different absolute performance measurements lead to different results and interpretations but this might also be due to our datasets which were larger than theirs. Another advantage of CC-HRR was that it clearly homogenized network characteristics from differently normalized RNA-seq datasets in addition to increase the number of co-occurring edges between microarrays and RNA-seq. The improved performance of HRR may be explained by its ability to integrate the whole transcriptional landscape of each gene thanks to the reciprocal rank calculation. In addition to normalize biases observed in raw CC values when using different data normalization procedures<sup>15</sup>, it probably added robustness to correlations because a low HRR value for a gene pair is observed only if the two genes had no higher correlations with other genes and this probably correct for noise in the expression dataset.

As revealed recently<sup>32</sup>, highlighting correlations between genes may require specific data processing or distance algorithms best suited to their query pathway. We also found that each of the five PLCs performed best





**Figure 7.** Characteristics of co-occurrence networks between microarrays and RNA-seq TPM. Percentage of guide genes (GG; A), modularity (B), normalized Chi<sup>2</sup> statistic (agreement with guide gene partitioning, (C) and NV AUROC (GO term performance, (D) were averaged over the 5 PLCs. Labels are ordered according to a hierarchical clustering. Co-occurrence networks obtained from phenylpropanoid PLC obtained with MI (E) or PCC-HRR (F). GG corresponds to guide gene number in the networks. Community numbers in F are indicated in deep blue and can be used to access Supplementary Table 4.

with specific RNA-seq normalizations (Fig. 6C,D) but RNA-seq TPM processed with PCC-HRR always provided informative networks which can be used as reliable starting point because they matched well expected pathway structure. In our case, the different data normalizations had a relatively weak effect on PLC characteristics especially when CCs were used with HRR. In a comparative analysis<sup>31</sup>, the authors have shown that PCC networks from VST normalized counts were more comparable to those from microarrays. In our case, VST normalization slightly improved the overlap between RNA-seq TPM and microarrays both at the global and targeted levels. This normalization can thus be further considered for co-expression studies. A fast greedy approach maximizing



modularity was used to detect communities within PLC subgraphs. Guide gene partitioning in these communities was compared to expected partitions in subpathways with a normalized Chi<sup>2</sup> test (Fig. 5A). We found that correct guide gene partitioning was negatively correlated with NV AUROCs but positively with modularity. Subnetworks with highest NV AUROCs but lower modularity such as those obtained with MI represented fewer guide genes and displayed large edge numbers. In these networks, guide genes formed inappropriate structures (Supplementary Fig. 5). We applied PLC to five pathways varying in size and nature. For the four metabolic pathways, PLC extracted from PCC-HRR based networks were able to cluster guide genes in the proper subpathways (Supplementary Fig. 5). Guide genes were associated in communities resembling subpathways and containing genes not included in the query gene set but known to be involved with the given pathway or being good candidates to be functionally validated (Supplementary Table 4). A similar PLC approach has been recently performed<sup>16</sup> using the Arabidopsis aliphatic glucosinolate pathway. In this previous work, the authors have successfully reconstructed this pathway and identified a new candidate glucosyltransferase that could be part of it. This demonstrated again that PLC is a powerful approach to complete biological pathways. When tested with a signaling pathway, we found that PLCs also displayed meaningful communities. For example, the CK signaling pathway is physiologically well known but its organization at the molecular level is far from being understood<sup>28</sup>. In particular, it is unclear how multi-family members of each signaling level (receptor, transducer and response regulator) interact with each other to drive a specific physiological response. In the PLC dedicated to the CK signaling pathway, PCC-HRR with microarrays suggested preferential transcriptional associations that have been described in the literature<sup>33</sup>. For example AHP2, AHP3 and AHP5 were grouped in the same module (module 7 Supplementary Fig. 5 and Table 4). These three AHPs have been reported to negatively regulate tolerance to abiotic stress<sup>34</sup>. The same community also contained AHK3, ARR1 and ARR2. Those three members are known to regulate primary root meristem activity and senescence<sup>35</sup>. AHK4, AHK2 and ARR14 which have been shown to regulate shoot apical meristem activity were grouped in the same community<sup>34</sup>. In addition, we saw clear associations between ETR1 and AHK3 in individual PLC subgraphs. Such association highlights crosstalk already known between CK and ethylene signaling pathways<sup>29</sup>. The co-occurrence pathway was relatively sparse in contrast to the metabolic pathways (Supplementary Fig. 5). It is possible that vertex number for this analysis (1,000) might have been too small to capture complex associations within this signaling pathway. Using VST normalized TPM increased edge and vertex number in the co-occurrence network (Supplementary Fig. 6 and Table 5). The above-described associations were also found in this co-occurrence network. While effective in revealing strong gene associations, merging PLC from microarray and RNA-seq data could miss other relevant associations. First, experimental conditions represented by each starting dataset are not completely overlapping. Together with inherent differences due to dynamic range, this leads to networks with very different edge compositions and node degrees<sup>19</sup>, explaining the relative weak overlap between networks. Second, RNA-seq expression data include genes that are not included in the GPL198 microarray. As an example, some important genes in aliphatic glucosinolate biosynthesis were not represented in a previous Arabidopsis microarray dataset but found in RNA-seq expression matrices from other related species<sup>16</sup>.

To capture transcriptional environment of a query gene list, distance calculations have to be performed on the whole transcriptome. Calculating partial correlations was particularly challenging but using a covariance shrinkage estimator worked well in terms of computing performance. It took less than 2 h for RNA-seq expression matrices but more than 12 h for the microarray dataset. By contrast, our program which is freely available at (<https://github.com/EA2106-Universite-Francois-Rabelais/Expression-network-analysis>) was able to calculate PCC-HRR in less than 3 h for both datasets. As PCC-HRR resulted in relevant networks, this tool can be useful for further studies requiring many computations such as analyzing sample size impact on PLC or testing other normalization methods.

The present work demonstrates that Pearson's Correlation Coefficients (PCC) on which highest reciprocal ranking (HRR) was applied can be used to construct reliable global and targeted networks. When considering Pathway Level Correlation (PLC) with a set of guide genes, three reliable measures can be used for evaluation: NV AUROC as a global indicator of GO recovery (expecting values > 0.5), modularity (between 0 and 1, 1 being the best network partition) and normalized Chi statistic (between 0 and 1, 1 indicating a perfect match with an expected partition). Clustering guide genes correctly was at the expense of capturing GO terms and dataset × distance measurement combination should be carefully selected to construct reliable PLC. Although specific RNA-seq data normalizations may be adapted to each pathway of interest, using TPM with PCC-HRR generated accurate and safe PLC. Using PCC with HRR also increased the quality of co-occurrence networks between RNA-seq and microarrays.

## Methods

**Microarray data preparation.** Experiment accessions (GSE) for GPL198 (Arabidopsis ATH1, 22,746 genes) were retrieved from ArrayExpress (Supplementary Table 6). Signal intensities per probe were generated with R<sup>36</sup> using the 'arrayexpress' package<sup>35</sup>. The function 'getAE' was used to convert the raw signal CEL files. Array normalization was performed per GSE using the 'justRMA' function of the 'affy' package. This procedure applies a background correction together with a quantile normalization to correct for biases within arrays and finally returns log2-transformed corrected signal intensities. All 10,095 arrays were combined into a single file and subjected to a quality control based on upper quartile dispersion (75%) and Kolmogorov-Smirnov statistical testing for outliers using an empirical cumulative distribution function as described previously<sup>37</sup>. A total of 142 arrays were considered outliers in the two tests and discarded from the final matrix. Each array was finally centered and scaled individually.

**RNA-seq data preparation.** 2,549 RNA-seq accessions obtained for *A. thaliana* were retrieved from ArrayExpress. Fastq files were obtained from the SRA after converting.sra files with the SRA ToolKit function



'fastq-dump' with the - split-files option for paired-end sequencing runs. Reads were systematically trimmed with Trimmomatic using adapter files according to the Illumina platform used for the runs<sup>38</sup>. Trimmed reads were pseudo-aligned to predicted transcripts from the representative gene models of Arabidopsis TAIR genome v10 (33,604 transcripts) with Salmon v0.7.2 using the variational Bayesian EM algorithm mode to improve abundance estimation<sup>39</sup>. Only samples displaying a mapping rate of reads >30% were kept, resulting in a final matrix containing 1,676 samples (Supplementary Table 6). RNA-seq counts were used as non-normalized raw counts or expressed as Transcript per Million to correct for sequencing depth. Normalization by Variance Stabilizing Transformation (VST) was performed with the DESeq2 R package. This normalization method aims at limiting the variance dependence to the mean<sup>40</sup>.

**Distance calculations.** Before calculations, zero-variance genes were discarded. CCs (Pearson or Spearman) are computationally intensive particularly in the case of large matrices. Highest Reciprocal Ranking (HRR) of CCs for genes A and B is calculated as  $\max(\text{rank(CC(A,B))}, \text{rank(CC(B,A))})$ . For each gene, all CC values are first transformed as ranks, with 0 corresponding to the gene rank against itself. Ranks are subsequently compared and the highest value is retained for each gene pair. We developed a tool written in C allowing the easy parallelization of these computations. Briefly, for a given initial matrix containing  $n$  genes and  $p$  samples, the number of cores  $c$  allocated is used to split the dataset into  $n/c$  submatrices. In case of non-integer value, the last line of the matrix is replicated (without incidence on PCC or rank values) so that  $n/c$  is an integer. PCC or HRR are then calculated for each gene pair using communication between CPUs with Message Passing Interface. The program delivers  $c$  files containing  $n/c \times n$  values corresponding to PCC or HRR. This program is freely available on Github (<https://github.com/EA2106-Universite-Francois-Rabelais/Expression-network-analysis>). To calculate SCCs, expression values were first ranked in R. Mutual information (MI) which is reported to better capture non linear relationships<sup>9</sup> were calculated with the 'knni.all' function of the Parmigene R package<sup>41</sup>. This function estimates MI using a  $k$ -nearest neighbor. Partial correlations were challenging to compute on genome scale expression matrices. Partial correlations are usually calculated from multiple linear regressions or by inverting the correlation matrix and used in Graphical Gaussian Models<sup>42</sup>. Our expression matrices had many more variables (genes) than samples therefore regression methods would have required a Lasso or Ridge penalization to estimate coefficients. However, this procedure generally leads to memory errors when considering more than 30,000 variables. We found that the most computationally appropriate method in our case was to estimate shrinkages of partial correlations with the R package 'pcorcor' (<http://strimmerlab.org/software/pcorcor/>). This package is maintained by Korbinian Strimmer's team<sup>43,44</sup>. We used 'pcor.shrink' function which relies on the inversion of the shrunken estimated covariance matrix to estimate partial correlations and which is suited for matrices with more genes than samples.

**Reference dataset.** We used the Arabidopsis Gene Ontology (GO) standard dataset to assess network quality. The annotation file provided by the AGRIGO database<sup>45</sup> and was filtered out to remove all terms with a IEA evidence code and keep only functionally attributed terms. We also removed GO terms represented by less than 5 genes or more than 100 to remove non-specific terms.

**Global Network analysis. Construction:** For each dataset  $\times$  distance combination, we dynamically set a threshold to obtain arbitrary lists of 10 million best gene pairs (with CC above or HRR below that threshold), *i.e.* less than 2% of the total possible edges. Networks were then constructed with the 1, 5, 10, 20, 40, 60 or 80% best pairs from these lists. Thresholds used to get the 10 million gene pairs are reported in Table S2. Global networks were analyzed as adjacency matrices in R. **Network characteristics:** besides classical topological characteristics such as vertex and edge numbers and mean node degree (the average number of connections for each vertex), we evaluated network quality by comparison with the reference dataset (Fig. 2). In a first approach, we built a confusion matrix by classifying edges as false or true positives, considering edges as valid if both genes were annotated with at least one same GO term. In this confusion matrix, true positives (TP) corresponded to gene pairs also found in the GO annotation, false positive (FP) to genes associated in the network but not in the GO annotation, false negatives (FN) to pairs in the GO annotation not predicted in the network and finally true negatives (TN) genes pairs not predicted in the network and the annotation table. This confusion matrix was used to calculate True Positive Rates (TPR) and False Positive Rates (FPR). TPR and FPR were obtained at various confidence thresholds (*i.e.* for networks differing in sizes) and used to draw a TPR vs FPR curve as described elsewhere<sup>15</sup>. These curves were only partial because we included only the first 10 million best pairs. This was useful to pinpoint the importance of low FPR<sup>46</sup>. In the second and third approaches, we relied on the guilt-by-association principle to estimate network predictability. In the second method, we used the 'predictions' function of EGAD R package<sup>47</sup>. For each gene, this function counts the number of connected genes annotated with an identical GO term and divides this count by the gene's degree. These scores are next ordered decreasingly to construct a TPR vs FPR curve for each network. It differs from the first approach described above because here TPR and FPR are not obtained from different confidence thresholds (and from different networks) but from all possible true positive and false positive edges in the current network. A global Area Under Receiver Operating Characteristic (global AUROC) was calculated from each of these TPR/FPR curves. In the third method, predictability was evaluated using a neighbor voting (NV) algorithm. In this case, an AUROC is calculated for each GO term from the ability of genes to predict the GO annotation of their direct neighbors in a 3-fold cross-validation<sup>19,48</sup>. A mean NV AUROC was calculated for each network. In addition to ROC analysis, we counted GO terms that were significantly enriched with gene pairs using a hypergeometric test with R.

**Pathway Level Correlation. Construction:** In PLC, subnetworks were constructed from global networks (see above) by keeping edges connecting at least one guide gene. Guide gene lists are indicated in Supplementary



Table 2. The R package ‘igraph’<sup>49</sup> v1.0.1 was used to construct and visualize these targeted networks with a force-directed layout (Fruchterman-Reingold). **Community Detection:** Modules containing densely connected vertices were estimated within each network by using a fast greedy approach which aims at maximizing modularity of the detected communities<sup>30</sup>. Modularity measures how good a network partition is by calculating for each gene the number of edges within its community against its total node degree. The fast greedy approach optimizes modularity over all possible divisions of the network and has been shown to perform well on large networks. Guide genes clustering within the communities was compared to expected partitions in sub-pathway with a Pearson’s Chi<sup>2</sup> test and Monte-Carlo simulated *p*-values with 2,000 replicates. This test was based on a contingency table with dimensions  $n \times m$  ( $n$ , sub-pathway number,  $m$  community number in the co-expression network) and each entry corresponding to the number of genes being in communities  $n_i$  and  $m_j$ , with  $i = 1$  to  $n$  and  $j = 1$  to  $m$ . Because Chi<sup>2</sup> statistic depends on sample number, values were normalized by dividing them to the maximal expected value (the ideal partition) of each pathway. This resulted in a score ranging from 0 to 1, 0 being a random distribution of guide genes in the network and 1 to the exact partitioning.

**Data availability.** All datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

## References

- Oliver, S. Proteomics: guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
- Lisso, J., Steinhäuser, D., Altmann, T., Kopka, J. & Müssig, C. Identification of brassinosteroid-related genes by means of transcript co-response analyses. *Nucleic Acids Research* **33**, 2685–2696 (2005).
- Wei, H. *et al.* Transcriptional coordination of the metabolic network in arabidopsis. *Plant physiology* **142**, 762–774 (2006).
- Ruiz-Sola, M. A. *et al.* Arabidopsis geranylgeranyl diphosphate synthase 11 is a hub isozyme required for the production of most photosynthesis-related isoprenoids. *New Phytologist* **209**, 252–264 (2016).
- Guerin, C. *et al.* Gene coexpression network analysis of oil biosynthesis in an interspecific backcross of oil palm. *The Plant Journal* **87**, 423–441 (2016).
- Coman, D., Rütimann, P. & Gruissem, W. A flexible protocol for targeted gene co-expression network analysis. *Plant Isoprenoids: Methods and Protocols* 285–299 (2014).
- Caputti, L. *et al.* Missing enzymes in the biosynthesis of the anticancer drug vinblastine in madagascar periwinkle. *Science* <https://doi.org/10.1126/science.aat4100> (2018).
- Maetschke, S. R., Madhamshettiar, P. B., Davis, M. J. & Ragan, M. A. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in bioinformatics* **15**, 195–211 (2013).
- de Siqueira Santos, S., Takahashi, D. Y., Nakata, A. & Fujita, A. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in bioinformatics* **15**, 906–918 (2013).
- De La Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**, 3565–3574 (2004).
- Li, Y., Pearl, S. A. & Jackson, S. A. Gene networks in plant biology: approaches in reconstruction and analysis. *Trends in plant science* **20**, 664–675 (2015).
- Serin, E. A., Nijveen, H., Hilhorst, H. W. & Ligterink, W. Learning from co-expression networks: possibilities and challenges. *Frontiers in plant science* **7** (2016).
- Blasi, M. F. *et al.* A recursive network approach can identify constitutive regulatory circuits in gene expression data. *Physica A: Statistical Mechanics and its Applications* **348**, 349–370 (2005).
- Chai, L. E. *et al.* A review on the computational approaches for gene regulatory network construction. *Computers in biology and medicine* **48**, 55–65 (2014).
- Obayashi, T. & Kinoshita, K. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA research* **16**, 249–260 (2009).
- Wisecaver, J. H. *et al.* A global co-expression network approach for connecting genes to specialized metabolic pathways in plants. *The Plant Cell Online* tpc-00009 (2017).
- Mutwil, M. *et al.* Assembly of an interactive correlation network for the arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiology* **152**, 29–43 (2010).
- Tsuchiya, M., Giuliani, A., Hashimoto, M., Erenpreisa, J. & Yoshikawa, K. Self-organizing global gene expression regulated through criticality: mechanism of the cell-fate change. *PloS one* **11**, e0167912 (2016).
- Ballouz, S., Verleyen, W. & Gillis, J. Guidance for rna-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* **31**, 2123–2130 (2015).
- Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* **13**, 328 (2012).
- Censi, F., Giuliani, A., Bartolini, P. & Calcagnini, G. A multiscale graph theoretical approach to gene regulation networks: a case study in atrial fibrillation. *IEEE Transactions on Biomedical Engineering* **58**, 2943–2946 (2011).
- Huang, S. Reprogramming cell fates: reconciling rarity with robustness. *Bioessays* **31**, 546–560 (2009).
- Besseau, S. *et al.* Flavonoid accumulation in arabidopsis repressed in lignin synthesis affects auxin transport and plant growth. *The Plant Cell* **19**, 148–162 (2007).
- Zhang, Y. *et al.* Phenolic compositions and antioxidant capacities of chinese wild mandarin (*citrus reticulata blanco*) fruits. *Food chemistry* **145**, 674–680 (2014).
- Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant physiology* **126**, 485–493 (2001).
- Elejalde-Palmett, C. *et al.* Characterization of a spermidine hydroxycinnamoyltransferase in malus domestica highlights the evolutionary conservation of trihydroxycinnamoyl spermidines in pollen coat of core eudicotyledons. *Journal of experimental botany* **66**, 7271–7285 (2015).
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457–D462 (2016).
- Hwang, I., Sheen, J. & Müller, B. Cytokinin signaling networks. *Annual review of plant biology* **63**, 353–380 (2012).
- Zdarska, M. *et al.* Illuminating light, cytokinin, and ethylene signalling crosstalk in plant development. *Journal of experimental botany* **66**, 4913–4931 (2015).
- Clauzel, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Physical review E* **70**, 066111 (2004).
- Giorgi, F. M., Del Fabbro, C. & Licausi, F. Comparative study of rna-seq-and microarray-derived coexpression networks in arabidopsis thaliana. *Bioinformatics* **29**, 717–724 (2013).



32. Uygun, S., Peng, C., Lehti-Shiu, M. D., Last, R. L. & Shiu, S.-H. Utility and limitations of using gene expression data to identify functional associations. *PLoS computational biology* **12**, e1005244 (2016).
33. Jiang, L. *et al.* Strigolactones spatially influence lateral root development through the cytokinin signaling network. *Journal of experimental botany* **67**, 379–389 (2015).
34. Wang, L. & Chong, K. The essential role of cytokinin signaling in root apical meristem formation during somatic embryogenesis. *Frontiers in plant science* **6** (2015).
35. Kauffmann, A. *et al.* Importing arrayexpress datasets into r/bioconductor. *Bioinformatics* **25**, 2092–2094 (2009).
36. R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria <http://www.R-project.org> (2018).
37. Feltus, F. A., Ficklin, S. P., Gibson, S. M. & Smith, M. C. Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an arabidopsis case study. *BMC systems biology* **7**, 44 (2013).
38. Bolger, A. M. *et al.* Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–2120 Oxford University Press (2014).
39. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).
40. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 550 (2014).
41. Sales, G. & Romualdi, C. *parmigene*—parallel r package for mutual information estimation and gene network reconstruction. *Bioinformatics* **27**, 1876–1877 (2011).
42. López-Kleine, L., Leal, L. & López, C. Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. *Briefings in functional genomics* **12**, 457–467 (2013).
43. Schäfer, J. & Strimmer, K. Learning large-scale graphical gaussian models from genomic data. In *AIP Conference Proceedings*, vol. 776, 263–276 (AIP, 2005).
44. Schaefer, J., Opgen-Rhein, R. & Strimmer, K. *Corpcor*: efficient estimation of covariance and (partial) correlation. R package version 1.4. 7 (2007).
45. Du, Z., Zhou, X., Ling, Y., Zhang, Z. & Su, Z. *Agrigo*: a go analysis toolkit for the agricultural community. *Nucleic acids research* **38**, W64–W70 (2010).
46. Schrynemackers, M., Küffner, R. & Geurts, P. On protocols and measures for the validation of supervised methods for the inference of biological networks. *Frontiers in genetics* **4** (2013).
47. Ballouz, S., Weber, M., Pavlidis, P. & Gillis, J. *Egad*: ultra-fast functional analysis of gene networks. *Bioinformatics* **33**, 612–614 (2016).
48. Gillis, J. & Pavlidis, P. The impact of multifunctional genes on “guilt by association” analysis. *PloS one* **6**, e17258 (2011).
49. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).

## Acknowledgements

We deeply acknowledge the Fédération CaSciModOT (CCSC Orléans-Tours, France), Jean-Louis Rouet and Laurent Catherine for help and access to the Région Centre computing grid. We also thanks Yann Jullian for access and help on University computer resources. This study was supported by the Région Centre-Val de Loire, France (SiSCyLi grant). Doctoral Fellow attributed to F.L. and D.D. was jointly funded by the Région Centre-Val de Loire, France and the Ministère de l’Enseignement Supérieur et de la Recherche, France.

## Author Contributions

F.L., O.P., J.C., N.G. and T.D.D.B. conceived the experiment(s), F.L., D.D., O.P., M.C., S.B., V.C. and R.D.D.B. conducted the experiment(s), F.L., S.B., G.G., J.C., J.O.C. and T.D.D.B. analyzed the results. All authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-29077-3>.

**Competing Interests:** The authors declare no competing interests.

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018



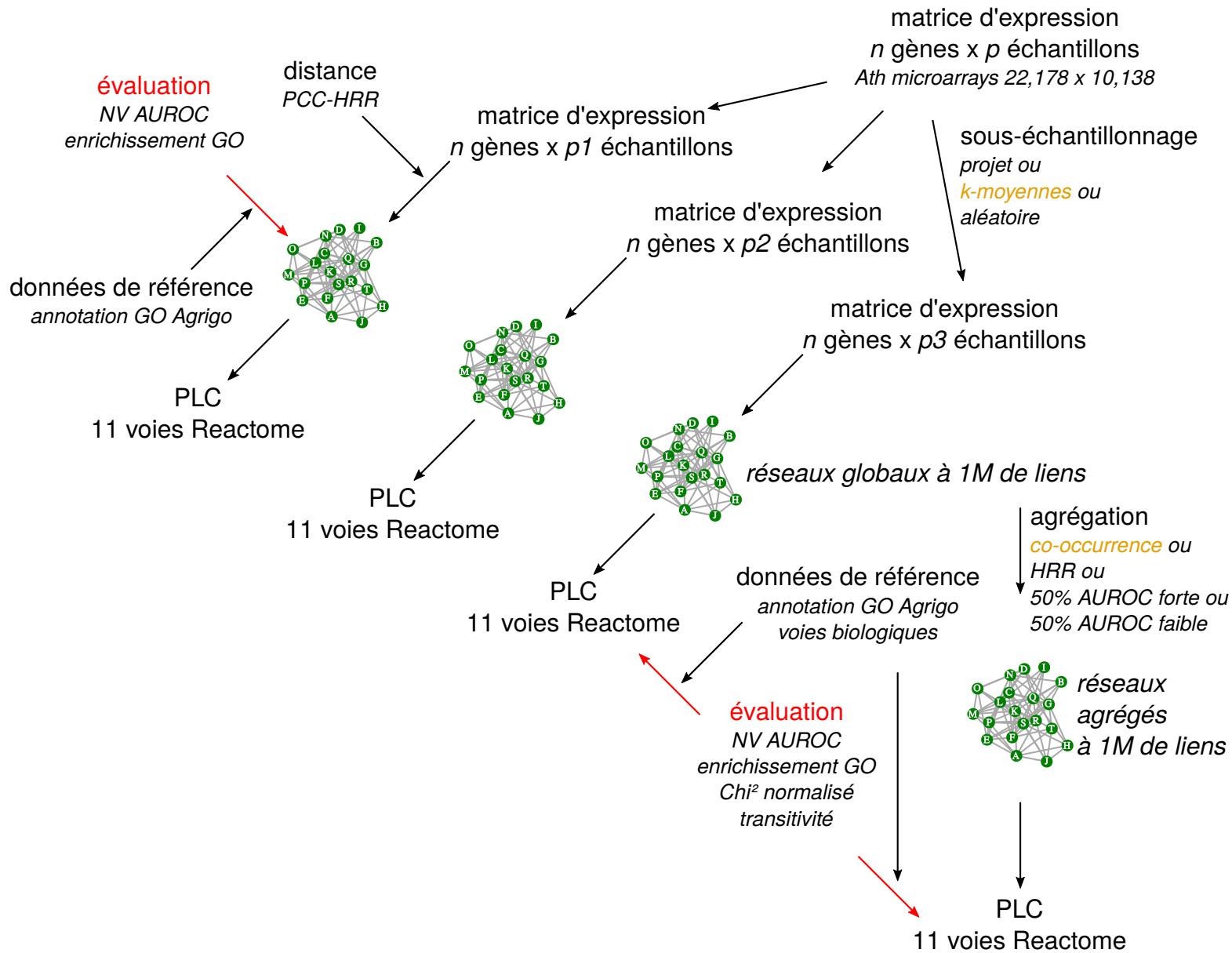
## Partie II

**Une affaire de taille : gestion d'échantillons de larges jeux de données d'expression dans la construction de réseaux de co-expression robustes**



## Partie II

**Une affaire de taille : gestion d'échantillons de larges jeux de données d'expression dans la construction de réseaux de co-expression robustes**



**Figure 24: Impact de la taille de matrice sur la performance de réseaux de coexpression.** L'évaluation a été faite sur 3 espèces, *Arabidopsis thaliana* (indiquée dans le schéma), *Solanum lycopersicum* et *Zea mays*. Pour chacune des espèces, données microarrays et RNA-seq ont été utilisées. Chacune des 6 matrices (3 espèces x 2 technologies) a été sous-échantillonnée par trois méthodes différentes: un regroupement des conditions associées à un même projet, un regroupement des conditions par clustering avec k-moyennes ou un regroupement aléatoire de taille variable (25, 50, 100, ...). Trois sous-matrices sont montrées pour l'exemple mais plus de 6000 matrices ont été travaillées au total (tous jeux de données confondus). Chacun des réseaux provenant des matrices sous-échantillonnées a été évalué individuellement ainsi que sur 11 PLC utilisant des gènes guides décrits dans la base de données Reactome. De plus, ces mêmes réseaux ont aussi été agrégés et pour chaque agrégat 1 million de liens ont été retenus selon l'un des 4 méthodes suivantes: nombre d'occurrences (les plus fréquents sont conservés), la valeur du HRR (les plus significatifs), la valeur du HRR en gardant 50% des réseaux avec les meilleures AUROCs, ou la valeur du HRR en gardant 50% des réseaux avec les plus faibles AUROCs. Les mêmes PLC ont alors été réalisées sur les différents agrégats puis évaluées. Les termes indiqués en orange montrent la combinaison la plus efficace pour l'analyse des voies biologiques.

La première partie de nos évaluations a indiqué l'intérêt d'utiliser les PCC-HRR pour la construction de PLC. Dans cette seconde partie, nous focalisons l'évaluation sur le jeu de données initial. Les réseaux obtenus microarray d'*A. thaliana* ont globalement montré une performance supérieure par rapport aux réseaux obtenus par RNA-seq. Cette précédente observation soulève deux possibilités : (i) la technologie microarray est plus adaptée pour l'inférence de réseaux de co-expression ou (ii) la présence de plus d'échantillons dans la matrice d'expression microarrays ( $>10\,000$ ) que dans la matrice RNA-seq (1 600) permet une estimation plus correcte des distances entre gènes. Pour vérifier ces deux possibilités, une stratégie complète d'évaluation a été mise au point (**Figure 24**). Tout d'abord, nous avons sélectionné 6 jeux de données différents représentant 3 espèces : *A. thaliana*, *S. lycopersicum* et *Z. mays*. Pour chaque espèce, des données microarray et RNA-seq ont été incluses. Concernant l'évaluation de l'impact du nombre d'échantillons sur l'inférence des réseaux, nous avons testé trois méthodes de sous-échantillonnage. Les échantillons ont été groupés (i) selon leur projet d'origine (Table III), (ii) selon leur similarité par une partitionnement en k-moyennes ou (iii) aléatoirement. Pour cette dernière méthode, plusieurs tailles ont été testées (25, 50, 100, 200,...) et pour chaque taille, beaucoup de combinaisons aléatoires incluses de manière à ce qu'un maximum d'échantillons soit au final représenté dans l'ensemble des sous-matrices d'une même taille. Plus de 6,000 réseaux ont ainsi été évalués avec des critères similaires à ceux utilisés en partie 1. La construction des PLC a été étendue à onze grands ensembles métaboliques ou de signalisation décrits dans la base de données Reactome. Tirant profit du grand nombre de réseaux générés, nous avons également observé que l'agrégation de réseaux issus de sous-matrices améliore la capture d'associations biologiquement réelles. Plusieurs méthodes d'agrégation ont été testées, la plus efficace pour capturer ces associations étant basée sur le nombre d'occurrences des liens dans les réseaux.

L'ensemble de ces évaluations montre que les réseaux construits à partir des matrices d'expression complètes sont plus performants que ceux obtenus à partir de matrices sous-échantillonées. Cependant, l'agrégation de réseaux inférés de matrices obtenues après partitionnement d'une matrice large par k-moyennes s'est avérée encore plus performante que les réseaux issus de matrices larges, surtout lorsque l'agrégation est faite par co-occurrence des liens.

Cette seconde partie se termine par une illustration de cette méthodologie sur la voie de biosynthèse de l'Acide Jasmonique, une phytohormone importante contrôlant des processus



développementaux et de défense. Cette illustration monte la force de la méthodologie pour identifier des gènes candidats qui pourraient avoir un rôle capital dans cette voie, notamment par le biais de l'analyse de co-occurrence de ces gènes entre les trois espèces.

Article soumis dans **Nucleic Acids Research**.



# A matter of size: how to deal with samples from a large gene expression dataset to construct a robust co-expression network.

Franziska Liesecke<sup>1</sup>, Johan-Owen de Craene<sup>1</sup>, Sébastien Besseau<sup>1</sup>, Vincent Courdavault<sup>1</sup>, Marc Clastre<sup>1</sup>, Valentin Vergès<sup>1</sup>, Nathalie Giglioli-Guivarc'h<sup>1</sup>, Gaelle Glévarec<sup>1</sup>, Olivier Pichon<sup>1</sup> and Thomas Dugé de Bernonville<sup>1,\*</sup>

<sup>1</sup>Université de Tours, EA2106 Biomolécules et Biotechnologies Végétales, 31 avenue Monge, Tours, F-37200, France

## ABSTRACT

Large-scale gene co-expression networks are an effective methodology to analyze sets of co-expressed genes and discover new gene functions or associations. Distances between genes are estimated according to their expression profile and are visualized in networks that may be further partitioned to reveal communities of co-expressed genes. Creating expression profiles is now eased by the large amounts of publicly available expression data (microarrays and RNA-seq). Although many distance calculation methods have been intensively compared and reviewed in the past, it is unclear how to proceed when many samples reflecting different conditions are available. Should as many samples as possible be integrated into network construction or be partitioned into smaller sets of more related samples? Previous studies have indicated a saturation in network performance to capture known associations once a certain number of sample is included in distance calculations. Here, we examined the influence of sample size on co-expression network construction from microarray and RNA-seq expression data using three plant species. We tested different down-sampling methods and compared network performance in recovering known gene associations to networks obtained from full datasets. We further examined how aggregating networks may help increase this performance by testing four aggregation methods.

## INTRODUCTION

Co-expression networks are proven to be efficient approaches to uncover biologically relevant gene pair associations. The starting expression matrix from which distance measurements are calculated to determine correlations between genes, is expected to drive the final shape and content of the network. In relevant co-expression networks, the ratio of known and experimentally proven associations (True Positives) over the total number of captured links is expected to be high. Evaluating the True Positive Rate (TPR) and False Positive Rate (FPR) of a given network with respect to a reference annotation set remains challenging but simple machine learning algorithms have been shown to efficiently calculate these rates and assess a network quality (1).

Combining individual datasets (from independent studies, SRP /ERP numbers in RNA-seq or GSE in microarrays) is expected to increase biological situation range and help capture transient associations. Real gene pair associations will be found in the network only if their common expression is detected in the starting dataset. Including more datasets in co-expression analyses should therefore add biological situations where such co-expression occurs. Contrastingly, increasing the sample number in an expression dataset may also result in increased noise together with decreased capacity to detect transient associations, following the garbage in garbage out principle. An open question remains about the number of expression datasets needed to build the most biologically relevant networks, *i.e.* capturing as many real associations as possible while keeping the number of false negatives low.

If including more samples to construct a network improves its quality, it is still unclear how many datasets are sufficient to capture relevant gene associations. For model species with many available expression datasets (*e.g.* more than 1,000 samples), global networks can be constructed from one expression matrix combining every available sample but does it capture more efficiently biological associations than networks obtained from smaller or down-sampled datasets? How adding or removing samples alters the network composition is not clear.

A pioneer study(2) used an *Escherichia coli* microarray data compendium. They analyzed dependencies among samples and found that compendium subsets perform better than the full one in transcriptional regulatory network inference. The low efficiency of the global network was attributed to sample redundancies but could be circumvented by calculating an optimal effective number of samples. In their work, the full compendium (376 samples) could be down-sampled to 50% without decreasing network quality.

Using a larger *E. coli* expression compendium (524 samples) as well as synthetic datasets (with up to 2,000 samples), Altay et al(3) tested several information theory based inference methods as well as the sample size effect. In this case, both simulated and real datasets showed that *ca.* 100 samples were sufficient to capture transcriptional genome-wide regulations. These previous reports(2, 3) took advantage of the well known *E.coli* regulatory network to evaluate their

\*To whom correspondence should be addressed. Tel: +33 247 367023; Email: thomas.duge@univ-tours.fr

© The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.



co-expression networks. In another study(4), co-expression networks were obtained after applying a Random Matrix Theory process to threshold similarity matrices calculated with Pearson Correlation Coefficients (PCC). The effects of both gene number and sample size were analyzed for 3 species: Human, Rice and Yeast. The authors have shown a high edge conservation between full and down-sampled networks. However, new edges appeared with smaller datasets (down to 25% of the initial size) while other edges were lost. This indicates that conserved associations between genes are easy to uncover while revealing more transient association typically depends on the nature of samples in the dataset. Conserved associations corresponded to functional associations, suggesting that genes added or removed in the down-sampled networks mostly were found in already densely connected modules and were weakly connected to others (*i.e.* they were not hub genes).

Supervised down-sampling of a large dataset by finding the most appropriate dataset has been proposed to improve pathway reconstruction. Using a set of query genes, Hibbs et al(5) calculated correlations among this set on SVD-transformed individual datasets of *S. cerevisiae*. Each dataset was weighted according to its relevance, *i.e.* datasets maximizing PCC are given more weight. These weights are used next to calculate PCC of every gene with each query gene. This procedure is known as the SPELL algorithm (Serial Pattern of Expression Levels Locator). It has also been reported that creating subsets of related samples using a k-means approach improves feature detection(6). However, it remains to be determined whether individual networks resulting from partitions of down-sampled datasets could be more efficient in capturing associations than a network calculated from all initial samples. Aggregated networks have also been shown to improve the recovery of biologically relevant associations(7, 8). The underlying idea is that conserved coexpression links between 2 genes over several datasets reinforce the existence of a true association between these 2 genes (7). A web-based tool named MEM was designed to merge co-expression lists obtained from individual datasets(11). This multi-species microarray-based tool allows users to find the best coexpressed genes with an input query gene and manually excludes less relevant datasets considering the input. In MEM, gene ranks calculated with a query gene are aggregated over the selected datasets by using a binomial distribution hypothesis to attribute p-values to ranks and by taking the minimum value of all p-value. Ballouz et al(1) have constructed individual networks for different experiments and subsequently aggregated them, either taking all datasets or only the most significant (as indicated by AUROC of GO terms). They have revealed a clear improvement over individual experiments, probably because it has the advantage of combining moderately significant or condition-specific relationships.

Our aims are: (i) to establish the impact of the sample size on the recovery of relevant associations at both global and targeted levels and (ii) get more insights on the way individual and smaller networks may be aggregated to generate stronger networks. The latter implies a trade-off between capturing the maximum number of known associations and limiting the total association number.

## MATERIALS AND METHODS

### Dataset preparation

Microarray data were obtained from signal intensities in .CEL files downloaded from ArrayExpress in R(12, 41). Raw signals were quantile normalized with the RMA procedure in R(14). Outlier arrays were detected by monitoring quartile distribution together with Kolmogorov-Smirnov testing against an empirical cumulative distribution curve(6). RNA-seq data were obtained by downloading raw .fastq files from the EBI ENA. Reads were quasi-mapped with Salmon(15) on reference transcript assemblies obtained from Ensembl Plant to quantify transcript abundance. Assemblies were TAIR10 for *Arabidopsis thaliana*, SL2.50.31 for *Solanum lycopersicum* and AGPV3.31 for *Zea mays*. All accessions are indicated in Supplementary Table 1. Dataset sizes are indicated in Table 1.

### Dataset partitioning

Initial expression matrices were down-sampled in three different ways (Figure 1). In the first, k-means partitioning was performed in R using with 200 random starts. The optimal value of k was graphically determined for each dataset using the elbow method and varying k between 10 and 100 by increments of 10. In the second, datasets were partitioned by grouping arrays or runs from the same study/project (GSE for microarrays, SRP, ERP or DRP number for RNA-seq). In the last, we down-sampled datasets by randomly selecting a given number of samples. For each sample size, the random sampling was performed many times until the resulting down-sampled datasets represented at least 90% of the samples contained in the full expression matrix.

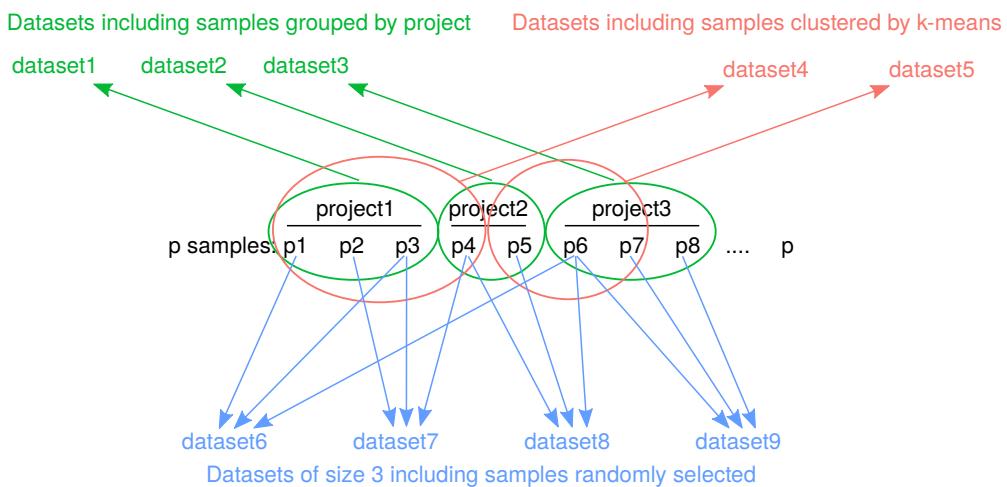
### Highest reciprocal ranking of Pearson Correlation Coefficients

Distances between transcript expression profiles were determined by calculating PCCs for each transcript pair. PCCs were next ranked so that for each gene, the rank value ranges from 0 (the gene itself) to N (the total number of genes), and the final rank value for a given gene pair was the highest of the two, *e.g.* for a gene pair A and B, HRR(A,B)=max(rank(cor(A,B)), rank(cor(B,A))). Computations were done in parallel with an MPI program written in C(16). The resulting distance matrices were arbitrary thresholded at a HRR<600 to obtain a large list of best co-expressed genes and these lists were further thresholded at different cut-off values to compare network performance at different sizes. As threshold choice has been shown to strongly influence network topology(18), the performance of networks obtained from different datasets

**Table 1.** Dataset sizes. Number of genes x number of samples.

	Microarrays	RNA-seq
<i>Arabidopsis thaliana</i>	22,178 x 10,138	33,602 x 1,676
<i>Solanum lycopersicum</i>	6,284 x 627	32,419 x 1,046
<i>Zea mays</i>	10,309 x 680	61,581 x 2,516





**Figure 1.** Down-sampling strategies. Starting from a given number of  $p$  samples retrieved from publicly available accessions, three methods were used to partition samples. Samples were grouped by their project accession, clustered by k-means or randomly selected (with possible redundancy between datasets).

(differing in size and the partitioning method) was evaluated for different network sizes.

## Network evaluation

Networks were evaluated for their ability to recover known or expected relationships between genes. Gene relationships obtained in a network were compared to gene associations described in Gene Ontology (GO) terms or the Reactome database(19). GO annotation files were downloaded from the Agrico database v2.0(20). Network ability to capture gene pairs associated with identical GO terms was evaluated by applying a neighbor voting algorithm which measures how well connections in the co-expression network predict a gene annotation. A three fold cross validation was performed and used to calculate an Area Under the Receiver Operating Characteristic (AUROC) for each GO term with the EGAD R package(21). Network performance was estimated with the GO AUROC which corresponded to the average of all GO term AUROCs. GO AUROCs of 0.5 and 1 respectively indicate random and perfect predictability. We also tested significant enrichment of networks with GO terms using an hypergeometric test. To test gene association into biological pathways, global networks with a 1 million edges were queried with guide gene (GG) sets described in the Reactome database. Genes co-expressed with the guides were captured to construct a sub-network, a process known as Pathway Level Coexpression(23). A total of 11 biological pathways were analyzed for each species. Subnetworks were evaluated in their ability to capture biologically relevant GO terms (GO AUROCs and significantly enriched GO terms) as well as to partition guide genes into their exact groups/pathway by measuring network modularity and a normalized Chi-squared metric(16). Networks were constructed and analyzed with the R igraph package(17), including several topological metrics such as transitivity (the probability that adjacent nodes of a given node are connected) and the log likelihood of node degree to fit a power law distribution.

## RESULTS

### Individual down-sampled matrices

Biologists aiming at constructing co-expression networks for their model species may face a large number of samples and experiments in public databases. A largely unresolved question is the way to process such data. Should the data be combined or processed as data subsets? We evaluated both methods by generating down-sampled expression datasets from large compendia (Figure 1) in three different ways (Figure 1). Samples were grouped according to their project accession number, according to their similarity using a k-means clustering or by random sampling at different fixed sample sizes. Because we allowed replacement in the sampling procedure, some samples were found in several matrices (5 matrices on average and rarely more than 10) (Supplementary Figure 1). A large variety of data subsets were therefore investigated (Figure 2A; 3,673 for *A. thaliana*, 947 for *Z. mays* and 913 for *S. lycopersicum*). The randomized sampling procedure generated many small data subsets (a minimum of 25 samples) and fewer large ones (Figure 2B). Data subsets obtained from project-grouped or k-means clustered samples rarely contained more than 100 samples (Figure 2B). Each data subset was used to construct a distance matrix using HRR ranked PCC thresholded at  $HRR > 600$  to get a large list of the best co-expressed gene pairs. These lists were next cut off at different confidence levels to construct networks and evaluate their performance in capturing GO terms. First, many more edges were retained for network construction with less stringent confidence thresholds (Figure 2C; Spearmans  $\rho > 0.95$ ,  $p\text{-value} < 2e-16$ ). A clear positive relationship was observed between edge number and GO AUROC (Spearmans  $\rho > 0.7$ ,  $p\text{-value} < 2e-16$ ) revealing that more true than false positive edges were included in networks at the considered network sizes. At a given confidence threshold (e.g. 0.1), sample size tended to be negatively correlated with edge number (average Spearmans  $\rho = -0.33$ ,  $p\text{-value} < 2e-16$ ) (Figure 2D). Positive significant correlations between sample size and GO AUROC were



observed for the random down-sampling and all 3 species (Spearmans  $\rho > 0.5$ ,  $p\text{-value} < 5e-14$ ), but for *A. thaliana* and *Z. mays* networks only with k-means down-sampled matrices (Spearmans  $\rho > 0.4$ ,  $p\text{-value} < 0.05$ ) (Figure 2E). In the other cases, there was no significant correlation between sample size and GO AUROC. These results suggested that networks from larger datasets potentially captured more biologically relevant GO terms. Although including more edges in networks (at less stringent confidence thresholds) clearly increased GO AUROCs, networks constructed with larger datasets required less edges to reach a similar GO AUROC. Taking the randomly sampled *A. thaliana* microarray dataset as an example, a high GO AUROC of 0.65 was obtained on average with 781,482 edges for matrices with 25 samples and 367,830 for matrices with 100 samples (Supplementary Fig2). This trend was less contrasted for *S. lycopersicum* (for a GO AUROC of 0.60, 547,494 edges with 25 samples vs 269,550 with 200 samples) and *Z. mays* (for a GO AUROC of 0.60, 192,261 edges with 25 samples vs 84,795 with 400 samples) datasets. In every case, strong significant effects of sample size and edge number on GO AUROC were observed (Supplementary Figure 2), indicating that smaller datasets might generate GO AUROCs as high as larger datasets by increasing the edge number. This likely indicated that the best associations found in smaller datasets were either false positives or new and transient associations which did not correspond to known GO associations.

We next compared GO AUROCs of networks constructed from datasets with more than 20 and less than 75 samples to evaluate the 3 down-sampling methods. Within this size range, we ensured that networks deriving from each method were comparable in terms of initial dataset size (Figure 2B). It revealed significantly higher GO AUROCs for randomly sampled networks but no difference between project-grouped or k-means clustered samples (Figure 3A). This indicated that using PCC-HRR with randomized matrices may be more informative than using thematically related samples. One would have expected that reducing complexity in expression matrices by combining related or similar samples might improve correlations between genes. To verify this hypothesis, we calculated Spearmans  $\rho$  correlations between samples for each matrix. We found a weak negative correlation between sample correlations and GO AUROC of the resulting networks (Figure 3B). Samples clustered per project or by k-means were in average significantly more correlated (0.92 and 0.85 respectively, calculated all data combined) than those selected randomly (0.76; Wilcoxon rank sum test,  $p\text{-value} < 2.2e-16$ ). Taken together, these results indicated that calculating a simple correlation between samples of a given dataset can be useful to partially predict performance of the resulting network. Data presented in Figure 3B also suggest that datasets with very weakly correlated samples (e.g., with a Spearmans  $\rho < 0.6$ ) should be associated with lower GO AUROC but this remains to be demonstrated.

At a 1 million edges, microarray based networks globally performed better than those based on RNA-seq (Figure 4). This trend was less clear for counts of significantly enriched GO terms (Supplementary Figure 3). For *S. lycopersicum* and *Z. mays*, networks derived from RNA-seq had significantly more enriched GO terms than those derived from microarrays while GO AUROCs were higher for microarray datasets.

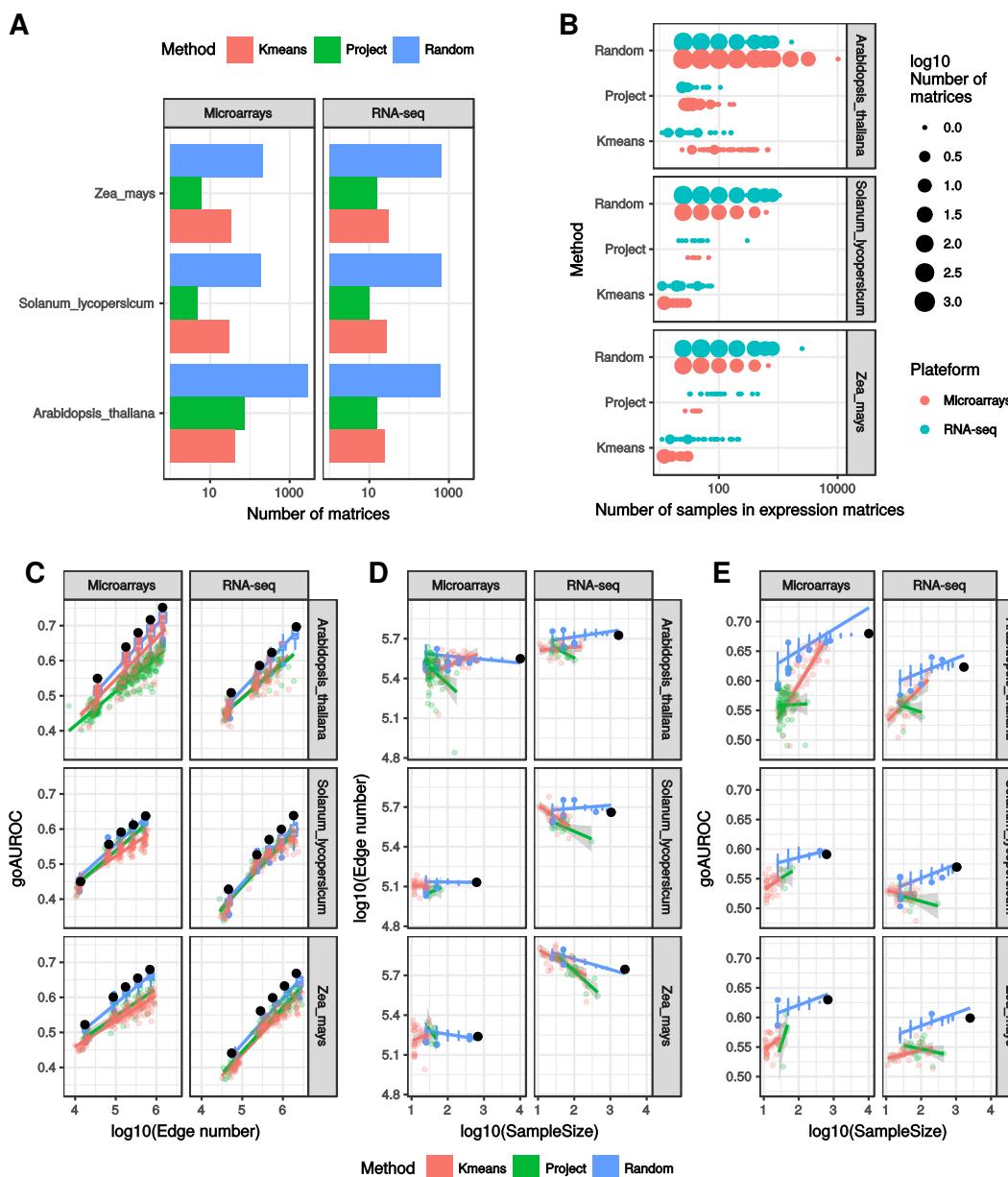
This was probably due to the incomplete transcriptome representation in microarrays for these two species (Table 1). While GO AUROCs were high for genes analyzed by microarrays, these genes inherently represent less different processes than those measured by RNA-seq. It also suggests that genes pairs with a same GO term are more likely to be direct neighbors in the microarray than in the RNA-seq networks. A weaker predictability for RNA-seq derived networks was surprising. Because networks were obtained at a same number of edges, it indicated that many edges were not represented into the GO reference annotation. These edges could be considered either as false positive or true interactions not captured in the current GO reference annotation. It is possible that the more exhaustive view with RNA-seq encompasses genes which associations could be false positives or yet unaccounted true associations resulting in a decreasing in RNA-seq network predictability.

### Aggregating networks

Networks obtained from smaller datasets generally had lower GO AUROCs than those with larger ones, but they could theoretically capture transient or local associations hidden in larger datasets. To allow networks obtained from small datasets to highlight both transient and more conserved gene associations, we analyzed the performance of their aggregation. Each aggregate generates a new network also named aggregated network. We compared aggregated networks from smaller datasets with those obtained from full datasets. Aggregation was expected to increase the global GO AUROC when using small datasets to construct networks. To aggregate individual networks, edge lists are combined and redundant edges are collapsed giving them a new score (Figure 5). This score is based on either the number of occurrences (therefore depending on the number of networks in the aggregate) or on the lowest HRR value. The first method is based on edge co-occurrence (CO) and considers most represented gene pairs as more robust than those found in only one network. Using the lowest HRR as a weight considers gene pairs with low HRR as significant, even if it is found in only one network. In this case, local associations are expected to be captured more efficiently. During the aggregation process, we asked whether all networks or only some of them should be considered. Inspired by a previous study(1), we tested complete and partial aggregations containing either the 50% highest GO AUROCs or the 50% lowest GO AUROCs. For these 2 types of aggregation, best edges were retained according to their lowest HRR value. A total of 4 aggregation methods were thus investigated: co-occurrence (CO), HRR-based (H), HRR-based 50% highest GO AUROCs (H-HGA, HRR-based aggregation Highest GO AUROC) and HRR-based 50% lowest GO AUROCs (H-LGA, HRR-based aggregation Lowest GO AUROC). Once redundant edges are collapsed, the 1M best gene pairs (either by their CO or by their HRR value) are retained for further network characterization.

**Aggregating networks from project or K-means partitioned data subsets** We first investigated edge conservation among aggregation methods. Correlation between HRR values and number of co-occurrences of each gene pair was low and





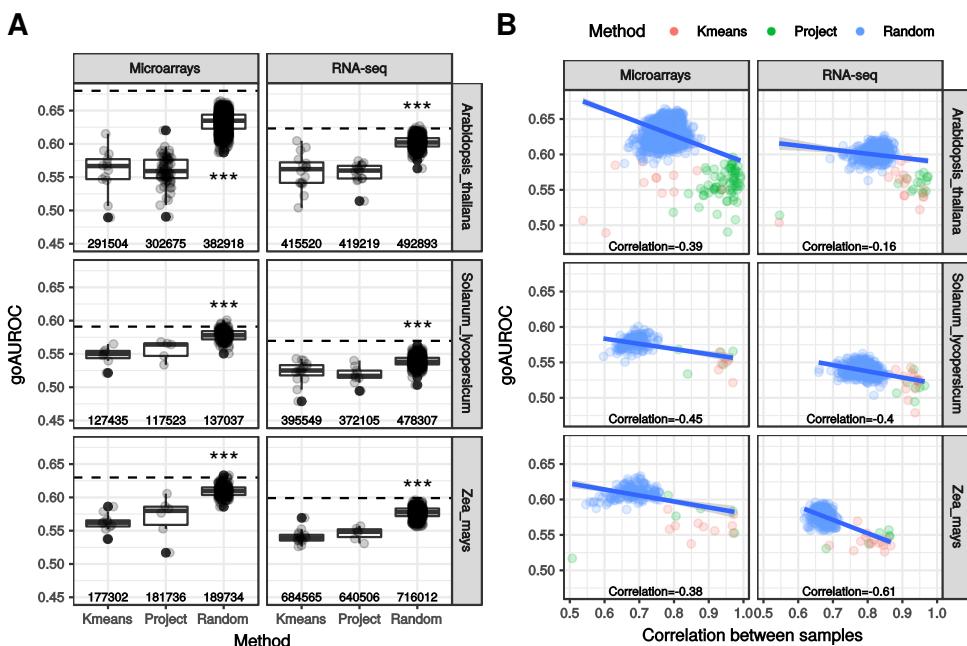
**Figure 2.** Performance of down-sampled expression matrices in capturing GO associations. Microarray and RNA-seq expression matrices from three species, *Arabidopsis thaliana*, *Solanum lycopersicum* and *Zea mays*, were prepared by combining all available RNA-seq data and subsequently down-sampled in three different ways, (i) random sampling, (ii) by project and (iii) by sample clustered by a k-means analysis, to construct networks at different confidence thresholds (1, 5, 10, 20 and 40% of the best co-expressed transcripts). The total number of down-sampled expression matrices is shown in A and the number of samples per table in B. In B, the highest number of samples in the Random category correspond to the full matrix. Pairwise relationships between GO AUROC, edge number and sample size are shown in C for all thresholds and in D and E at a threshold of 0.1. For networks obtained from randomly sampled expression matrices, data are summarized as boxplots (in blue). Lines correspond to regression lines with their 95% confidence interval in grey areas. Black dots correspond to data for networks inferred from full datasets.

generally negative (Figure 6A, maximum  $\rho=0.15$ , average  $\rho<-0.089$ ), indicating that most frequent gene pairs had uncorrelated HRR values. This explained that only partial overlaps were found between network aggregates obtained with the two methods (H and CO) (Figure 6B). For microarray based-networks, aggregates obtained by co-occurring edges shared between 30 and 50% of edges with those obtained by the minimal HRR value. For RNA-seq based networks, common edges represented generally between 10 and 20%.

Contrastingly, there was a higher number of common edges between HRR-based and H-HGA or L-LGA aggregated network while less than 10% were conserved between H-HGA and L-LGA based aggregates (Figure 6B). This indicated that HRR-based aggregates contained best edges from both best performing and worst performing networks.

We next measured aggregate performance in capturing GO associations. Networks derived from full matrices and CO aggregates had statistically higher GO AUROCs than single





**Figure 3.** Comparison between down-sampling methods. (A) GO AUROCs of networks obtained at a threshold of 0.1 from datasets with more than 20 and less than 75 samples are summarized as boxplots for each down-sampling method. Asterisks show significant difference in median between randomly sampled datasets and the two other methods (Wilcoxon rank sum test,  $p<0.001$ ). Average edge numbers are indicated below boxplots. (B) For each down-sampled matrix, correlation between samples was calculated (Spearman's  $\rho$ ) and plotted against the GO AUROC of the resulting network. Correlation between the two variables was calculated with the Pearson coefficient. Blue lines correspond to regression lines with their 95% confidence interval in grey areas. Dashed lines correspond to data for networks inferred from full datasets.

networks (Wilcoxon rank sum test,  $p$ -value<0.01) (Figure 6C). Although statistical differences in GO AUROCs between co-occurrence aggregates and individual networks were not confirmed by counts of significantly enriched GO terms (Supplementary Figure 4), it was likely that CO aggregates displayed the highest performance and were at least as efficient as networks derived from full matrices for both performance measures. All data combined, only networks aggregated by co-occurrence and 50% lowest GO AUROC (H-LGA) resulted in significantly different GO AUROCs (Figure 6D). We observed a substantial but not significant improvement in GO AUROCs between aggregates containing networks with the 50% best GO AUROCs or the 50% lowest GO AUROCs suggesting that a prior selection of networks to combine could substantially improve GO term recovery. No significant difference in performance was observed between aggregates of project or k-means clustered datasets, although in almost all cases the GO AUROC of the CO aggregates of k-means networks was higher than that of project aggregates.

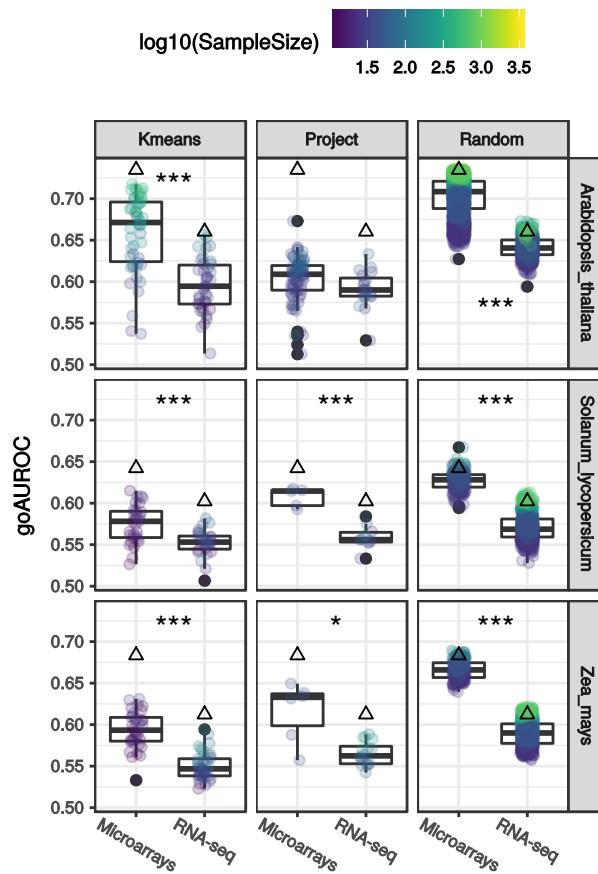
To further characterize the different aggregates, we subsequently analyzed 13 biological pathways from the Reactome database in Pathway Level Coexpression (PLCs(23)) subnetworks. Pathway reconstruction was evaluated according to guide gene partitioning into specific communities and biological processes associated within these communities (Supplementary Table 2). PLC size was set at a 1,000 vertices (1,000 genes best co-expressed with the query guide genes) obtained from the 1 million best edges. Over the 13 biological pathways, PLC from the full dataset networks or aggregated networks according to edge co-occurrence

appeared to better capture GO terms as reflected by high GO AUROC and counts of significantly enriched GO terms (Wilcoxon rank sum test,  $p$ -value<2e-16)(Supplementary Figure 5). No significant difference was observed among the three remaining aggregation methods (HRR-based, H-HGA and H-LGA). It is noteworthy that all aggregated networks from a same species x platform combination had the same number of edges (set at a 1 million) and almost the same number of vertices. Although PLC size was set at a 1,000 vertices, PLC networks had considerably variable edge numbers. In particular, PLCs based on co-occurrence aggregates had many more edges resulting in a higher mean node degree. As indicated by their higher GO AUROCs, these edges are likely to correspond to biologically relevant gene associations.

By contrast, networks aggregated according to edge HRR values had a lower mean node degree and the 1,000 vertices needed to construct the PLC were reached with a few edges, revealing that each node is connected to a few other ones. Concerning the pathway reconstruction quality, guide gene distribution into communities better matched the expected partition for the three best HRR based methods than for the co-occurrence one as revealed by their higher normalized Chi-squared values (Wilcoxon rank sum test,  $p$ -value<0.01)(Supplementary Figure 5). All data combined, normalized Chi-squared values were indeed the lowest in the co-occurrence aggregate ( $p$ -value<1e-05). This highlighted a trade-off between GO capture and pathway reconstruction.

It was likely that capturing gene associations annotated with GO terms was optimal with the most represented edges,





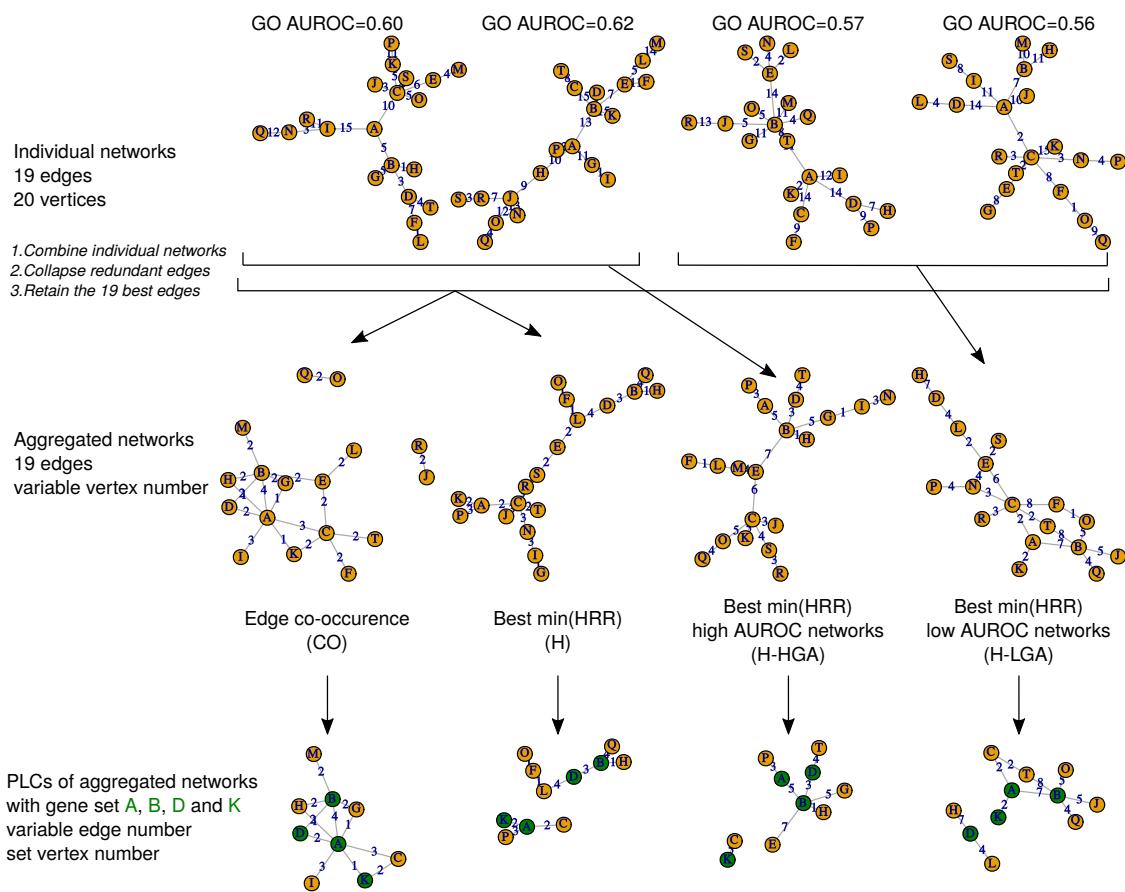
**Figure 4.** Performance comparison between microarray and RNA-seq. The performance of networks with a 1 million edge to capture GO terms was measured with GO AUROC. Asterisks denote a significant difference between the two platforms (Students *t* test, \*, *p*-value<0.05, \*\*, *p*-value<0.01, \*\*\*, *p*-value<0.001). Each point represent one individual network and boxplots summarize data all sample sizes confounded. White triangles correspond to data for networks inferred from full datasets.

while correctly associating genes from a same pathway requires transient and/or unknown associations with other genes. PLC network topology was evaluated by calculating the clustering coefficient which measures a probability that a given node is connected to other nodes in the network. This coefficient indicates the module structure of the network(24). It clearly appeared that the HRR-based aggregation procedure resulted in significantly lower clustering coefficients than co-occurrence aggregated and full dataset derived PLC networks (Wilcoxon rank sum test, *p*-value<1e-05)(Supplementary Figure 5). In accordance with this, degree distribution of these two kinds of PLC networks also had a better fit to the power law as revealed by the calculated log likelihood (Supplementary Figure 5). Taken together, these results suggest that for the full dataset or edge co-occurrence aggregates, partitions of guide genes into communities did not strongly reflect the expected partitions in pathways but their higher clustering coefficients revealed a more modular structure (see Supplementary Figure 6 for an example with the secondary metabolite Reactome pathway). This might be in turn explained by interdependencies between sub-pathways (guide genes that could be theoretically simultaneously found in several communities) as well as incomplete pathways in the

database. Concerning HRR value based aggregation methods (H, H-HGA and H-LGA), project network aggregates had globally better metrics than those using datasets with k-means clustered samples (Supplementary Figure 5). However, these aggregates never outperformed CO aggregates or full dataset networks. On the whole, PLC networks from full dataset had very good features in terms of GO AUROC, modularity, normalized Chi-squared and clustering coefficient.

During the aggregation process, we were also interested in aggregating microarray and RNA-seq based networks. Following the co-occurrence principle, we considered that gene pairs highlighted in networks deriving from different technologies may be more robust. We therefore determined co-occurring genes between PLCs obtained with the two technologies and analyzed intersection networks for each aggregation method. We found that PLCs from co-occurrence aggregates displayed the highest degree of conservation between microarrays and RNA-seq as revealed by the number of co-occurring edges and vertices (Figure 7A). In fact, the other aggregation methods did not allow to correctly align microarray and RNA-seq PLC networks. Microarray and RNA-seq intersection network of HRR aggregates had very few edges and vertices as exemplified with the secondary





**Figure 5.** Network aggregation procedures. In this example, single networks had 20 vertices (named with A to T) and 19 edges (numbers indicate a hypothetical HRR value sampled with replacement between 1 and 15) and were generated according to the model of Barabasi and Alberts(38). Aggregation was either total or partial. Total aggregation resulted in 20 vertices and 57 non redundant edges. The 19 best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For these two partial aggregations, the 19 best pairs were retrieved by taking the best HRR (H-HGA or H-LGA).

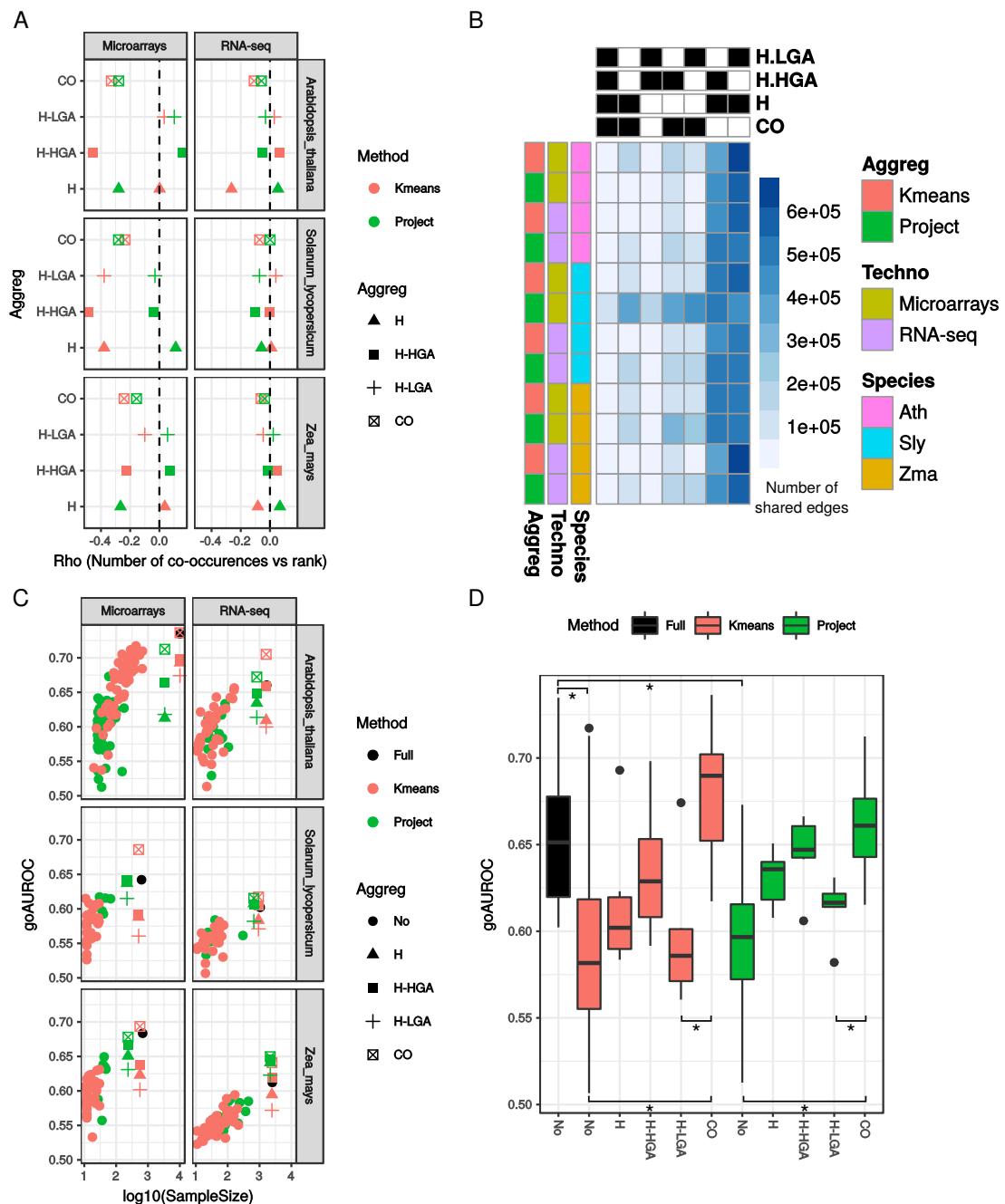
metabolite pathway in *S. lycopersicum*(Figure 7B). This reinforces the observation that gene pairs with lowest HRR did not frequently co-occur. Combining microarray and RNA-seq aggregates resulted in a substantial drop of GO AUROC in contrast to technology specific aggregates (an average of 0.54 vs 0.57 respectively,  $p$ -value<0.001). This was explained by the loss of vertices and edges during the combination of aggregates. Concerning data partitioning methods, no significant difference was observed between k-means and project on GO AUROC.

Taken together, these results suggest that, according the reference annotation sets used here, aggregating networks using the edge co-occurrence method is useful to improve the capture of relevant gene pairs. Final networks containing edges found in both microarray and RNA-seq based networks are likely to provide an appropriate transcriptional map of a given pathway at the expense of a slightly lower GO AUROC.

**Aggregating networks from randomly sampled data subsets.** Data subsets obtained by random combinations of samples resulted in networks performing better than those originating from subsets corresponding to project or k-means partitions

of the initial dataset (Figure 3A). Given the large sample size in each initial matrix (Table 1), aggregation possibilities were numerous. We evaluated different aggregate sizes with multiple sample combinations for each and we aggregated networks according to the size of the datasets they originated from. Aggregate performance was evaluated with GO AUROC and the number of significantly enriched GO terms. Aggregates differed by both the number of networks they contained (x-axis on Figure 8), the combination of networks included (error bar on y-axis on Figure 8) and by the sample size of the initial dataset used to generate individual networks (different colors on Figure 8). Plotting mean GO AUROCs of aggregates obtained from different sample combinations revealed an overall low variation whatever the aggregate size or the initial dataset sample size (Figure 8). Largest variations in GO AUROC were observed for aggregates of networks obtained with datasets having 25 and 50 samples, especially in *S. lycopersicum* microarray networks. This suggests that edge content differed substantially between different combinations of a same aggregate size. This was relatively unexpected for *S. lycopersicum* microarray based aggregates because the corresponding arrays contain less than 7,000 genes which



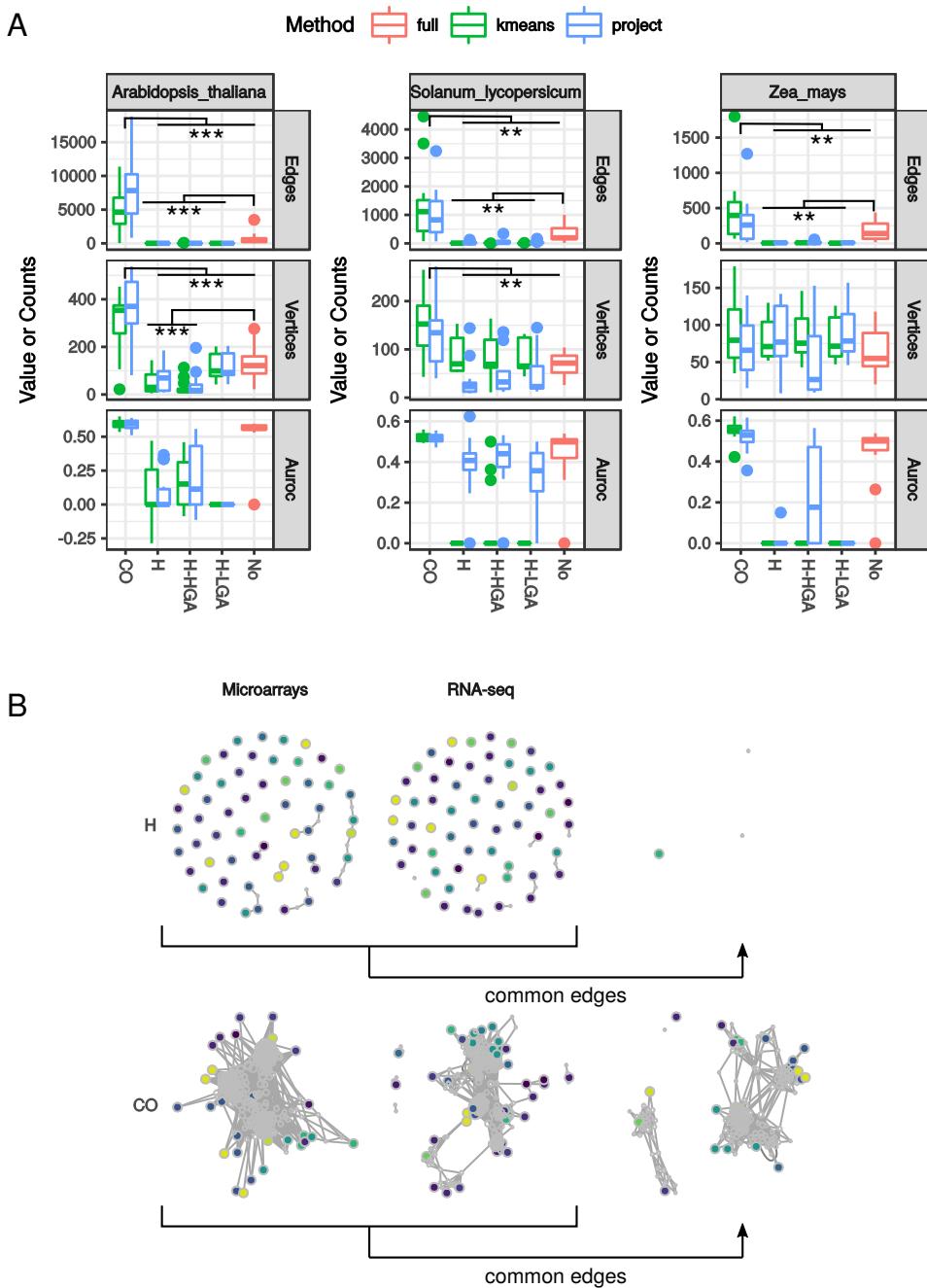


**Figure 6.** Aggregated networks (1 M edges) from expression matrices down-sampled by grouping samples by their project or by clustering them by k-means. Aggregation was either total or partial. For total aggregation, the 1 M best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For the two partial aggregations, the 1 million best pairs were retrieved by taking the best HRR (H-HGA or H-LGA). (A) For each 4 aggregation methods and each dataset, correlation (Spearmans rho) between gene pair HRR value and number of co-occurrence was calculated. (B) Number of edges found among aggregates obtained by different methods. Black and white boxes for columns respectively indicate that an aggregate is included or not in the comparison. (C) Network performance measured by GO AUROC. Single non aggregated (No) networks with 1 million edges are also reported. (D) Data from all species and platforms summarized in boxplots. Asterisks denotes significant differences between two procedures (Wilcoxon rank sum test,  $p$ -value < 0.05).

might have led to more homogeneous edges. This higher heterogeneity shows that correlations are largely impacted by the starting dataset when using partially represented transcriptomes.

Networks aggregated according to edge HRR values rarely outperformed the network from the full dataset. Edge CO aggregation generally improved GO AUROC for RNA-seq based networks but the effect was less visible for microarray data. Larger dataset sizes generally decreased differences



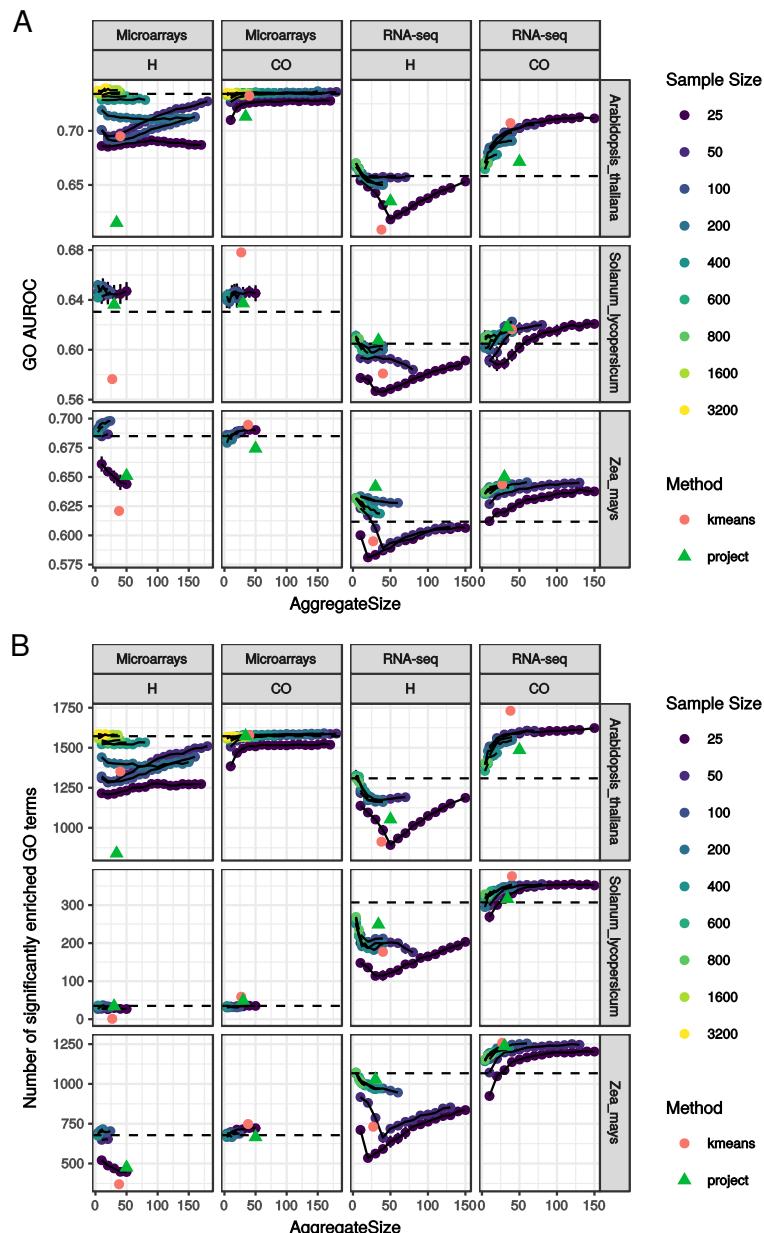


**Figure 7.** Characteristics of merged Pathway Level Coexpression (PLC) obtained with microarray and RNA-seq data. For total aggregation, the 1 M best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For these two partial aggregations, the 1 million best pairs were retrieved by taking the best HRR (H-HGA or H-LGA). No aggregation indicates that networks from the full datasets were compared. Network size and performance measured with GO AUROC are presented in (A). Asterisks indicate significant differences (Wilcoxon rank sum test, \*\*,  $p$ -value < 0.01; \*\*\*,  $p$ -value < 0.001). In (B) are shown two examples with *Solanum lycopersicum* and the secondary metabolite pathway. Microarray and RNA-seq networks were obtained by aggregating networks from k-means partitioned data subsets with either the CO (edge co-occurrence) or the H (best HRR) method. The third row displays the edges and vertices common to microarray and RNA-seq deriving PLCs.

between aggregation methods, especially when few networks (<10) are combined, suggesting that networks constructed from large dataset sizes (e.g. >50) are more robust. GO AUROC was increased with an average of 0.02 and a

maximum of 0.05 for *A. thaliana* RNA-seq when using the CO method (Supplementary Figure 7). Aggregation of networks from randomly selected samples increased the number of significantly enriched GO terms on average in 21 and 182 for





**Figure 8.** Aggregate performance (A, GO AUROC; B, counts of significantly enriched GO terms) with networks obtained from randomly sampled data subsets. Aggregate size was set at a 1 million edges. Only two aggregation methods are shown. Redundant edges in the aggregate were collapsed according to their co-occurrence (CO) in several individual networks or according to their best HRR value (H). Horizontal dashed lines indicate the number of GO terms for networks obtained with the full datasets. All points correspond to average measures over replicates and vertical black bars to standard deviation. Pink points and green triangles indicate GO AUROCs of networks aggregates using k-means or by project respectively.

microarrays and RNA-seq respectively, with a minimum of 2 for *S. lycopersicum* microarrays and a maximum of 314 for *A. thaliana* RNA-seq in comparison to the full dataset derived networks (Figure 8, Supplementary Figure 7).

As a general trend, including more networks with the CO aggregation method helped to increase GO AUROCs (Spearman's  $\rho > 0.5$ ,  $p\text{-value} < 0.05$ ) whatever the initial sample size until a plateau was reached (Figure 8). Using larger sample sizes had contrasting effects on GO AUROCs according to the initial dataset considered, increasing GO AUROCs for *Z. mays* and *S. lycopersicum* RNA-seq

aggregates ( $\rho > 0.75$ ,  $p\text{-value} < 1e-05$ ) but decreasing it for *A. thaliana* RNA-seq aggregates ( $\rho = -0.80$ ,  $p\text{-value} < 1e-09$ ). When looking at aggregate size x dataset size combination maximizing GO AUROC, the CO method performed better with more aggregates of networks from small datasets (Supplementary Figure 7). Contrastingly, the HRR method required less aggregates but larger dataset sizes.

The aggregates of networks inferred from k-mean partitioned datasets performed generally at least as good as the best random aggregates (Figure 8). This trend was even more pronounced when considering the number of significantly



enriched GO terms. We examined whether it was also the case for PLC (Supplementary Figure 8). PLC performance varied considerably according to the pathway considered as shown by min and max pathway metrics in Supplementary Figure 8. This was probably due to the important differences in Reactome pathway sizes and compositions (Supplementary Table 2). The CO aggregation method was the best to improve GO term recovery in comparison to full dataset networks (+0.014 vs -0.03 for the HRR method for GO AUROC; +23 vs -54 for the number of significantly enriched GO terms, Welch two sample *t*-test *p*-value<2e-16 for both parameters) but modularity and normalized Chi-squared statistic were higher when the aggregation was based on HRR (modularity: 0.58 for CO vs 0.66 for HRR; normalized Chi-squared statistic: 0.17 vs 0.25, Welch two sample *t*-test *p*-value<2e-16 for both parameters) (Supplementary Figure 8). PLC GO AUROC was only significantly influenced by aggregate size, not by the initial sample size (ANOVA, *p*-value<0.00573, expected for *Z. mays* microarray based networks), while counts of significantly enriched GO terms were influenced by aggregate size for RNA-seq based networks only (ANOVA, *p*-value<0.0206). GO capture in PLC with aggregates of networks from randomized datasets could therefore be maximized by using small dataset and large aggregate sizes. Differences between randomly sampled datasets or partitioned datasets in GO term capture varied according to the species and the platform. On the whole, aggregates of networks constructed from k-mean partitions had a higher number of significantly enriched GO terms for RNA-seq as well as a higher average GO AUROC (excepted for *Zea mays*) (pairwise *t*-tests *p*-value<0.05). This was true for small aggregate sizes, because in all cases, larger aggregate sizes (obtained with smaller dataset sizes, e.g. 25 or 50) generated PLCs with GO capture performance similar to k-means partitioned dataset network aggregates.

Concerning topological metrics, modularity was not significantly influenced by either aggregate size or sample size in *S. lycopersicum* and *Z. mays* microarray based aggregates (ANOVA, *p*-value>0.05) (Supplementary Figure 8). Strongest effects were observed for RNA-seq based aggregates, with larger sample sizes correlating with higher modularity (Spearman's *rho*>0.61, *p*-value<2e-16). A very similar observation was made for the clustering coefficient, although CO based aggregates had higher values than HRR based aggregates (average of 0.58 for CO, 0.22 for HRR; *t*-test, *p*-value<2e-16). For all species x platform combinations, PLCs had a significantly higher log likelihood of fitting to a power law when aggregated with the CO method (*t*-test, *p*-value<0.03). These topological metrics suggest that CO based aggregates have a more clustered structure with hubs corresponding to few genes having many connections with other, as exemplified with the Fatty Acid pathway in Supplementary Figure 9. Partitioning guide genes into communities was similar to expected partitions when using the HRR aggregation method (average of 0.25 for HRR and 0.16 for CO; *t*-test *p*-value<2e-16). This was mainly due to the higher number of guide genes represented in HRR based aggregates (average of 51 for HRR aggregates vs 38 for CO aggregates; *t*-test *p*-value<2e-16). However, the lower clustering coefficients of HRR-based aggregates showed a less evident interpretation of the communities (Supplementary

Figure 9). With the exception of *A. thaliana* microarray based networks, the clustering coefficient was higher in aggregates of randomly selected samples than in aggregates of partitioned datasets (Supplementary Figure 8; *t*-test *p*-value<0.014). Although aggregates of networks constructed with k-means partitioned datasets had a very good performance in GO term capture, the PLC topology may be less biologically relevant, such as exemplified in *Z. mays* RNA-seq based k-means aggregates vs random aggregate (Supplementary Figure 9).

### Application to jasmonic acid (JA) biosynthesis in plants

To further demonstrate why aggregating networks is useful to study biological pathways, we extracted PLC from RNA-seq networks containing a 1 million edges with genes involved in the biosynthesis of JA (9). For this example, we focused on three networks for each species: (i) a network inferred from the full RNA-seq dataset, (ii) a co-occurrence (CO) aggregate of individual networks inferred from the k-means partitioned full dataset and (iii) a HRR-value based aggregate of individual networks inferred from the k-means partitioned full dataset. Each 1 million edge network (a total of 9, 3 construction modes in 3 species) was queried with guide genes (GG) obtained from the Plant Metabolic Network databases(10) (26 for *A. thaliana*, 44 for *S. lycopersicum* and 33 for *Z. mays*) and included lipoxygenases (LOX), allene oxide synthases (AOS) and cyclases (AOC) among others (Supplementary Table 3). Among the 1 million edge networks of *A. thaliana*, *S. lycopersicum* and *Z. mays*, 2,050, 575 and 1,935 edges on average contained one guide gene respectively. It was surprising that although more guide genes were considered for *S. lycopersicum*, the edge number was lower than in the two other species, the maximum edge number (863) being observed for the full dataset derived network. The n best edges involving at least one guide gene (having the lowest HRR values or the highest CO) in each network were retained so that the final PLCs contain ca. 30 vertices (Figure 9A; Supplementary Table 4). This cut-off value was chosen for three reasons: (i) to allow comparisons among networks and among species, (ii) it represented no more than 10% of the vertices found in PLCs performed on the 1 million edge networks and (iii) in CO or H aggregates, it retained edges with a CO weight >2 (*i.e.*, each edges was found in at least two individual network before aggregation) or a HRR value <10 respectively. This HRR value is a stringent threshold(39). Hence, the 9 resulting PLCs (3 construction modes for 3 species) with 30 vertices are expected to be high confidence closely focused on JA metabolism and signaling.

Genes with evident relationships such as TIFY transcription factors, known modulators of the JA signaling pathway(22), were classified as associated genes (AG) while those with no previously described relationship to this pathway were classified as other genes (OG) (Figure 9B). Within each species, the number of GG retained in the 30-vertices PLCs was similar but they differed among networks. In fact, gene content strongly differed between PLCs because less than 20% of the genes were conserved in a same species between each pairwise PLC comparison, showing the strong impact of the construction procedure on the resulting PLCs. Although differences between numbers of AG among aggregation methods were not statistically significant, CO



aggregates of networks inferred from k-means partitioned datasets appeared to contain more AG (11.3 in average over the 3 species) than networks inferred from full datasets (3.6 in average) or from H aggregates (3.3 in average). Mapping gene accessions to the Uniprot databases clearly revealed that CO aggregated networks contained more genes related to response to wounding for *A. thaliana* and *S. lycopersicum* (Supplementary Table 4).

In *S. lycopersicum* and *Z. mays*, CO aggregates contained JA biosynthesis related genes not included as guide genes and not found in the other networks, suggesting that these aggregates were likely to give a more exhaustive picture of transcriptional relationship within the JA biosynthesis and signaling than the other construction methods. It was the case for Solyc07g007870.2 encoding a 12-oxophytodienoic reductase (OPR3; found in the community containing Solyc04g079730.1 corresponding to an AOS) and GRMZM2G136857 encoding a putative JA methyltransferase (in the community containing GRMZM2G104843 corresponding to a LOX). For the two species, patatin-like proteins (Solyc04g079250.2 and GRMZM2G154523) potentially involved in releasing linolenic acid from membrane phospholipids as a JA biosynthetic precursor were detected in the CO aggregates(9). This highlights CO aggregate ability to capture close relationships from biological pathways and to prioritize candidate genes to be functionally tested. The present CO aggregates also indicate that unexpected functions might be associated with the JA biosynthesis or signaling. For example, different hydrolases were found in the transcriptional neighborhood of the guide genes (AT1G31550 encoding a GDSL esterase/lipase, Solyc03g083010.2 encoding a alpha/beta fold hydrolase and GRMZM2G032160 encoding a glycoside hydrolase). Together with the presence of amino-acid metabolism related genes (such as AT4G08870 encoding an arginase and Solyc03g013160.2 encoding an amino-acid transporter), these candidate genes may reveal unexpected but key connections of the JA pathway with primary metabolism.

In *A. thaliana*, the CO aggregate contained 5 TIFY, 2 WRKY as well as the basic helix-loop-helix (bHLH) AtMYC2 transcription factors, families containing known JA signaling regulating factors(26, 30). Similarly, the CO aggregate of *S. lycopersicum* contained 4 TIFY, 2 bHLH and 1 WRKY. For *Z. mays*, the community containing a lipoxygenase encoded by GRMZM2G104843 did not display significant GO term enrichment but included several transcription factors with bHLH (2), TIFY (2) or WRKY (1) domains suggesting they could be the functional orthologs of the two other species. For example, transcripts for WRKY40 (AT1G80840 and Solyc03g116890.2) were also found in *Z. mays* (GRMZM2G120320), suggesting this would be the *Z. mays* functional WRKY40 ortholog. This was confirmed by BlastP analysis against nr NCBI database in which Solyc03g116890.2 and GRMZM2G120320 had best similarities with AT1G80840 (respective max score 261 and 188), although AT1G80840 had better homologies to Solyc06g068460.3 than to Solyc03g116890.2 (score 261 vs 241 with and Solyc03g116890.2) and to GRMZM2G111711 than to GRMZM2G120320 (score 185 vs 181 with GRMZM2G120320). This revealed that direct orthology would have probably failed at finding the correct

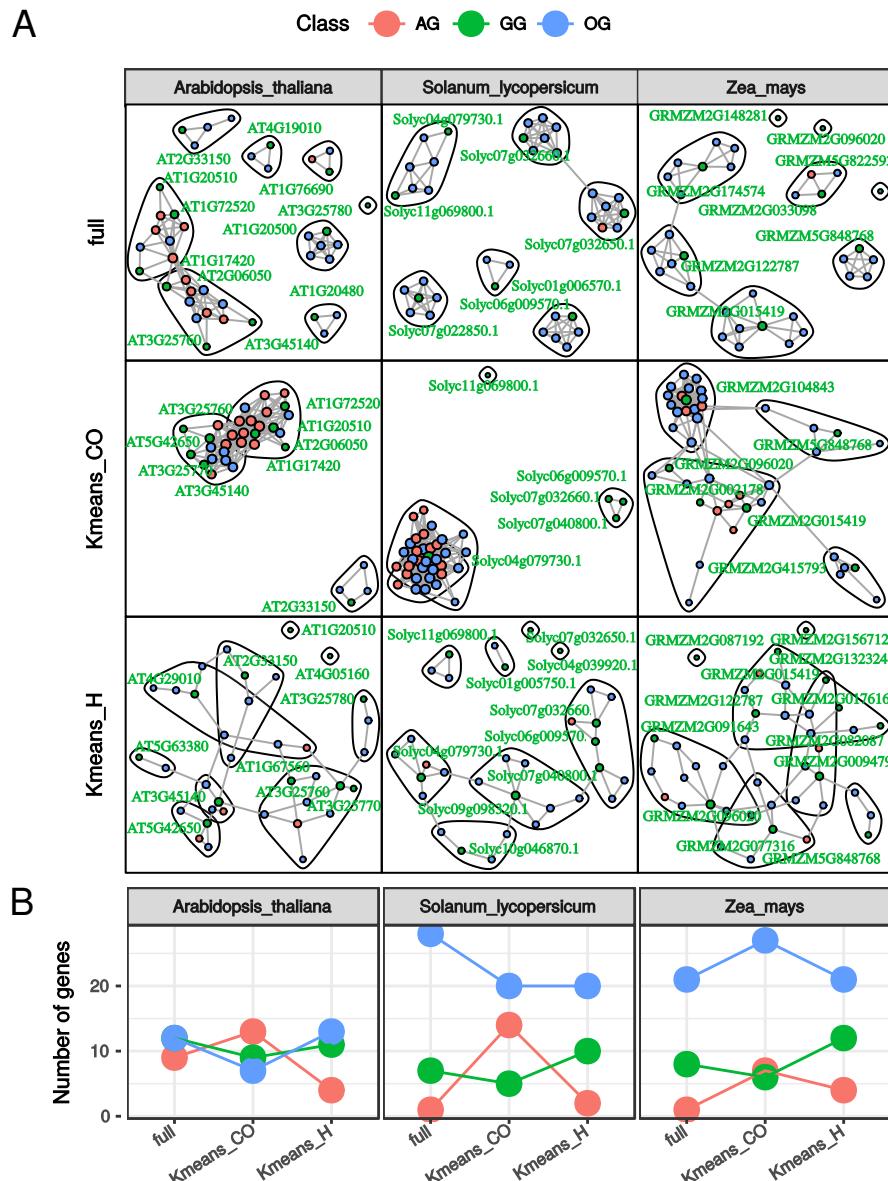
functional WRKY factors whereas our network approach succeeded. These transcripts in *S. lycopersicum* and *Z. mays* are currently annotated as putative WRKY factor or uncharacterized protein in Uniprot, but their presence in the JA PLCs obtained here strongly suggests their involvement in JA signaling. Also of interest were transcripts encoding putative bHLH transcriptions factors found in CO aggregates of *S. lycopersicum* (Solyc09g083360.2) and *Z. mays* (GRMZM2G301089). Both of them had best similarities with *A. thaliana* bHLH92 (AT5G43650). The only bHLH transcription factor identified in CO aggregate of *A. thaliana* was MYC2 (AT1G32640). Whether orthologs of bHLH92 may play a similar role than MYC2 in the two other species deserve more investigations. On the whole, our networks clearly indicate an involvement of a few specific members from large families of transcription factors (on average: 231 bHLH, 25 TIFY and 111 WRKY for the 3 species as indicated in the PlantTFDB(13)), that would have not been as efficiently identified by other approaches.

## DISCUSSION

Using gene co-expression is an efficient approach to predict a gene function or find candidate genes involved in a given pathway(25, 27, 28, 29, 31, 32). In many cases, the authors generate their own data and infer co-expression from them. It is also possible to re-use published data to visualize transcriptional relationships in large expression compendia, such as those available on ATTED II(33) and PlaNet(40) databases. In any case, it is not clear how many samples should be included and/or how related they should be to calculate distances between genes. Specific datasets may highlight specific correlations while sharing many transcriptional relationships(34). Whether a limited number of datasets is more appropriate than maximal datasets (containing all available samples) to capture gene transcriptional relationships remains to be determined.

Our results clearly showed that individual networks from down-sampled datasets (Figure 1) do not outperform networks obtained from the largest datasets (Figure 2), in accordance with previous studies(1, 2, 3, 4). For both microarray and RNA-seq, the more samples contained in the initial dataset, the higher the performance in capturing edges matching GO terms is expected to be. Single networks from randomly down-sampled data subsets had a higher GO AUROC than single networks obtained from project or k-means partitioned data subsets probably because samples in these latter were more correlated between themselves (Figure 3). In addition, it appeared that microarray data resulted in higher GO AUROC at equivalent sample sizes (Figure 4). Aggregation has already been used in co-expression analysis(7, 8), however without extensive testing of aggregation methods. We found a substantial performance gain when aggregating individual networks. Retaining edges co-occurring in several networks to construct the aggregate clearly improved recovering known gene pairs (Figure 6 and 8). This method allowed a good overlap between microarray and RNA-seq derived aggregates using project or k-means partitioned dataset networks (Figure 7). Such co-occurring edges in networks deriving from two technologies and differing initial sample sets could be considered as more robust. Taken altogether, these results





**Figure 9.** Pathway Level Coexpression (PLC) for the JA biosynthesis pathway in *Arabidopsis thaliana*, *Solanum lycopersicum* and *Zea mays*. Gene accessions were obtained from the Plant Metabolic Pathway Databases. PLCs were constructed by retrieving guide genes with their best co-expressed genes from large networks constructed from full RNA-seq datasets or from RNA-seq datasets partitioned with K-means and subsequently aggregated according to edge co-occurrence (CO) or edge HRR value (H). PLC size was set at 30 genes (A). Communities were detected with a fast greedy algorithm and are delimited by black polygons. Genes in PLC were classified as guide genes (GG, green, with their respective locus tags), associated genes (AG, non-guide genes with evident relationships to guide genes, red) or other genes (OG, non-guide genes without evident relationship to guide genes, blue) and counted in each PLC (B).

suggest that inferring networks from large compendia with PCC and HRR are biologically relevant for both microarray and RNA-seq data.

Aggregating individual networks constructed from randomly sampled datasets resulted in meta networks with satisfying performance, especially when using RNA-seq data where some aggregates performed much better than the network derived from the full dataset (Figure 8). However, aggregates of k-means partitioned datasets generally had a better GO AUROC than other aggregates. This indicates that

the strategy proposed by Feltus et al(6) is very appropriate for both microarray and RNA-seq data. These networks appeared to highlight a wide range of transcriptional relationships (Figure 8). This was further confirmed by both GO AUROC and PLC topological metrics. In addition, results obtained for the k-means partitioned datasets appeared to be more robust than for randomly sampled datasets for which trends differed between species (Figure 8). Further experiments will be necessary to determine the minimal number of samples required for a k-means based partitioning.



In the present article, we did not test other distance calculation methods. Whether our results can be directly transposed to other more complex calculations (such as in supervised methods for regulatory network inference) remains to be determined. It is possible that the ranking procedure used here is tolerant to error and adapted to larger datasets. Mutual information based networks of *E. coli* expression data processed with different algorithms displayed similar trends during down-sampling(3). Down-sampling had also a similar impact on network build with Spearmans rank correlations(1). Although more investigations will be required when using other distance measurements, it is likely that similar results will be observed with other methods.

In our previous work(16), we showed that PLC quality could be monitored with the GO AUROC, normalized Chi-squared and modularity. This has previously been demonstrated using KEGG pathways on Arabidopsis. In the present article, we extended our set of validation pathway by considering the Reactome database. We found that it included more complete information especially for hormone signaling pathways and was more convenient to compare pathways between the three plant species investigated here. However, we found very low normalized Chi-squared values in contrast to the KEGG pathways. We impute this difference to the complexity of the Reactome pathways which sometimes included many different subpathways. To further characterize PLC quality, we showed that the clustering coefficient also called transitivity(24) was useful to measure how networks were marked by few hubs (high clustering coefficient value) or by transcripts homogeneously connected to each others (low clustering coefficient value).

The higher AUROCs obtained with microarrays may reveal that single color arrays are well adapted to co-expression networks or that our annotation sets were more appropriated for genes effectively represented on each array. As gene models evolve with genome annotation refinement, RNA-seq, which is based on a mapping of reads on a reference transcriptome, allows to quantify more comprehensively gene expression(34). For the three species investigated here, reference transcriptomes represented more genes than arrays. It is possible that the reference annotation sets used here reflected more array gene content than the more comprehensive RNA-seq based transcriptomes, suggesting that some associations in the RNA-seq network could be true positives but are considered as false positives because not found in the reference annotation sets.

## CONCLUSION

Taken altogether, our results suggest that co-expression networks using PCCs ranked with HRR clearly benefit from increasing sample size of the initial expression dataset. Small sized datasets (with less than 100 samples) had variable performance which was probably due to the samples they contained. We observed that differences between networks decreased when constructed from datasets with more than 100 samples (Figure 2 and 4). As a consequence, any combination of more than 100 samples may generate robust networks. When more than 500 samples are available (as it was the case in our 6 combinations), more biologically relevant networks can even be obtained by creating single networks after

partitioning the whole dataset with a k-means algorithm and aggregating them according to co-occurring edges.

## SUPPORTING MATERIAL

Supplementary Figure1: Occurrence of samples in randomly down-sampled expression matrices. For each sample, we counted the number of matrices containing it.

Supplementary Figure 2: Significant interactions between edge number and sample size on GO term recovery. Statistical effects were analyzed by ANOVA, asterisks denoting a significant effect (\*,  $p$ -value<0.05; \*\*\*,  $p$ -value<0.001). The 4 groups of points represent networks obtained at 4 different significance thresholds.

Supplementary Figure 3: Performance comparison between microarray and RNA-seq. The performance in capturing GO terms of networks with a 1 million edges was measured by the number of significantly enriched GO terms (hypergeometric test,  $q$ -value<0.05). Asterisks denote a significant difference between the two plateform (Students  $t$  test, \*,  $p$ -value<0.05, \*\*,  $p$ -value<0.01, \*\*\*,  $p$ -value<0.001). Each point represent one individual network and boxplots summarize data all sample sizes combined. White triangles correspond to data for networks inferred from full datasets.

Supplementary Figure 4: Performance of aggregated networks from expression matrices down-sampled by grouping samples by their project or by clustering them by k-means. Performance is evaluated by the number of significantly enriched GO terms in each network (hypergeometric test,  $q$ -value<0.05). Single non aggregated (No) networks with 1 million edges are also reported. Aggregation was either total or partial. For the total aggregation, the 1 M best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For the partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For these two partial aggregations, the 1 million best pairs were retrieved by taking the best HRR (H-HGA or H-LGA).

Supplementary Figure 5: Application of PLC on aggregated networks from expression matrices down-sampled by grouping samples by their project or by clustering them by k-means. Boxplots summarize values obtained for 13 biological pathways from the Reactome Database. Performance in GO recovery is evaluated by GO AUROC and the number of significantly enriched GO terms (hypergeometric test,  $q$ -value<0.05). Aggregation was either total or partial. For the total aggregation, the 1 M best pairs were retrieved either by taking the best HRR (H) or the most co-occurring (CO) edges. For the partial aggregation, we combined either the 50% of networks with the highest GO AUROCs (HGA) or 50% with the lowest GO AUROCs (LGA). For these two partial aggregations, the 1 million best pairs were retrieved by taking the best HRR (H-HGA or H-LGA).

Supplementary Figure 6: PLC Networks obtained full datasets and aggregated networks for the secondary metabolites Reactome pathway.

Supplementary Figure 7: Maximum performance improvement of aggregated networks (A, GO AUROC; B, counts of significantly enriched GO terms) over full dataset. Aggregate size was set at a 1 million edges. Only



two aggregation methods are shown. Y values correspond to the mean GO AUROC or counts of significantly enriched GO terms from the best combination of sample and aggregate size (from Figure 7) to which the value of the full dataset network was subtracted. Redundant edges in the aggregate were collapsed according to their co-occurrence (CO) in several individual networks or according to their best HRR value (H).

Supplementary Figure 8: Application of PLC on aggregated networks from expression matrices randomly down-sampled. Individual networks (1M edges) from a same sample size were aggregated in various number (x axis, aggregate size) and PLCs were performed using 13 different gene sets from the Reactome database. Characteristics were averaged over all replicates from a same sample size x aggregate size combination. Each point corresponds to averaged measure of 13 pathways in the Reactome database and vertical bars range from min and max values. Performance in GO recovery is evaluated by GO AUROC and the number of significantly enriched GO terms (hypergeometric test,  $q\text{-value} < 0.05$ ) and both measures are expressed as the average of each pathway difference between aggregates and the corresponding full dataset network. Other measures, modularity, clustering coefficient and the log likelihood of a Power law fit are used to analyze PLC topologies. Normalized Chi-squared value measure the quality of guide gene partitioning into expected subgroups as depicted in the Reactome database. Pink and blue points respectively correspond to aggregates of networks obtained with k-means or project partitioned datasets.

Supplementary Figure 9: PLC with the Fatty acid metabolism gene set on aggregated networks. For clarity purposes, edges are not shown. Communities are delimited by black polygons and colored vertices represent guide genes from a same sub-pathway. Redundant edges in the aggregate were collapsed according to their co-occurrence (CO) in several individual networks or according to their best HRR value (H). All PLCs contain the best edges allowing the representation of no more than 300 vertices.

Supplementary Table 1: Sample and their corresponding study accession number.

Supplementary Table 2: Gene accessions for Reactome pathways.

Supplementary Table 3: Guide genes related to Jasmonic Acid biosynthesis. Gene accessions were retrieved from the Plant Metabolic Network database.

Supplementary Table 4: Gene content and functional enrichment of Pathway Level Coexpressions with Jasmonic Acid (JA) biosynthesis related genes. GG, guide genes; AG, associated genes (evident association with JA biosynthesis or signaling); OG, other genes (non-evident association with JA biosynthesis or signaling).

## ACKNOWLEDGEMENTS

We deeply acknowledge the Fdration CaSciModOT (CCSC Orlans-Tours, France), Jean-Louis Rouet and Laurent Catherine for help and access to the Rgion Centre computing grid. We also thanks Yann Jullian for access and help on University computer resources.

**Funding.** Doctoral Fellow attributed to F.L. was funded by the Rgion Centre-Val de Loire, France and the Ministre de l'Enseignement Suprieur et de la Recherche, France.

**Conflict of interest statement.** None declared.



## REFERENCES

1. Ballouz, S. and Verleyen, W. and Gillis, J. (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*; **31**, 2123–2130.
2. Cosgrove, E.J. and Gardner, T.S. and Kolaczyk, E.D. (2010) On the choice and number of microarrays for transcriptional regulatory network inference. *BMC bioinformatics*; **11**, 454.
3. Altay, G. (2012) Empirically determining the sample size for large-scale gene network inference algorithms. *IET systems biology*; **6**, 35–43.
4. Gibson, S.M. and Ficklin, S.P. and Isaacson, S. and Luo, F. and Feltus, F.A. and Smith, M.C. (2013) Massive-scale gene co-expression network construction and robustness testing using random matrix theory. *PloS one*; **8**, e55871.
5. Hibbs, M.A. and Hess, D.C. and Myers, C.L. and Huttenhower, C. and Li, K. and Troyanskaya, O.G. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*; **23**, 2692–2699.
6. Feltus, F.A. and Ficklin, S.P. and Gibson, S.M. and Smith, M.C. (2013) Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an Arabidopsis case study. *BMC systems biology*; **7**, 44.
7. Lee, H.K. and Hsu, A.K. and Sajdak, J. and Qin, J. and Pavlidis, P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome research*; **14**, 1085–1094.
8. Gillis, J. and Pavlidis, P. (2011) The impact of multifunctional genes on “guilt by association” analysis. *PloS one*; **6**, e17258.
9. Wasternack, C. and Feussner, I. (2018). The oxylipin pathways: biochemistry and function. *Annual Review of Plant Biology*; **69**, 363–386.
10. Schlapfer, P. and Zhang, P. and Wang, C. and Kim, T. and Banf, M. and Chae, L. and Dreher, K. and Chavali, A. K. and Nilo-Poyanco, R. and Bernard, T. and Kahn, D. and Rhee, S.Y. (2017). Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. *Plant Physiology*; **173**:2041–2059.
11. Adler, P. and Kolde, R. and Kull, M.s and Tkachenko, A. and Peterson, H. and Reimand, J. and Vilo, J.K. (2009) Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome biology*; **10**, R139.
12. Kauffmann, A. and Gentleman, R. and Huber, W. (2008) arrayQualityMetricsa bioconductor package for quality assessment of microarray data. *Bioinformatics*; **25**, 415–416.
13. Jin, J.P. and Tian, F. and Yang, D.C. and Meng, Y.Q. and Kong, L. and Luo, J.C. and Gao, G. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*; **45(D1)**:D1040–D1045.
14. Gautier, L. and Cope, L. and Bolstad, B.M. and Irizarry, R.A. (2004) affyanalysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*; **20**, 307–315.
15. Patro, R. and Duggal, G. and Love, M.I. and Irizarry, R.A. and Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*; **14**, 417.
16. Liesecke, F. and Daudu, D. and Dug de Bernonville, R. and Besseau, S. and Clastre, M. and Courdavault, Vt and De Craene, J.-O. and Crche, J. and Giglioli-Guivarc'h, N. and Glvarec, G. and Pichon, O. and Dug de Bernonville, T. (2018) Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Scientific Reports*; **8**, 10885.
17. Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, **1695**, 1–9.
18. Couto, C.M.V. and Comin, C.H.e and da Fontoura Costa, L. (2017) Effects of threshold on the topology of gene co-expression networks. *Molecular BioSystems*; **13**, 2024–2035.
19. Naithani, S. and Preece, J. and DEustachio, P. and Gupta P. and Amarasinghe, V. and Dharmawardhana, P.D. and Wu, G. and Fabregat, A. and Elser, J.L. and Weiser, J. and Keays, M. and Fuentes, A.M. and Petryszak, R. and Stein, L.D. and Ware, D. and Jaiswal, P. (2017) *Nucleic Acids Research*; **45(D1)**:D1029-D1039.
20. Tian, T. and Liu, Y. and Yan, H.u and You, Q. and Yi, X. and Du, Z. and Xu, W. and Su, Z. (2017) agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic acids research*; **45**, W122–W129.
21. Ballouz, S. and Weber, M. and Pavlidis, P. and Gillis, J. (2016) EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics*; **33**, 612–614.
22. Bai, Y. and Meng, Y. and Huang, D. and Qi, Y. and Chen, M. (2011) Origin and evolutionary analysis of the plant-specific TIFY transcription factor family. *Genomics*; **98**, 128–136.
23. Wei, H. and Persson, S. and Mehta, T. and Srinivasasainagendra, V. and Chen, L. and Page, G.P. and Somerville, C. and Loraine, A. (2006) Transcriptional coordination of the metabolic network in Arabidopsis. *Plant physiology*; **142**, 762–774.
24. Horvath, S. and Dong, J. (2008) Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*; **4**, e1000117.
25. Righetti, K. and Vu, J.L. and Pelletier, S. and Vu, B.L. and Glaab, E. and Lalanne, D. and Pasha, A. and Patel, R.V. and Provart, N.J. and Verdier, J. and others. (2015) Inference of longevity-related genes from a robust coexpression network of seed maturation identifies regulators linking seed storability to biotic defense-related pathways. *The plant cell*; **27**, 2692–2708.
26. Hickman, R. and van Verk, M.C. and Van Dijken, A.J.H. and Pereira Mendes, M. and Vroegop-Vos, I.A. and Caarls, L. and Steenbergen, M. and Van Der Nagel, I. and Wesseling, G.J. and Jironkin, A. and Talbot, A. and Rhodes, J. and de Vries, M. and Schuurink, R.C. and Denby, K. and Pieterse, C.M.J. and Van Wees, S.C.M. (2017) Architecture and dynamics of the jasmonic acid gene regulatory network. *The Plant Cell*; **tpc-00958**.
27. Ruiz-Sola, M. and Coman, D. and Beck, G. and Barja, M.V. and Colinas, M. and Graf, A. and Welsch, R. and Rütimann, P. and Bühlmann, P. and Bigler, L. and others. (2016) Arabidopsis GERANYLGERANYL DIPHOSPHATE SYNTHASE 11 is a hub isozyme required for the production of most photosynthesis-related isoprenoids. *New Phytologist*; **209**, 252–264.
28. Guerin, C. and Joet, T. and Serret, J. and Lashermes, P. and Vaissayre, V. and Agbessi, M.D.T. and Beule, T. and Severac, D. and Amblard, P. and Tregebar, J. and others. (2016) Gene coexpression network analysis of oil biosynthesis in an interspecific backcross of oil palm. *The Plant Journal*; **87**, 423–441.
29. Tantong, S. and Pringsulaka, O. and Weerawanich, K. and Meeprasert, A. and Rungrotmongkol, T. and Sarnthima, R. and Roytrakul, S. and Sirikantaramas, S. (2016) Two novel antimicrobial defensins from rice identified by gene coexpression network analyses. *Peptides*; **87**, 7–16.
30. Birkenbihl, R.P. and Liu, S. and Somssich, I.E. (2017) Transcriptional events defining plant immune responses. *Current opinion in plant biology*; **38**, 1–9.
31. Caputi, L. and Franke, J. and Farrow, S.C. and Chung, K. and Payne, R.M. E. and Nguyen, T.-D. and Dang, T.-T. T. and Soares Teto Carqueijeiro, I. and Koudounas, K. and Dugé de Bernonville, T. and Ameyaw, B. and Jones, D. M. and Vieira, I. J. C.o and Courdavault, V. and O’Connor, S. E. (2018) *Science*; **360**, 1235–1239.
32. Sibout, R. and Proost, S. and Hansen, B.O. and Vaid, N. and Giorgi, F.M. and Ho-Yue-Kuang, S. and Legée, F. and Cézart, L. and Bouchabké-Coussa, O. and Soulhat, C. and others. (2017) Expression atlas and comparative coexpression network analyses reveal important genes involved in the formation of lignified cell wall in *Brachypodium distachyon*. *New Phytologist*; **215**, 1009–1025.
33. Obayashi, Takeshi and Aoki, Yuichi and Tadaka, Shu and Kagaya, Yuki and Kinoshita, Kengo. (2017) ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant and Cell Physiology*; **59**, e3–e3.
34. Schaefer, R.J. and Michno, J.-M. and Myers, C.L. (2017) Unraveling gene function in agricultural species using gene co-expression networks. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*; **1860**, 53–63.
35. Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *science*; **286**, 509–512.
36. Bolger, A.M. and Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*; **30**, 2114–2120.
37. Kauffmann, A. and Rayner, T.F. and Parkinson, H. and Kapushesky, M. and Lukk, M. and Brazma, A. and Huber, W. (2009) Importing arrayexpress datasets into r/bioconductor. *Bioinformatics*; **25**, 2092–2094.
38. Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*; **286**, 509–512.
39. Mutwil, M. and Klie, S. and Tohge, T. and Giorgi, F.M. and Wilkins, O. and Campbell, M.M. and Fernie, A.R. and Usadel, B. and Nikoloski, Z. and Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell*; **tpc.111.083667**.
40. Proost, Sebastian and Mutwil, Marek. (2017) BPlaNet: comparative co-expression network analyses for plants. In Editor,A. and Editor,B. (eds),



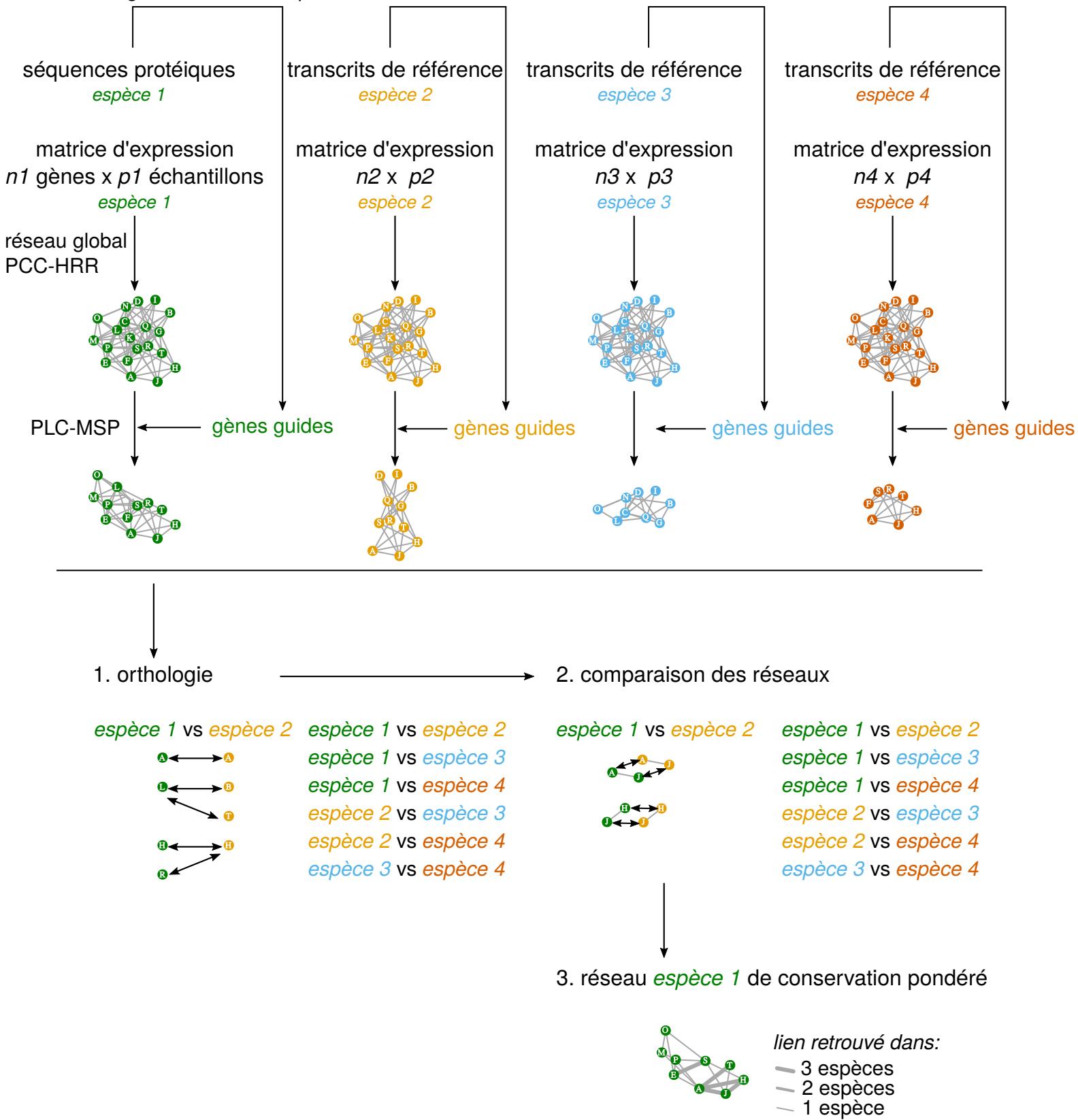
- Plant Genomics Databases*, Springer, pp. 213–227.
41. R Core Team. (2018) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, <https://www.R-project.org/>.



## Partie III

**Associations préférentielles au sein des voies de signalisation de type phospho-relais à étapes multiples : étude comparative de réseaux de co-expression ciblés chez 15 espèces végétales**

recherche de domaines fonctionnels  
homologies Pfam: HisK, Hpt, RR



**Figure 25: Comparaison multi-espèce de réseaux de co-expression.** L'exemple présenté ci-dessus porte sur 4 espèces différentes. La première étape consiste à produire les réseaux global après calcul des distances par PCC-HRR et à en extraire les gènes co-exprimés avec les acteurs de la voie de signalisation des MSP par l'approche PLC. Les gènes utilisés en tant que guides ont été identifiés en scannant les séquences protéiques de référence contre la base Pfam (Kinase A domain (PF00512.22), Hpt (PF01627.20) et Response regulator receiver domain (PF00072.21)). Les PLC sont alors successivement comparées à deux niveaux, tout d'abord par définition des liens d'orthologie entre espèces (1.) puis par comparaison des liens de co-expressions impliquant des orthologues (2.). Une visualisation efficace des liens conservés consiste à utiliser la PLC d'une espèce et à pondérer les liens en fonction du nombre de comparaisons dans lesquels ce lien est retrouvé.

La voie de signalisation répondant aux cytokinines (CK) chez les plantes est du type MultiStep Phosphorelay (MSP). De manière à mieux comprendre comment fonctionne cette voie décrite dans le contexte, notre méthodologie de construction de réseaux de co-expression et de PLC a été appliquée à cette voie. Les premières PLC ciblées sur la voie CK ont été décrites chez *A. thaliana* dans la partie 1. Bien qu'ayant une performance apparente moins grande que celle observée pour des voies métaboliques, ces PLC ciblées sur la voie CK ont clairement montré des spécificités transcriptionnelles pour chaque acteur de cette voie (récepteurs de type Histidine Kinase, HK ; phosphotransférase, Hpt ; régulateurs de réponses, RR). Dans cette 3<sup>e</sup> partie, nous avons étendu l'approche PLC à plusieurs espèces végétales pour étudier la conservation de ces associations spécifiques dans la lignée verte.

Quinze espèces végétales représentant une large diversité taxonomique (algue unicellulaire verte, mousse, gymnospermes et angiospermes) ont été retenues sur la base (i) d'une disponibilité d'une séquence génomique et d'un transcriptome de référence et (ii) du nombre d'échantillons RNA-seq disponibles dans les bases de données. La méthodologie utilisée pour cette analyse de génomique comparative est expliquée dans la **Figure 25** en prenant un exemple avec 4 espèces. La première étape a été de déterminer les sets de gènes guides. Nous avons choisi d'étendre l'analyse de la voie CK à l'ensemble des voies MSP chez les plantes, en incluant les récepteurs à l'éthylène et à la lumière (phytochromes) pour mettre en évidence d'éventuelles associations transcriptionnelles entre ces voies. Les gènes guides ont été identifiés chez les 15 espèces végétales grâce au criblage des séquences protéiques prédictives contre des domaines Pfam retrouvés dans les acteurs des voies MSP. Des PLC utilisant ces gènes guides ont alors été extraites des réseaux globaux de chaque espèce inférés avec les PCC-HRR. La comparaison des réseaux entre espèces implique deux phases : (i) déterminer les liens d'orthologie entre tous les gènes retenus dans les PLC (l'orthologie peut aussi se déterminer en amont sur les protéomes entiers) et (ii) déterminer les liens de co-expression conservés entre espèces. La visualisation de la conservation de ces liens s'est faite en s'appuyant sur la PLC d'*A. thaliana*.

Bien que présentant un état d'avancement moindre que les deux parties précédentes, cette 3<sup>e</sup> partie ouvre une perspective très intéressante sur le fonctionnement des voies MSP chez les plantes. Les résultats décrits dans cette partie devront encore faire l'objet d'un approfondissement. En particulier, il est prévu de construire les réseaux non plus sur les matrices complètes mais selon l'agrégation de réseaux décrite dans la partie 2. De plus,



l'analyse d'orthologie et l'alignement des réseaux pourraient être améliorer en utilisant de nouveaux algorithmes publiés récemment. Toutefois, en l'état actuel de cette analyse, certaines associations transcriptionnelles ont été retrouvées dans plus de 9 espèces différentes. Une telle conservation au travers d'espèces phylogénétiquement distincts, mise en évidence par des jeux de données vraisemblablement très différents, suggère très fortement que ces associations aient un sens biologique. Ces premiers résultats renforcent l'intérêt des analyses de réseaux de co-expression et de les comparer entre espèces si possible.

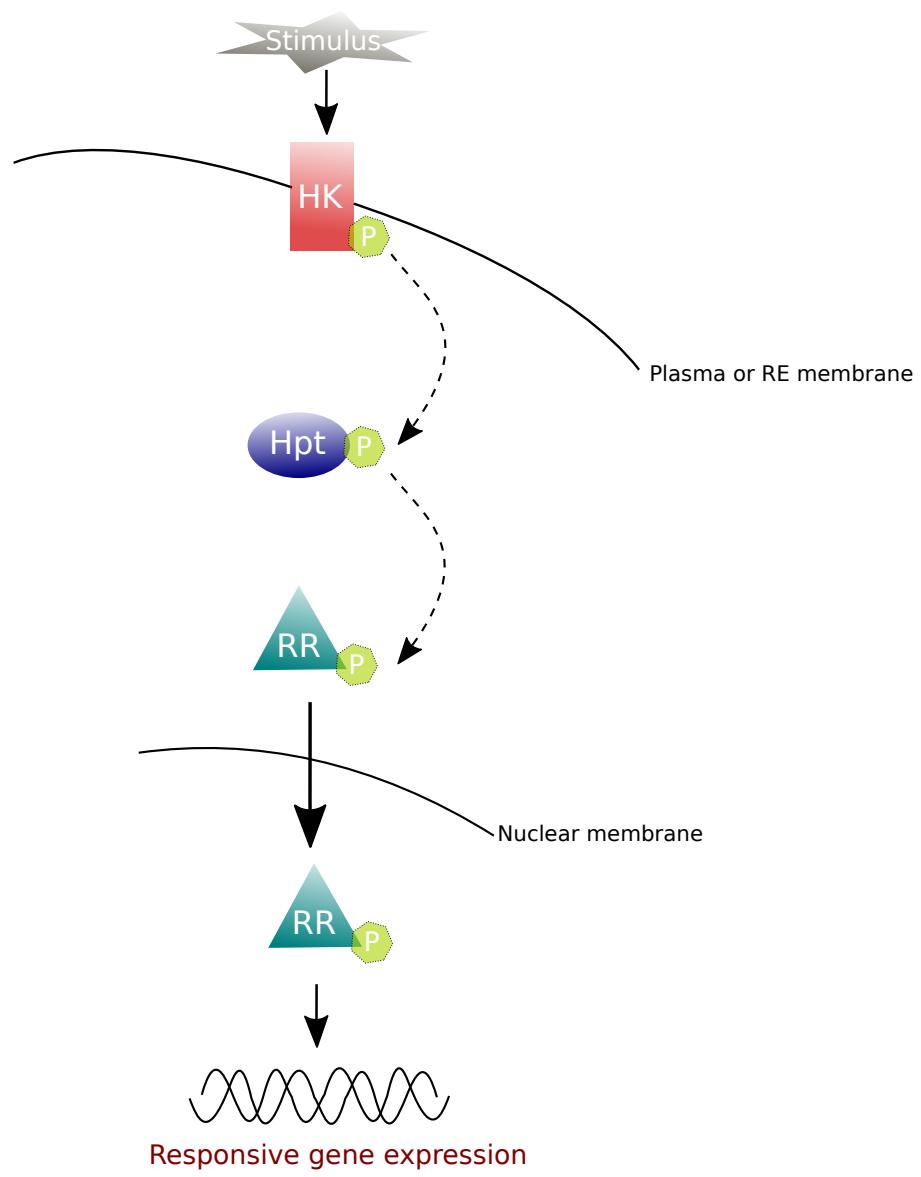


# **Insights in the multiple-step phosphorelay pathway: comparative study of targeted co-expression networks in 15 plant species**

## **Introduction**

Two-component systems (TCS) are an ancient signaling transduction mechanism found in archaea, bacteria as well as in eukaryotes. Originally, these pathways involve two proteins: a receptor which is an Histidine Kinase (HK) and a Response Regulator (RR). In multi-step phosphorelays (MSP) found in eukaryotes, Hybrid Histidine Kinase also carry a REC domain and involve a Histidine Phosphotransferase (Hpt) acting as a go-between for the receptor and the response regulator (Schaller, Kieber & Shiu, 2008). Bacteria use TCS to respond to a large spectrum of stimuli such as stress signaling, nutrient regulation, pathogen-host interactions and chemotaxis (Chang & Stewart, 1998). In eukaryotes, TCS are found specifically in many phylum but not in higher animals where this type of signal transduction was probably replaced by Serine/Threonine and Tyrosine kinases (Stock, Robinson & Goudreau, 2000). In higher plants, as the model species *Arabidopsis thaliana*, MSP pathways are involved in a wide range of physiological processes involving phyhormone (ethylene and cytokinin) signaling, responses to light and osmosensing required for proper plant growth and development (Schaller, Kieber & Shiu, 2008).

In canonical MSP pathway (**Figure 1**), the homodimeric HK binds or senses the signaling pathway activating molecule which leads to a dimerization of the receptor followed by an teautophosphorylation on the histidine residue. This phosphate group is then transferred to the receptor's aspartate residue in order to be shifted on the aspartate residue of the Hpt. This protein is able to cross the nucleus' membrane to transfer the phosphate group to the last actor of the signaling pathway: the RR. These proteins can act as transcription factors and bind the DNA (RR-B) or interact with other proteins and leading to negative feedback mechanisms (RR-A). In 2004, were identified a third class



**Figure 1: Schematic representation of the Multi Step Phosphorelay (MSP) pathway.** The receptor containing a Histidine Kinase (HK) domain binds/senses the stimulus and subsequently auto-phosphorylates. This leads to a phosphorylation cascade via a Histidine-Phosphotransferase (HPT) up to a Response Regulator (RR) which interacts with DNA and elicits transcription of target genes.

of response regulators, RR-C which functions are still not entirely established (Kiba et al., 2004). Last but not least, there is the small family of Pseudo Response Regulators (PPR) which are known to be involved in circadian rhythm control (Nakamichi et al., 2005). This linear phosphorylation cascade is typically observed for the cytokinin transduction pathway (Hwang, Sheen & Müller, 2012) after binding of cytokinin on the HK's CHASE domain. Concerning ethylene signaling, although the role of HK and aspartate receiver domain's phosphorylation remains unclear, it seems that it may play a crucial role in signaling modulation (Etheridge, Hall & Schaller, 2006). Direct interactions of ethylene Histidine Kinases ETR1 and ERS1 between Histidine Phosphotransferases and response regulators from the MSP pathway have been highlighted in yeast as well as in plant (Urao et al., 2000; Cho & Yoo, 2006; Binder et al., 2018).

Both, Cytokinin and Ethylene are largely involved in developmental processes such as leaf senescence, germination and flowering (Riefler, 2006; Bartrina et al., 2011; Raines et al., 2016). As these activities are intimately related to light and circadian regulation, a link between light signaling through Phytochromes carrying HK domains and these signaling pathways appears as obvious. For example, Phytochrome A expression has been shown to be directly up-regulated by cytokinins (Brenner et al., 2005) and Phytochrome B to interact with ARR-A 4 to modulate red-light signaling in response to cytokinin in photomorphogenesis (Sweere et al., 2001; Mira-Rodado et al., 2007). How these crosstalks occur is currently not understood, but different HK may recruit specific Hpt or RR.

Even more complexity is added by the high number of different actors at each level of the signaling pathway. In *Arabidopsis thaliana* for example, sixteen HK (cytokinin, ethylene signaling pathways receptors, osmosensors and phytochromes), six Hpt and around twenty RR has been identified (Kieber & Schaller, 2014). The large amount of proteins and the resulting number of possible combinations among them, leads to the expectation that preferential and specific interactions among them are involved in particular physiological



responses. Indeed, some studies have highlighted such specific interactions as for example Mirado-Rodado et al. who had shown the involvement of AHK5, AHP2 and ARR4 in stomatal closure upon H<sub>2</sub>O<sub>2</sub> and ethylene in *Arabidopsis* and Sun et al. who had demonstrated the role of OsAHP1 and OsAHP2 in response to salt and drought stresses in rice (Mira-Rodado et al., 2012; Sun et al., 2014).

Considering the huge number of possible combinations among MSP actors, it appears necessary to prioritize candidate associations for wet lab validation. Following the “guilt by association” (GBA) principle (Wolfe, Kohane & Butte, 2005), genes which are co-expressed are likely to play a role in a same physiological process and therefore to interact, directly or indirectly, in a given biological response (Wolfe, Kohane & Butte, 2005). Thereby we can expect that genes coding for MSP-related proteins will cluster together if specific associations exist among them in particular processes. In a previous work, we had highlighted preferential associations in the cytokinin signaling pathway in *Arabidopsis* by using Pathway Level Co-expression (PLC) networks based on large-scale microarray and RNA-seq data (Wei et al., 2006; Liesecke et al., 2018).

In this study, we focus on specific interactions of MSP actors in several plant species. The question is how are these putative interactions conserved among different species in the plant kingdom? Co-expression is here a useful tool as it permits to put in evidence those preferential associations and moreover to compare the different organisms. As a matter of fact, comparing co-expression patterns based on large-scale data, permits to overcome limitations of comparative expression analysis which may be biased by non equivalent experimental conditions in the input data set and allows to bring out genes sharing similar functions and involved in a same biological process as well as studying their evolution (Stuart et al., 2003; Bergmann, Ihmels & Barkai, 2004; Tirosh, Bilu & Barkai, 2007; Ruprecht et al., 2016). Feltus et al. for example have compared global co-expression networks in maize and rice to identify modules of genes who share similar functional terms and that are putatively conserved between these two species (Ficklin & Feltus, 2011). Hansen et al. for their part have focused on a specific process: cellulose



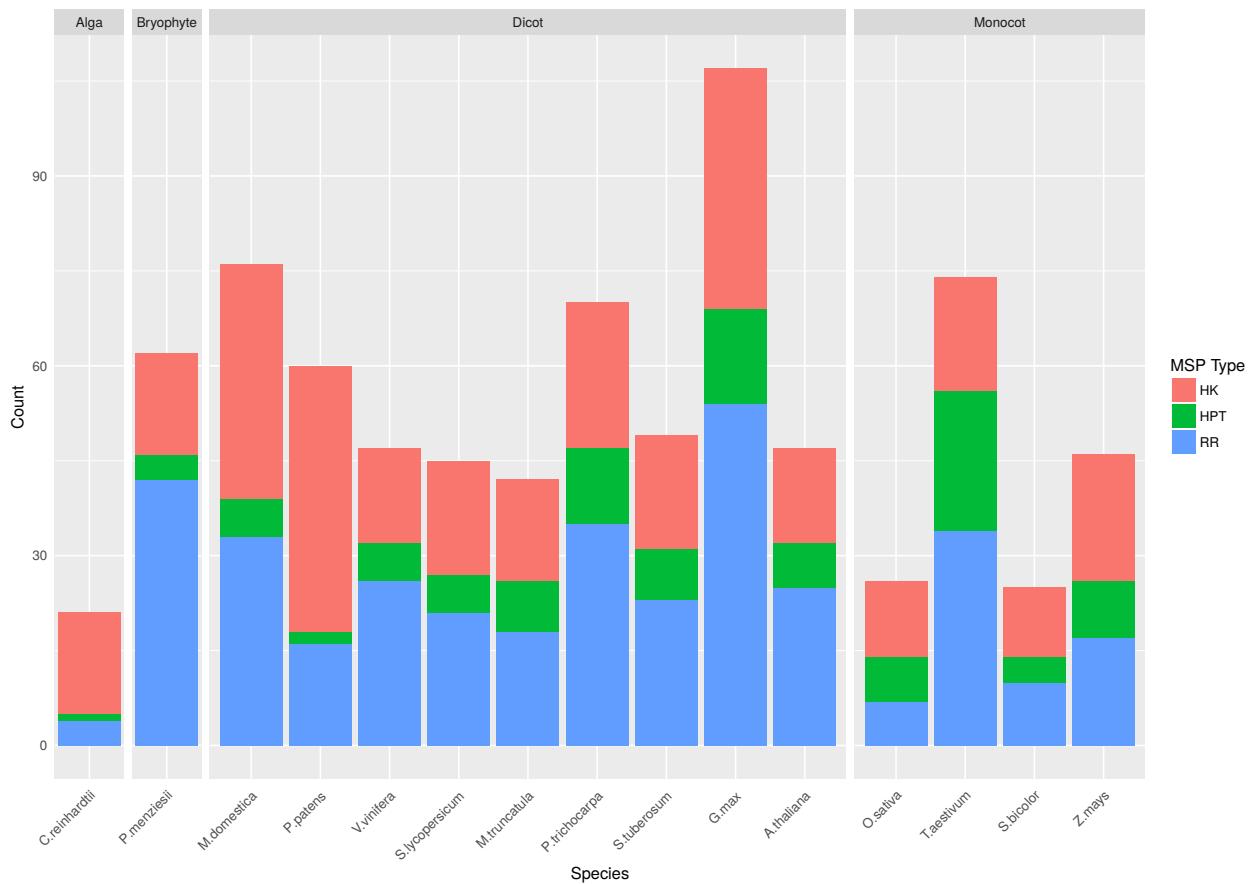
biosynthesis in *Arabidopsis* and rice by comparing co-expression networks (Hansen et al., 2014). An online tool called “Network-Comparer” which permits to compare conserved co-expression networks in seven plant species is available on the PlaNet platform (<http://aranet.mpimp-golm.mpg.de>) (Mutwil et al., 2011).

Using inter-species comparative co-expression has a double purpose in this study. Firstly, beyond simple sequence similarity analysis, the aim was to investigate the conservation and putative evolution of preferential interactions among actors of the MSP pathway based on their co-expression profiles. This may lead to a better understanding of the MSP pathway’s organization in the different species but also help to annotate unknown genes in non model species. We also expect that co-occurring edges between different species may help to limit the rate of false positive results as most biologically relevant associations are more likely to be conserved among different species (Hansen et al., 2014).

## Results

### 1. From species to genes

We considered plant species with a large evolutionary diversity in the green lineage. Our sample species contain an alga (*Chlamydomonas reinhardtii*), a bryophyte (*Physcomitrella patens*), a coniferous (*Pseudotsuga menziesii*), monocot model species (*Oryza sativa*, *Sorghum bicolor*, *Zea mays* and *Trichocarpa aestivum*) and finally different dicots (*Solanum lycopersicum* and *tuberosum*, *Vitis vinifera*, *Medicago truncatula*, *Malus domestica*, *Populus trichocarpa*, *Glycine max* and of course the model plant *Arabidopsis thaliana*). In order to determine genes involved in MSP, we first identified genes which sequences contain the Pfam domains corresponding to the His Kinase A (phospho-acceptor) domain (PF00512.22), the Hpt domain (PF01627.20) and the Response regulator receiver domain (PF00072.21) (but not the His Kinase A domain as receptors are often Hybrid HK which contains RR domains (Stock, Robinson & Goudreau, 2000)(See Material and Methods). We observed



**Figure 2: Distribution of MSP related genes.** For each species, potential guide genes for Pathway Level Co-expression network construction were identified based on their PFAM domains. The different species show an heterogeneous amount of MSP-related genes though display a constant distribution at the different pathway levels (amount of RR > amount of HK > amount of HPT).

a large discrepancy in the amount of MSP related genes across the different species. The number of identified MSP-related genes seems to be directly related to the organism's amount of predicted genes as plants with larger genomes tend towards having more genes containing MSP-related domains (**Supplemental data 1**). Noteworthy, we found no relation between the number of identified genes and the organism's phylogenetic position (**Figure 2**). In the selected dicots for example, this number can vary in a ratio of more than one to two (*V. vinifera* 54 versus *G. max* 190 identified genes) and no clear tendency was observed among the four groups. Notably, the most ancient organism of this study, *C. reinhardtii* displays a small amount of genes compared to the remaining species. All the species display more or less the same distribution of genes at each level of the pathway as already described in *Arabidopsis* (Schaller, Kieber & Shiu, 2008), *i.e.* a large amount of response regulators, less histidine kinases and a relatively small number of histidine phosphotransferases. It might be conceivable that the response specificity is mainly driven by the receptor and the response regulator while the Hpt only acts as a shuttle between these two elements and thus its function is mainly limited to translocate the phosphate group from the cytoplasm to the nucleus.

### **1.1 Identifying MSP-related genes orthologs in 15 species**

As the MSP pathway is the most well-known in terms of annotated genes and preferential associations in the model plant *A. thaliana*, we used this information to annotate the genomes of the other species (Table I). We performed an ortholog analysis with OrthoDB using the sequences of the whole proteome of each species as input. The OrthoDB orthology performs a delineation procedure between genes from each species pair, determined from all-against-all Smith-Waterman protein sequence comparisons (Waterhouse et al., 2013). The genes containing relevant Pfam domains and found in the same sequence cluster than an *Arabidopsis* MSP-related gene (**Supplemental data 2**) were retained as guide genes for the further PLC network construction. Due to high sequence similarity some *Arabidopsis* MSP-related genes were attributed to the same cluster. This was the case of the phytochromes B and D,



the ethylene receptors ETR1 and ETR1 as well as EIN4 and ETR2. Unsurprisingly, as their sequences are strongly conserved, all histidine phosphotransferases excepting the pseudo-HPT AHP6 (probably because of the mutation in the conserved histidine residue (Muller & Sheen, 2007) were attributed to a same cluster. This lower divergence support a shuttle role of AHPs rather than a role as a specificity actor. The large part of the ARR-A set (ARR3, ARR4, ARR5, ARR6, ARR7, ARR8, ARR9, ARR15, ARR16, ARR17) was also found in a unique OrthoDB cluster revealing strong identity between these proteins. Other non unique clusters regrouped ARR1 and ARR2, APRR13 and APRR21 and finally APRR5 and APRR9. This illustrates the difficulty of precise gene annotation based on nucleic or proteic sequences alone. We can expect that studying co-expression will help to fine-tune the annotation of these elements as genes which share common conserved functions are expected to display similar co-expression patterns.

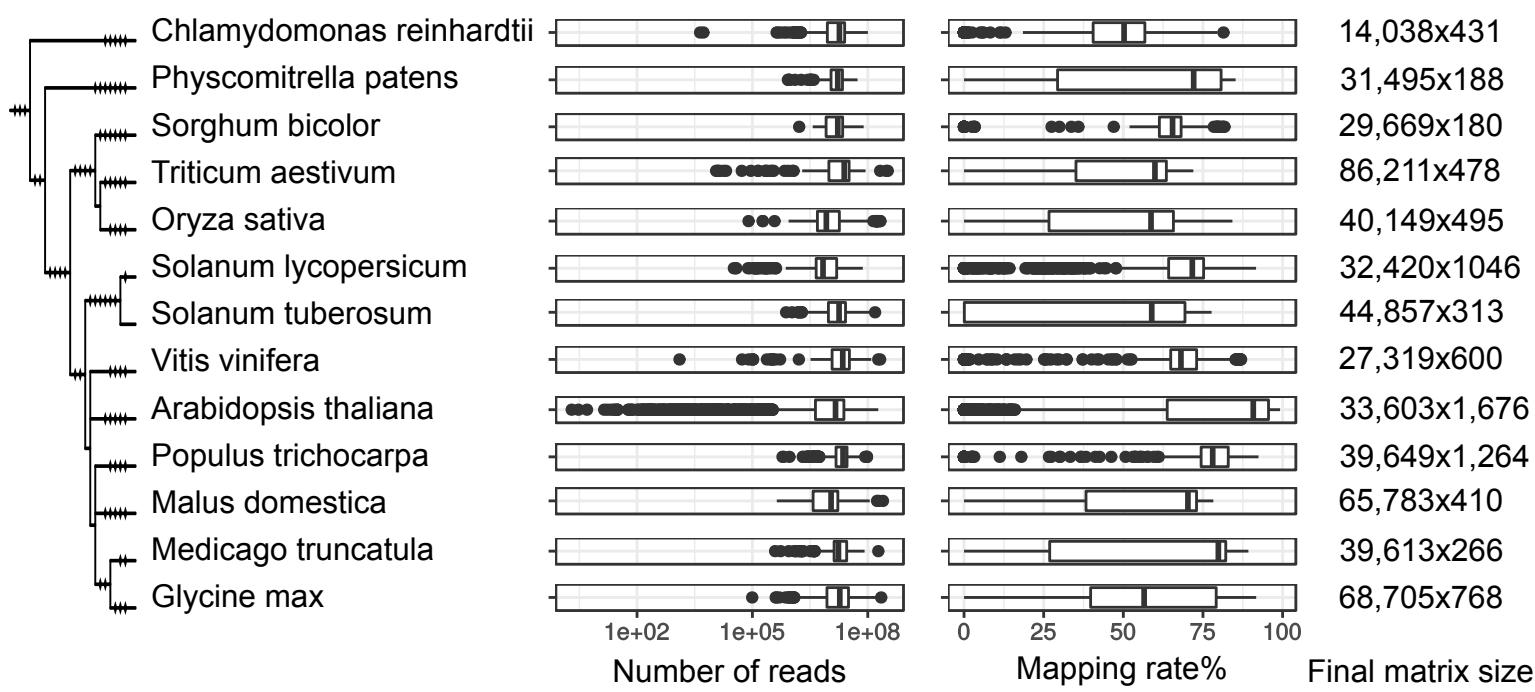
**Table I: MSP-related genes in *Arabidopsis thaliana* used as reference for OrthoDB clustering.**

Locus tag	Gene name	MSP level	Type
AT2G17820.1	AHK1	Receptor	Osmosensor
AT5G35750.1	AHK2	Receptor	Cytokinin signaling
AT1G27320.1	AHK3	Receptor	Cytokinin signaling
AT2G01830.2	AHK4	Receptor	Cytokinin signaling
AT5G10720.1	AHK5	Receptor	Cytokinin independent
AT2G47430.1	CKI1	Receptor	Cytokinin independent
AT1G09570.1	PHYA	Receptor	Phytochrome
AT2G18790.1	PHYB	Receptor	Phytochrome
AT5G35840.1	PHYC	Receptor	Phytochrome
AT4G16250.1	PHYD	Receptor	Phytochrome
AT4G18130.1	PHYE	Receptor	Phytochrome
AT2G40940.1	ERS1	Receptor	Ethylene signaling
AT1G04310.1	ERS2	Receptor	Ethylene signaling
AT3G04580.1	EIN4	Receptor	Ethylene signaling
AT3G23150.1	ETR2	Receptor	Ethylene signaling
AT1G66340.1	ETR1	Receptor	Ethylene signaling
AT3G21510.1	AHP1	Transmitter	-
AT3G29350.1	AHP2	Transmitter	-
AT5G39340.1	AHP3	Transmitter	-
AT3G16360.2	AHP4	Transmitter	-
AT1G03430.1	AHP5	Transmitter	-
AT1G80100.1	AHP6	Transmitter	-
AT4G04402.1	AHPx	Transmitter	-
AT3G16857.2	ARR1	Response Regulator	Type-B
AT4G16110.1	ARR2	Response Regulator	Type-B
AT1G59940.1	ARR3	Response Regulator	Type-A



AT1G10470.1	ARR4	Response Regulator	Type-A
AT3G48100.1	ARR5	Response Regulator	Type-A
AT5G62920.1	ARR6	Response Regulator	Type-A
AT1G19050.1	ARR7	Response Regulator	Type-A
AT2G41310.1	ARR8	Response Regulator	Type-A
AT3G57040.1	ARR9	Response Regulator	Type-A
AT4G31920.1	ARR10	Response Regulator	Type-B
AT1G67710.1	ARR11	Response Regulator	Type-B
AT2G25180.1	ARR12	Response Regulator	Type-B
AT2G27070.1	ARR13	Response Regulator	Type-B
AT2G01760.1	ARR14	Response Regulator	Type-B
AT1G74890.1	ARR15	Response Regulator	Type-A
AT2G40670.2	ARR16	Response Regulator	Type-A
AT3G56380.1	ARR17	Response Regulator	Type-A
AT5G58080.1	ARR18	Response Regulator	Type-B
AT1G49190.2	ARR19	Response Regulator	Type-B
AT3G62670.1	ARR20	Response Regulator	Type-B
AT5G07210.1	ARR21	Response Regulator	Type-B
AT3G04280.1	ARR22	Response Regulator	Type-A
AT5G62120.1	ARR23	Response Regulator	Type-B
AT5G26594.1	ARR24	Response Regulator	Type-A
AT5G61380.1	APRR1	Pseudo Response Regulator	-
AT4G18020.1	APRR2	Pseudo Response Regulator	-
AT5G60100.2	APRR3	Pseudo Response Regulator	-
AT5G49240.1	APRR4	Pseudo Response Regulator	-
AT5G24470.1	APRR5	Pseudo Response Regulator	-
AT1G68210.1	APRR6	Pseudo Response Regulator	-
AT5G02810.1	APRR7	Pseudo Response Regulator	-
AT4G00760.1	APRR8	Pseudo Response Regulator	-
AT2G46790.1	APRR9	Pseudo Response Regulator	-

Phylogenetic trees were constructed for genes associated to each step of the MSP pathway (i.e. receptor, transmitter and response regulator (respectively **Supplemental data 3A, B and C**)). Complementary to the OrthoDb sequence clustering and in order to visualize sequence based ortholog relationships. Once again, *Arabidopsis* genes were used as reference in order to fine-tune gene annotation. Concerning the receptors, an overall good attribution of a majority of genes in relationship with *Arabidopsis* were obtained except for 10 genes coming from Douglas fir, wheat, rice, maize and sorgho (which could be annotated by the previous OrthoDB analysis). For the transmitter genes, a distinction among the different Histidine Phosphotransferases was possible only between AHP6 and the remaining AHP using OrthoDB as the *Arabidopsis* sequences are highly similar, whereas the analysis based on the phylogenetic tree for this family of MSP actors permits at least a distinction between AHP1, AHP4, AHP6 and the other histidine phosphotransferases. The tree based on



**Figure 3: RNAseq datasets characteristics.** For each species, available RNAseq data were retrieved and pre-processed to establish expression matrices (see Material and Methods). Here the distribution of the number of reads, the mapping rate as well as the dimensions of the resulting expression matrix are displayed for each one.

Response regulator sequences permits a better segregation for some genes, in particular the ARR-A genes which are all attributed to a same cluster by the clustering algorithm. Nevertheless, some RR are found at a huge distance from *Arabidopsis* RR making it difficult to establish the correspondence.

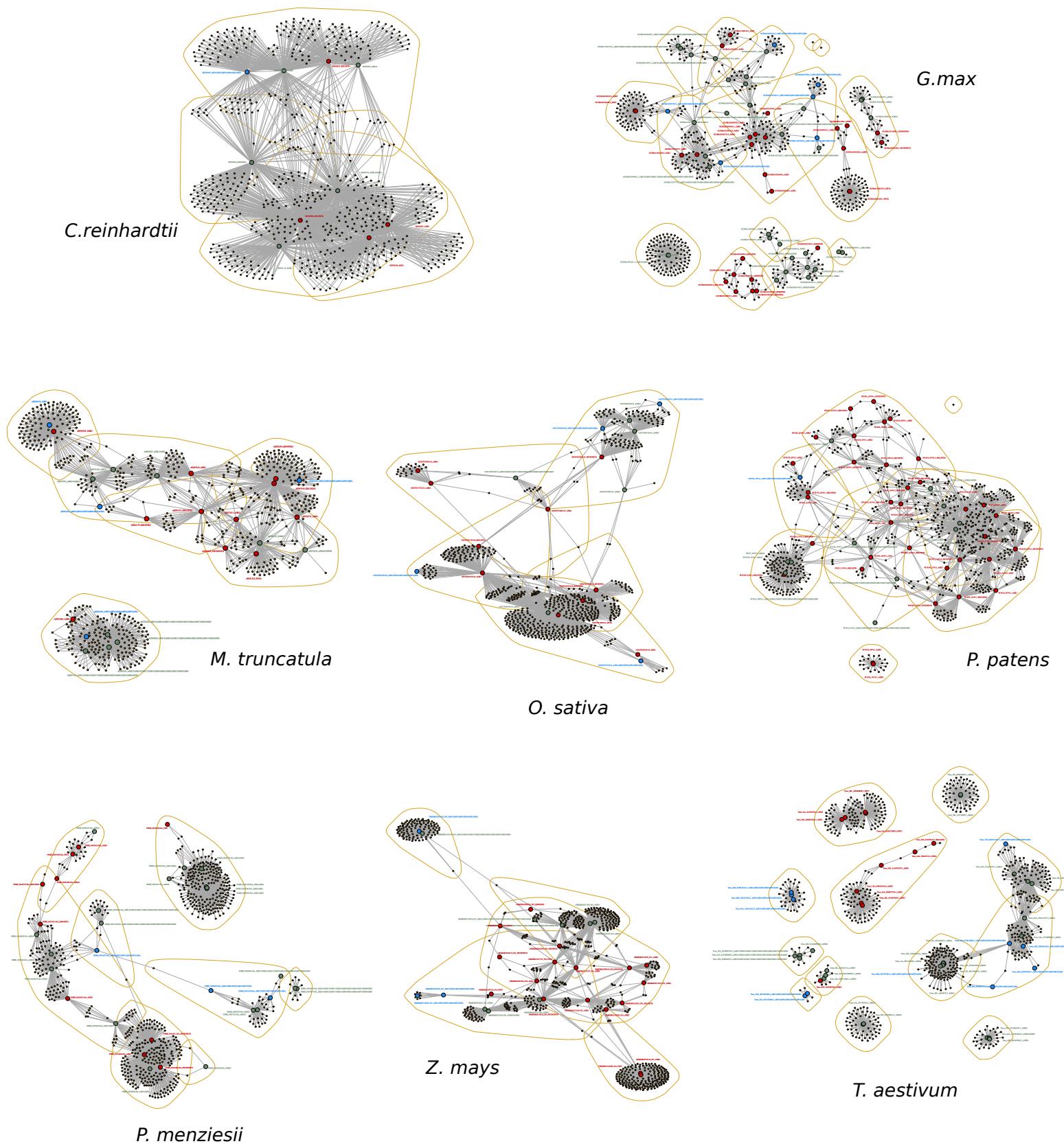
Even if the OrthoDB annotation based *Arabidopsis* identifiers failed to annotate very similar protein sequences, we further used the OrthoDB clusters to refine gene locus tags. Moreover this method appears more robust as it takes in account each species' whole genome sequences. Altogether, both methods show that a classification in order to deduce potential gene functions based only on sequence similarity with genes of a model species is not sufficient.

## 2. From genes to Pathway Level Co-expression networks

### 2.1 Network construction and analysis

For each species, available RNA-seq data were retrieved, processed and combined to an expression matrix (See Materiel and Methods for details). The resulting matrices were heterogeneous in number of genes (with a minimum of 14,038 genes for the green algae *Chlamydomonas rheinhardtii* and a maximum of 86,211 genes for the hexaploid *Triticum aestivum*) as well as in the number of experimental conditions varying from 180 for *Sorghum bicolor* to 1,676 for the model plant *Arabidopsis thaliana* (**Figure 3**).

Co-expression among every possible gene pair in the considered genome was evaluated by the highest reciprocal rank (HRR), a robust PCC based ranked co-expression measurement method (Mutwil et al., 2011; Liesecke et al., 2018). We used the Pathway Level Co-expression (PLC) approach to build targeted networks focused on the MSP pathway. Genes identified as actors of the MSP pathway for each species and their respective co-expressed genes were retrieved from global networks (see Material and Methods). In order to make the species networks as comparable as possible, a HRR threshold was set that the networks contain approximatively the same number of vertices (or genes). Considering the large number of studied species, an acceptable



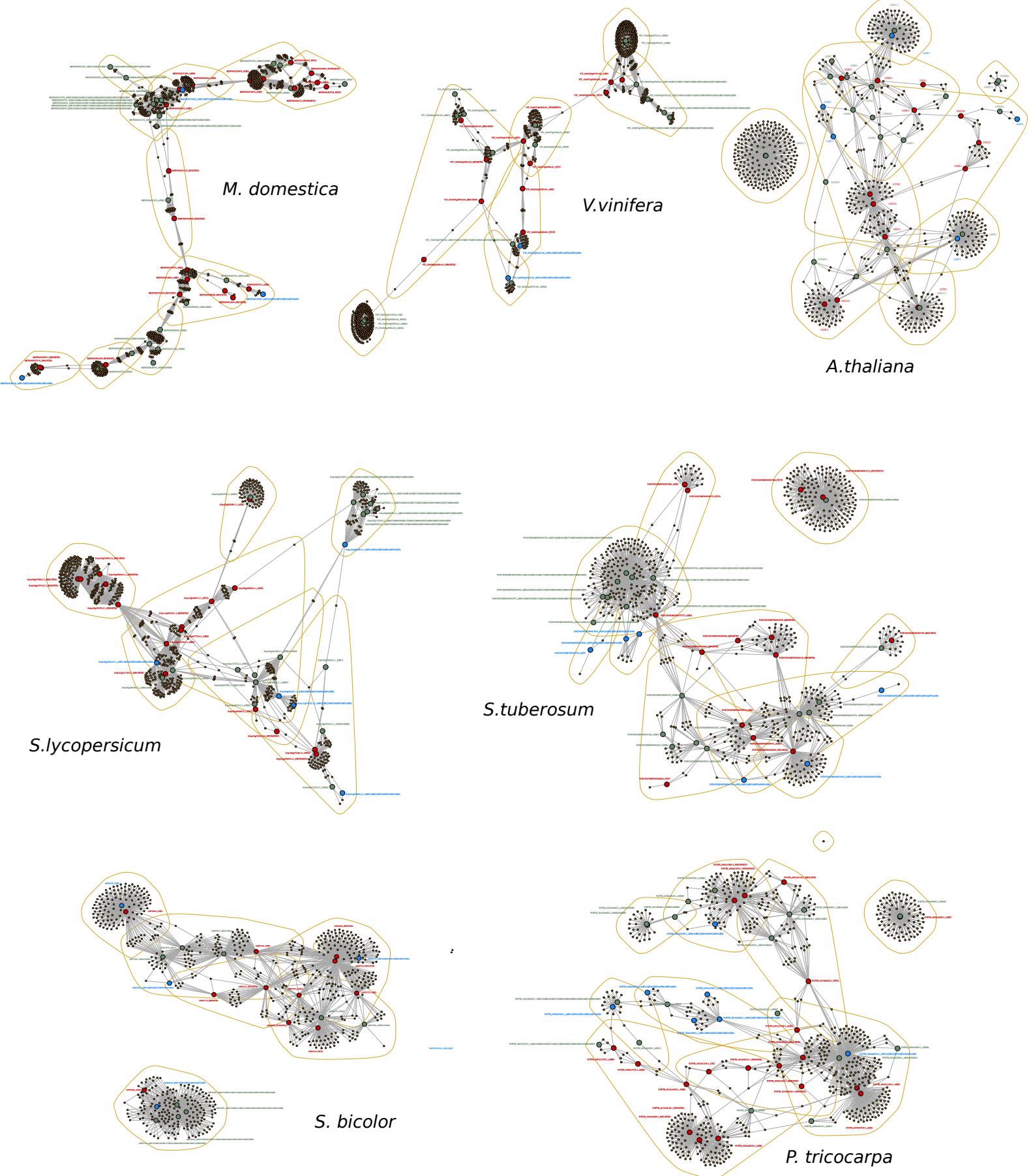
**Figure 4A: Pathway Level Correlation networks from the 15 species.** PLC networks were build to contain ca. 750 vertices. Receptors are displayed in red, Histidine Phosphotransferases in blue and Response Regulators in green. Co-expressed genes shared among guide genes (marker genes) are drawn in orange. Communities are surrounded in orange as identified by the clustering algorithm. Genes are annotated with their respective locus tag and their Arabidopsis ortholog's names as identified by OrthoDB.

compromise has been found for PLC networks containing *ca.* 750 vertices leading to networks with a satisfying topology, in particular concerning the modularity score, for all of them (**Supplemental data 4**). A community analysis was performed on the networks to identify highly connected gene modules using the fast greedy modularity optimization algorithm (Clauset, Newman & Moore, 2004) in order to highlight preferential associations among guide genes. In our case, the ideal community contains members of each MSP pathway's level (**Figure 4A and B**).

The *Arabidopsis* PLC network displays a huge community containing the four receptors AHK1, AHK4, AHK5, EIN4, the following ARR-A: ARR3, ARR4, ARR5, ARR 6, ARR7, ARR8, ARR9, and an ARR-B ARR10. A second community emerges from the network: it contains the cytokinin receptor AHK2, the phytochromes B and C, the histidine- phosphotransferase AHP6 which has the particularity to inhibit cytokinin signaling (Muller & Sheen, 2007) and the pseudo response-regulators APPR2, 7 and 9. A third community contains the ethylene receptors ERS1 and ETR2, PHYA, AHP2, AHP3 and ARR1. PHYD and AHK3 associated to the Pseudo Response Regulators. APRR5 and APRR1 form another group as well as ETR1 and ARR12. Two communities did not contain any receptor: AHP5 with ARR2 and AHP1 associated with ARR11. The remaining communities contained only genes from a same level of the MSP pathway: ARR21 associated with ARR13 and APRR6 with ARR22. This indicates that in spite of their sequence identity, AHP may be preferentially associated with specific receptors and RR.

In the other species, interesting communities emerged similarly. The model green algae *Chlamydomonas reinhardtii* groups orthologs of ERS1/ETR1, AHP, and ARR1/2 and ARR24. In the bryophyte *Physcomitrella patens*, two orthologs of ERS1/ETR1 are associated in a same community than two ARR24 orthologs and one of ARR3/4/5/6/7/8/9/15/16/17.

Concerning monocots, in rice's co-expression network, a community is formed by the OrthoDB orthologs PHYB/D, two APRR1 and APPR7 related genes and a histidine phosphotransferase (for locus tags see **Supplemental**



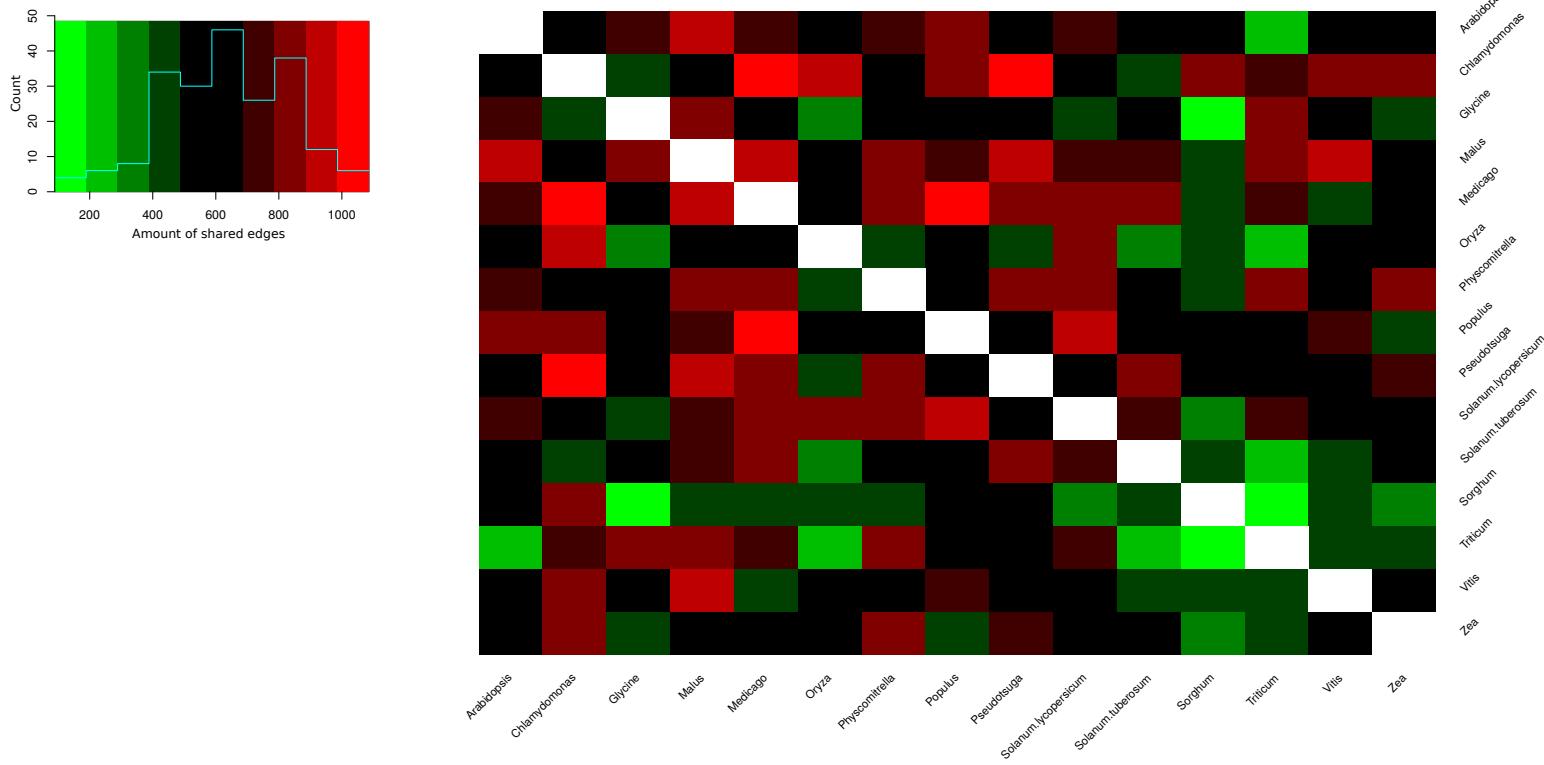
**Figure 4B: Pathway Level Correlation networks from the 15 species.** PLC networks were build to contain ca. 750 vertices. Receptors are displayed in red, Histidine Phosphotransferases in blue and Response Regulators in green. Co-expressed genes shared among guide genes (marker genes) are drawn in orange. Communities are surrounded in orange as identified by the clustering algorithm. Genes are annotated with their respective locus tag and their Arabidopsis ortholog's names as identified by OrthoDB.

**data 5).** The wheat PLC network presents a majority of mono-level communities but a small one is formed by an AHK3 ortholog associated to three APPR7 orthologs. In Sorghum, the ortholog of ethylene receptors EIN4 and ETR2 clusters with three AHP orthologs and a APRR7 ortholog. A huge community is formed by orthologs of phytochromes A, B/D and C, ERS1/ETR1, AHK3, EIN4.ETR2 and APPR5/9. Finally in maize, an interesting community is composed of the orthologs of PHYA, PHYC, AHK3, two PHYB/D, two AHP orthologs and APPR5/9 and APPR1/7.

In dicots, the grape network presents inter alia a community containing orthologs of ERS1/ETR1, ARR1/2, ARR11 and ARR 10/12, another one grouping the PHYA, PHYB/D, ERS1/ETR1, AHK3, EIN4/ETR2, APPR1 and ARR11 orthologs. In tomato, an interesting community is formed by ERS1/ETR1, AHP, ARR1.2 and APRR1 orthologs. The closely related potato presents among others a huge community with orthologs of AHK3, AHK1, ESR1/ETR1, three AHP, APRR1 and APPR5/9 and a group containing PHYE, PHYB/D and APPR5/9 orthologs. The soybean co-expression network displays numerous mono-level communities, but also a community containing orthologs of PHYA, AHK2, PHYE and APPR2 and another one with two PHYB/D and three APPR2. The model legume *Medicago truncatula* also shows a community grouping orthologs of phytochromes (PHYA and PHY B/D) with Pseudo Response Regulator orthologs (APPR 5/9 and APPR7).

In the conifer *Pseudotsuga menziesii* better known as the Douglas fir, one ortholog of AHK4 grouped with ERS1/ETR1, EIN4/ETR2, two ARR24, ARR1/2 and APRR1, while another AHK4 ortholog is grouped with PHYC and two other ARR24. This study also included two other ligneous species: the poplar which displays a community with a PHYA, a PHYB/D, an APRR5/9 and an APPR7 ortholog and the apple tree which network shows interesting communities grouping for example AHK3, PHYB/D, PHYE and APPR2 or PHYA, AHK2, AHK5, APRR5/9 and APPR7 orthologs.

We were able to build informative PLC networks for each species. Even although some species displayed communities grouping only genes from a same level of the MSP pathway, in a majority of networks, communities



**Figure 5: Amount of conserved network edges among species.** Conserved edges between species as identified by pair-wise Pinalog comparison on MSP PLC networks of the different species containing approximatively 1250 vertices.

contained genes from different levels which may match to preferential associations could be highlighted.

Considering the number of species in this study and the subsequent colossal number of analyzed genes, a direct comparison among the co-expression networks was impossible. In order to facilitate such comparison, we performed pairwise alignments between all obtained networks.

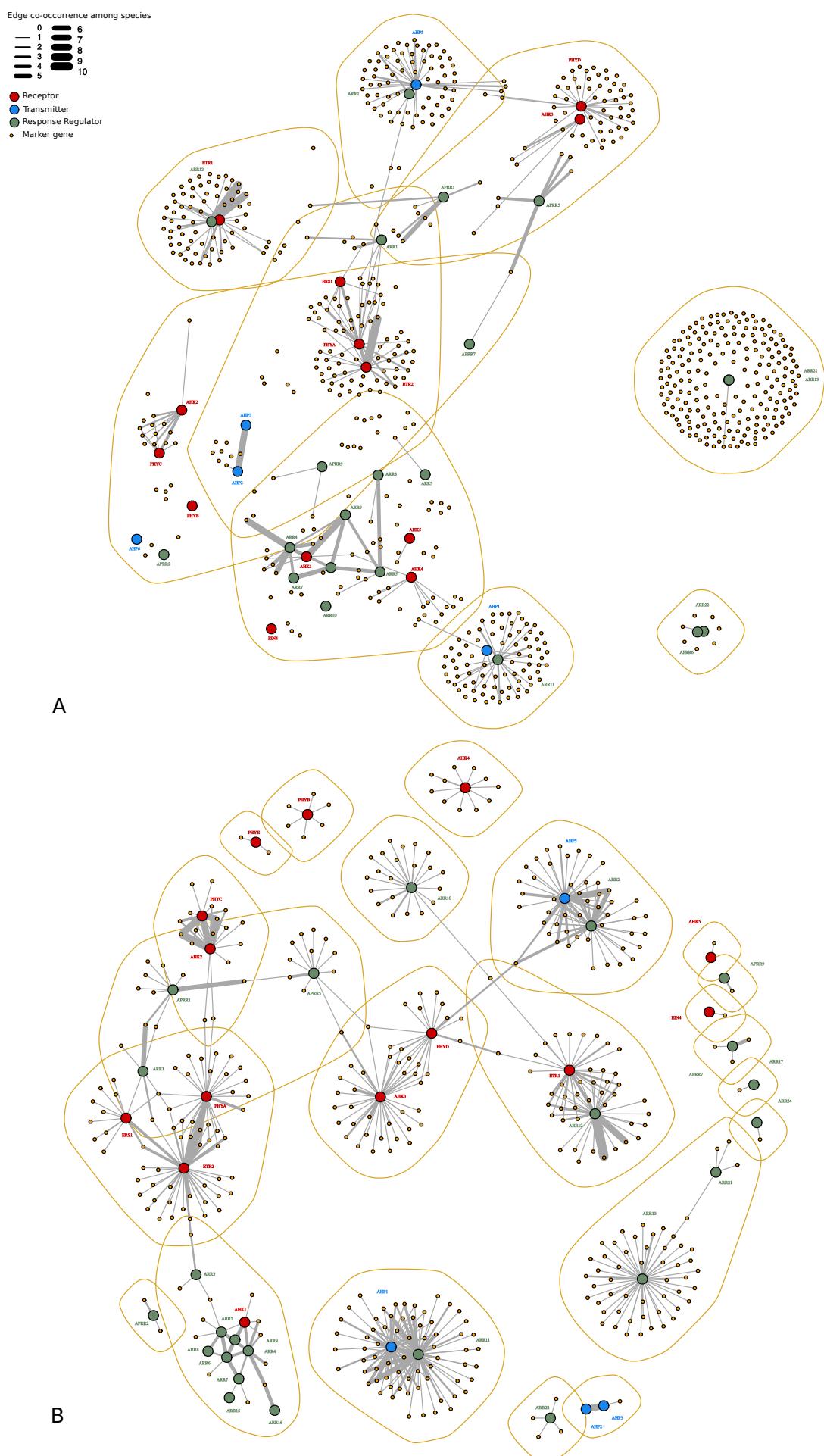
## 2.2 Inter-species network alignment

All the previous networks based on *ca.* 750 vertices were aligned two-by-two using the Pinalog software in order to highlight edges between guide and marker genes and among guide genes which are conserved over different species. As using inter-species edge co-occurrence as condition sine qua non to be retained for network elaboration is highly stringent and expected to decrease false positive rate (Hansen et al., 2014), a lower threshold was set to obtain networks with approximatively 1250 vertices for each species.

On the whole, pairwise aligned initial PLC networks (750 vertices) shared an average of 364 edges with a maximum for the *Pseudotsuga* and *Chlamydomonas* networks which shared 673 edges and a minimum for the intersection of *Triticum* and *Oryza* which shared 52 edges. The larger networks shared an average of 637 edges (1085 edges maximum between *Chlamydomonas* and *Pseudotsuga* and 89 edges at a minimum between *Triticum* and *Sorghum*). Surprisingly, number of co-occurring relationships were not related to evolutionary distances. For example *S. lycopersicum* and *S. tuberosum* shared 586 edges in the initial networks and 703 in the larger ones but larger intersections could be observed with other species (**Figure 5**).

## 2.3 Inter-species network comparison for improved network construction

In order to construct more robust PLC networks, network edge co-occurrence information was integrated to the *Arabidopsis* network to weight edges according to their representation in the other species. (**Figure 6A**). Specific



**Figure 6: *Arabidopsis* PLC networks including cross-species edge conservation information.** Pinalog based on initial networks containing ca. 750 vertices. Receptors are displayed in red, Histidine Phosphotransferases in blue, Response Regulators in green and marker genes are drawn in orange. Communities are surrounded in orange as identified by the clustering algorithm. Edge thickness is proportional to inter-species co-occurrence. A. Initial *Arabidopsis* PLC network with co-occurrence weighted edges B. Conserved *Arabidopsis* PLC network: Based on the initial network but only vertices implied in edges found in *Arabidopsis* and at least one other species were retained for network construction

associations between guide and marker genes clearly emerged (Table II). The edge between the ethylene receptor ETR2 and AT2G44080.1 coding for an ARGOS-like protein involved in organ growth was found in 9 different species. Interestingly, the effects of this gene on plant growth was described as being mediated by ethylene signaling (Rai et al., 2015). APPR1 linked to the abscisic acid activated serine/threonine-protein kinase SRK2D (AT3G50500.2) was found in five species (Medeiros et al., 2015). APPR1 (also called TOC1) has been identified as a key player in circadian rhythm control in interaction with abscisic acid signaling ((Pokhilko, Mas & Millar, 2013). Some guide genes present several highly conserved associations. For example, the Pseudo Response Regulator APPR5, known to be involved in multiple circadian-associated biological processes, displayed 5 conserved edges with more or less described genes. In 4 different species, this gene is associated to AT5G24470.1 coding the GIGANTEA protein which promotes flowering under long days in a circadian clock-controlled flowering pathway (Park et al., 1999). The relationship between APRR5 and the WRKY26 transcription factor represented in 3 species could be an indicator for the existence of crosstalk between light and temperature signaling pathways in plants as pointed out by Song et al. (Song et al., 2016). Indeed, this gene was identified to be potentially involved in heat acclimation (Li et al., 2011).

Further we constructed the *Arabidopsis* network based on co-expression relationships which are not exclusive to this species to evaluate how retaining only edges found in multiple species affects the network topology and in this way the community organization (**Figure 6B**).

The network modularity as well as the transitivity were slightly increased (Modularity = 0.8695 and Transitivity = 0.0377 versus respectively 0.8103 and 0.0138 in the initial network) and 461 vertices linked by 582 edges were represented in at least one other species. A majority of guide gene associations were conserved as described in Table III. The reorganized network also displayed a relationship which didn't appear in the initial network among ARR1, APRR5 and APRR1.

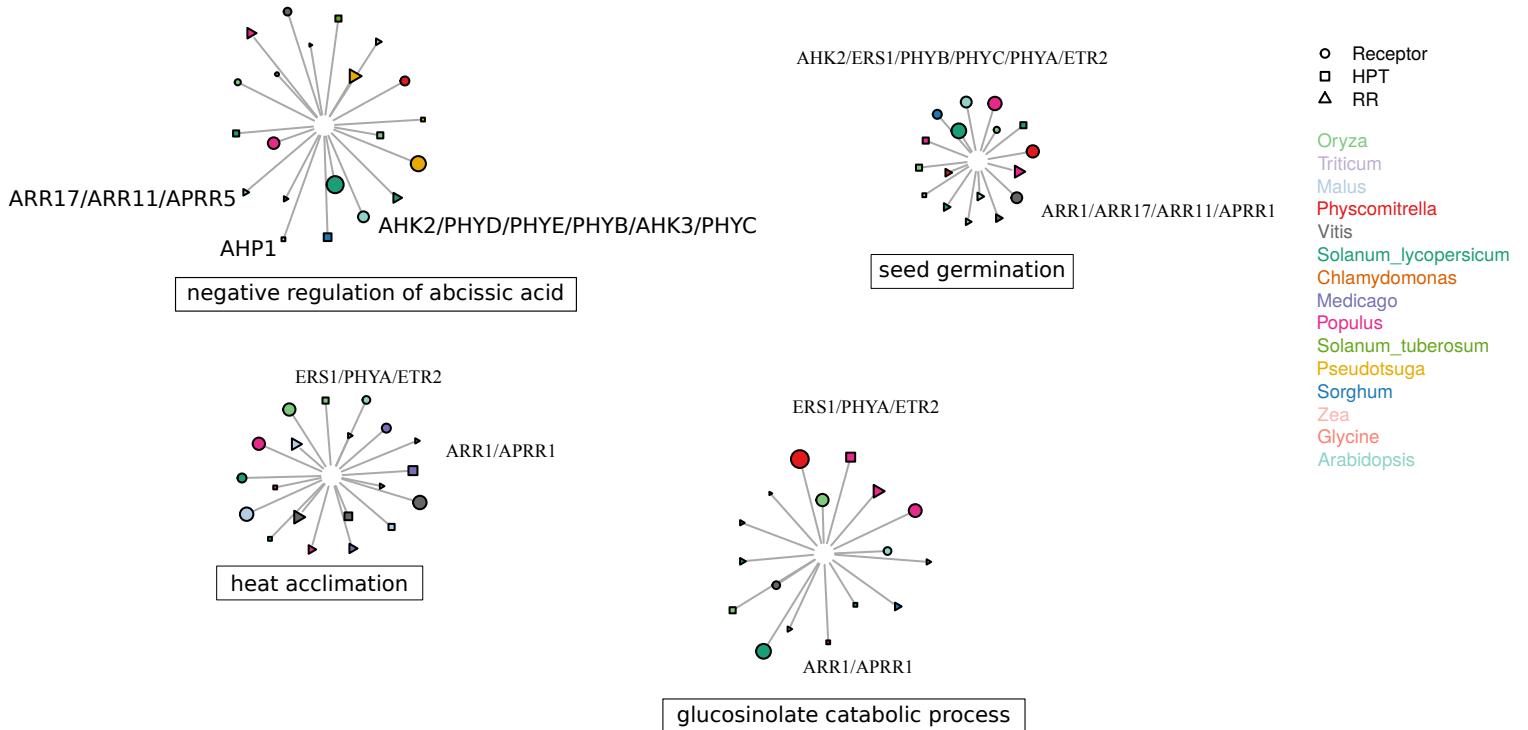


**Table II: Best conserved edges in the *Arabidopsis* PLC network (based on ca. 750 vertices).**

From	To	Nr species co-occurrence	From Name	To Name
AT2G44080.1	AT3G23150.1	9	ARGOS-like protein	ETR2
AT1G65950.1	AT2G25180.1	8	Protein kinase superfamily protein	ARR12
AT3G29350.1	AT5G39340.1	7	AHP2	AHP3
AT2G17820.1	AT3G57040.1	7	AHK1	ARR9
AT1G67195.1	AT1G10470.1	7	NA	ARR4
AT3G63110.1	AT1G10470.1	6	Adenylate isopentenyltransferase 3 Cytokinin biosynthesis	ARR4
AT5G62920.1	AT1G19050.1	5	ARR6	ARR7
AT1G74890.1	AT1G19050.1	5	ARR15	ARR7
AT5G61380.1	AT3G50500.2	5	APRR1-TOC1	Serine/threonine-protein kinase SRK2D
AT2G41310.1	AT3G48100.1	4	ARR8	ARR5
AT5G62920.1	AT3G48100.1	4	ARR6	ARR5
AT5G62920.1	AT3G57040.1	4	ARR6	ARR9
AT1G66340.1	AT2G25180.1	4	ETR1	ARR12
AT2G17820.1	AT1G10470.1	4	AHK1	ARR4
AT4G39770.1	AT1G10470.1	4	Probable trehalose-phosphate phosphatase H	ARR4
AT4G00880.1	AT1G10470.1	4	putative SAUR-type hormone signal effector (AtSAUR31)	ARR4
AT1G22770.1	AT5G24470.1	4	Protein GIGANTEA	APRR5
AT2G40940.1	AT3G23150.1	3	ERS1	ETR2
AT5G62920.1	AT1G10470.1	3	ARR6	ARR4
AT3G48100.1	AT3G57040.1	3	ARR5	ARR9
AT1G66340.1	AT4G27950.1	3	ETR1	Ethylene-responsive transcription factor CRF4
AT3G16857.2	AT3G26100.2	3	ARR1	Regulator of chromosome condensation (RCC1) family protein
AT5G35840.1	AT5G35750.1	3	PHYC	AHK2
AT1G19050.1	AT1G10470.1	3	ARR7	ARR4
AT1G21680.1	AT5G24470.1	3	DPP6 N-terminal domain-like protein	APRR5
AT5G62910.1	AT5G24470.1	3	MQB2.23 RING/U-box superfamily protein	APRR5
AT1G74040.1	AT5G24470.1	3	2-isopropylmalate synthase 2, chloroplastic IPMS2	APRR5
AT5G07100.1	AT5G24470.1	3	Probable WRKY transcription factor 26 WRKY26 cellular heat acclimation	APRR5

**Table III : Community conservation in *Arabidopsis* inter-species co-occurrence network.**

Guide gene community in initial network	Conservation in co-occurrence network	Partially conserved associations
ETR1/ARR12	Yes	
ARR21/ARR13	Yes	
ERS1/PHYA/ETR2/AHP2/ AHP3/ARR1	Partial	ERS1/PHYA/ETR2 AHP2/AHP3
AHK1/AHK4/AHK5/EIN4/ ARR10/ARR3/ARR9 ARR5/ARR7/ARR6/ARR8/	Partial	AHK1/ARR3/ARR9/ARR5/ ARR7/ARR6/ARR8/ARR4
AHK2/PHYB/PHYC/AHP6/ APRR7/APRR9/APRR2	Partial	AHK2/PHYC
APRR6/ARR22	No	
AHP1/ARR11	Yes	
AHP5/ARR2	Yes	



**Figure 7: Putative relationships between Gene Ontology terms and MSP related genes.** A selection of GO terms linked through marker genes to MSP-related guide genes for the different species. Representation size of the different MSP steps in the pathway (receptor, HPT, RR) is proportional to the number of genes found associated to a given GO term (see Supplemental data). As used as a reference in this study, Arabidopsis gene names are indicated for each functional cluster when relevant.

## 2.3 Inter-species network comparison for expressolog annotation

Attributing functions to genes in non model species remains a tricky and hazardous process. Genes produced by duplication events and sharing high identity can be either functionally redundant or play different roles. Thus comparison of genes or proteins according to their sequence similarity comparison using well annotated model species may give a glimpse of a gene's function and its corresponding protein activity but may be not sufficient to infer a concrete physiological function.

A further step to highlight orthologous genes which share common functions is the identification of expressologs i.e genes with high sequence similarity but which also display similar expression profiles (Patel et al., 2012). Here, we experienced such approach based on the PLC networks. To this purpose, marker genes connected to guide genes in the networks were replaced by their associated Gene Ontology (GO) terms to construct the same network. Using such approach was useful to overcome orthology and complex relationships such as many-to-many by taking a gene as a generic functional process display. However, as genes are generally annotated with several GO terms, we manually selected a variety of biologically informative GO terms represented in several species. The associations among MSP-related genes and the retrieved functions were linked to various physiological processes such as crosstalk with other phytohormone signaling pathways, development, adaptation to environment and finally primary and specialized metabolism (**Figure 7**).

As the MSP pathway is involved in complex cross-talk with other signaling pathways (Grefen & Harter, 2004), well represented GO terms linked to other phytohormones were represented. The GO term “negative regulation of abscisic acid–activated signaling pathway” (GO:0009788) was related to cytokinin receptors (AHK2 and AHK3) and phytochromes B, D and E, AHP1 as well as ARR11, 17 and APRR5 in *Arabidopsis*. Receptor and response regulator orthologs associations linked to this term were also found for *Populus*, *Pseudotsuga* and *S. lycopersicum*. Interestingly AHK3, PHYD and E, ARR11 were also linked to “basipetal auxin transport” (GO:0010540) and



APPR5 to “auxin polar transport” (GO:0009934). Bearing in mind the large panel of developmental and growth processes involving MSP signaling through cytokinin, ethylene and light, specific actors of this pathway must preferentially act together in a given process. In Arabidopsis, marker genes associated to seed germination for example were linked to the receptors AHK2, ERS1,ETR2 PHYB, PHYC and PHYA, AHP1 and ARR1,ARR11, ARR17 and APRR1 as well as MSP actors of different levels in other plants as *Solanum lycopersicum* or *Vitis*. Concerning response to environment, PHYA, ERS1 and ETR2 associated to ARR1 and APRR1 are linked to “heat acclimation”. Both phytochromes and ethylene are involved in heat stress response (Larkindale, 2002; Jung et al., 2016). MSP genes from *Populus*, *Medicago*, *Vitis*, *Malus* and *Oryza* has been linked to the same process. Concerning the cooperation between MSP pathways and metabolism, an interesting association was found for “glucosinolate catabolic process” (GO:0019762) and PHYA, ERS1, ETR2, ARR and APPR1 in Arabidopsis. Ethylene has been described by the past to regulate glucosinolate content in this plant (Mewis et al., 2006). *Oryza*, *Physcomitrella*, *S. lycopersicum*, *Populus* and *Sorghum* also display associations between MSP actors and this GO term.

## Discussion

Comparative genomics have been carried out on a large variety of species (Stuart et al., 2003; Bergmann, Ihmels & Barkai, 2004; Lu, Huggins & Bar-Joseph, 2009). Recently, Saul et al. highlighted conserved co-expressed gene modules in mice, sticklebacks, and honey bees in such a complex process as social challenge response (Saul et al., 2017). We therefore compared MSP co-expression networks between several species from the green lineage.

Such co-expression networks are expected to provide additional information above sequence similarity since genes despite showing highly comparable sequences and stemming from the same ancestor gene may have acquired new functions or evolved to pseudo-genes by losing their function (Wagner,



2008). A presumptive bias could be the heterogeneity in term of sample size and sample conditions among the datasets. We used here the largest expression matrices possible for each species since a large panel of different experimental samples is expected to cover a wide broad of physiological conditions and therefore give an averaged overview of transcript relationships. Moreover, we have shown in our previous work that sample size of *ca.* 200 samples is sufficient to construct robust co-expression networks. As shown previously in *Arabidopsis*, PLC networks based on a set of cytokinin signaling pathway related guide genes highlighted putative preferential paths which can be associated to given biological functions by analyzing the marker genes linked to guide gene associations appearing in the network (Liesecke et al., 2018). The prerequisite to construct such targeted networks is to define a set of guide genes used as baits to determine shared co-expressed genes among this latter. Thus, using OrthoDB and Pfam domains, we identified MSP related orthologs in each species. In order ease gene annotation according to experimental validations, the model plant *Arabidopsis* was used as a reference for the other species. This method rapidly showed the limits of comparison based on sequence similarity only. Indeed, the approach did not segregate the different ARR-A from *Arabidopsis* which were found in a same cluster of orthologs. To bring out putative preferential associations among the different levels of the transduction pathway, we constructed PLC networks containing a similar number of vertices for each species (**Figure 5**). These networks showed interesting MSP guide gene combinations reflecting the complex interactions which exist within this pathway. At least partial redundancy has been largely described by mutant analysis in *Arabidopsis* for cytokinin receptors (Nishimura, 2004), phytochromes (Franklin, 2003), histidine phosphotransferases (Hutchison et al., 2006) and for ARR-A, ARR-B as well as for PRR (Salomé et al., 2005; Ishida et al., 2008; To & Kieber, 2008). Hence it is not surprising to find several genes of a pathway level in a same community. The same can be pointed out for communities displaying as well ARR from the A-type and from the B-type as transcription of ARR-A is regulated to a certain extend by ARR-B (Hwang & Sheen, 2001; Sakai et al., 2001). The fact that cytokinin and ethylene receptors as well as phytochromes are found in



common communities reflects the complex signaling cross-talk occurring among these phytohormone transduction pathways and between this latter and light (Mira-Rodado et al., 2007; Zdarska et al., 2015). To fulfill our purpose, the most interesting communities were those displaying at least receptor and response regulator genes. But even if mono-level communities do not permit to map specific sub-pathways, they can nevertheless be useful to deduce potential biological responses involving guide genes by GO enrichment tests performed on the communities' marker genes and facilitate their functional annotation. Comparison among communities remained difficult because of the large number of related sequences in receptors and response regulators among species. To overcome complex orthology relationships between marker genes, these latter were replaced by functional GO terms to establish putative links between biological processes and specific MSP-related genes in the different species (**Figure 7**). Actors from different levels of the pathway linked to a common GO term may be considered as potential interplaying in the associated process.

The stake of this cross-species procedure was two-pronged. Besides highlighting preferential associations in the MSP pathway and their potential evolutionary conservation, co-expression networks from multiple species can be used to infer weight to specific associations by edge co-occurrence and thus to improve network quality. The use of model organisms relies on the fact that orthologous gene function is expected to be mainly conserved through evolution. Gene associations found in several species' co-expression networks are more likely to be biologically relevant whereas true negative associations are plausible to not appear over different independent datasets (Hansen et al., 2014). As among all species used in this study, the model plant *Arabidopsis thaliana* is by far the most well described, we chosed to re-construct its PLC network with only edges co-occurring in at least another species (**Figure 6A**). As a consequence of the vertex number reduction, network topology and GO representation are likely to change leading to a reorganization of communities and their associated functions bringing the most relevant ones out. Therefore retaining only the strongest interactions between guide and marker genes



(*i.e.* those represented in a maximum of co-expression networks of different species) permits to prioritize gene selection for further functional validation.

The general purpose of this study was to give insights in the intrinsic organization of the MSP transduction pathway by using RNAseq data from 15 different plant species. As expression changes has been detected in genes coding for actors of the MSP pathway under stress conditions for example (Urao et al., 1998; Choi et al., 2011; Singh et al., 2015), building targeted co-expression network focusing on MSP-related genes appears as an interesting approach to get an overview of putative preferential associations occurring in response to a specific stimulus and to get a better idea of gene function in non model species rather than by sequence similarity comparison alone. The high amount of putative associations highlighted in the different species allows an insight in the intricacy of Multi Step Phosprelay signaling. It perfectly reflects the complexity of signal transduction in plants as sessile organisms have to adapt quickly to a large amount of environmental changes to ensure their survival. Functional validation of candidate genes and elucidating which molecular mechanisms other than transcriptional regulation, including phytohormone ratios for example, contribute to the large panel of responses implying this transduction pathway will lead to knowledge enhancement about the MSP transduction pathway in the future.

## **Material and methods**

All analyses using R language were performed with R v3.4.4 (R Core Team, 2014).

### **1. RNA-seq data preparation**

For each of the fifteen species, available RNA-seq accessions were retrieved from ArrayExpress. Fastq files were obtained from the SRA after converting .sra files with the SRA ToolKit function ‘fastq-dump’ with the -split-files option for paired-end sequencing runs. Reads were systematically trimmed with



Trimmomatic v0.38 using adapter files according to the Illumina platform used for the run (Bolger, Lohse & Usadel, 2014). Trimmed reads were pseudo-aligned to predicted transcripts from the representative gene models of latest genomes with Salmon v0.7.2 using the variational Bayesian EM algorithm mode to improve the estimation of abundances (Patro et al., 2017). Only samples displaying a mapping rate of reads  $>50\%$  were kept, resulting in a final expression matrices of variable dimensions *i.e.* different number of genes and experimental conditions for each species (see **Figure 3** for details).

## 2. Identification of MSP-related orthologs

### 2.1 Ortholog clustering

Protein sequences corresponding to the CDS were retrieved with NCBI Blast+ (2.2.29 (Camacho et al., 2009)). This latter, were annotated with associated Pfam domains (Finn et al., 2016) with HMMER (v3.1, (Mistry et al., 2013)) which uses profil hidden Markov models on Pfam database.

To identify gene orthologs involved in the MSP for each species, we used OrthoDB 2.4.4 with standard parameters (BRHCLUS 2.2.1 and ORTHOPIPE 6.2.4) (Waterhouse et al., 2013) to cluster protein sequences (versions annexe). As *Arabidopsis thaliana* is the plant species where MSP pathways are best described, we considered the *Arabidopsis* MSP genes (annexe) as reference: genes from other species found in the same cluster than known *Arabidopsis* MSP genes were kept and checked for the presence of MSP related Pfam domains.

### 2.2 Phylogenetic trees

Phylogenetic trees were generated with R. Protein sequences containing PFAM domains related to the MSP pathway were retrieved in FASTA format for each species. These latter were aligned using the multiple alignment tool Cobalt v2.0.2 (Papadopoulos & Agarwala, 2007) using the fast minimum evolution algorithm (Desper & Gascuel, 2002). Pairwise distances were



computed with Phangorn R package v2.2.0 (Schliep, 2011) using the JC69 model (Jukes & Cantor, 1969).

### **3. Gene annotation**

For each species, complete genome annotations were retrieved from Uniprot database (<http://www.uniprot.org/>) (Bateman et al., 2017).

### **4. Co-expression network construction**

#### **4.1 Distance calculation**

We used the Highest Reciprocal Rank (HRR) (Mutwil et al., 2011) as a distance measure to establish co-expression between the candidate genes and the remaining genome.

To establish the HRR between two genes of interest: gene A and gene B, Pearson Correlation Coefficients (PCC) are computed for each gene and the remaining genes of the genome. Then an increasing rank is attributed to each gene following the decreasing PCC values. The HRR between gene A and gene B is calculated as follow:

$$\text{HRR}(\text{GeneA}, \text{GeneB}) = \max(\text{rank}(\text{GeneA} \rightarrow \text{GeneB}), \text{rank}(\text{GeneB} \rightarrow \text{GeneA}))$$

To compute HRR for each species we used the respective RNAseq expression tables as input with our distance calculation program available under <https://github.com/EA2106-Universite-Francois-Rabelais/Expression-network-analysis>.

#### **4.2 Co-expression network construction**

The R package igraph v1.0.1 was used to build targeted networks (Csárdi & Nepusz, 2006). They were constructed as unweighted and undirected networks from two columns matrices with the first column containing guide



Based on PLC networks containing ca. 1250 vertices, for each species, GO terms represented more than 5 and less than 50 times in the genome were retained. Marker genes were replaced by associated GO terms generating a “From-To table” containing guide genes and GO terms associated to connected marker. GO terms represented at least 15 times and at the most 40 times over the 15 species and further represented in at least seven different species were retrieved. Biologically relevant GO terms linked to development, response to environment, crosstalk with other phytohormones or primary and specialized metabolism and the associated guide genes (*i.e.* found in the same community as marker genes associated to a given GO term) were linked to this latter.

## Figures

**Figure 1: Schematic representation of the Multi Step Phosphorelay (MSP) pathway.** The receptor containing a Histidine Kinase (HK) domain binds/senses the stimulus and subsequently auto-phosphorylates. This leads to a phosphorylation cascade via a Histidine-Phosphotransferase (HPT) up to a Response Regulator (RR) which interacts with DNA and elicits transcription of target genes.

**Figure 2: Distribution of MSP related genes.** For each species, potential guide genes for Pathway Level Co-expression network construction were identified based on their PFAM domains. The different species show an heterogeneous amount of MSP-related genes though display a constant distribution at the different pathway levels (amount of RR > amount of HK > amount of HPT).

**Figure 3: RNAseq datasets characteristics.** For each species, available RNAseq data were retrieved and pre-processed to establish expression matrices (see Material and Methods). Here the distribution of the number of reads, the mapping rate as well as the dimensions of the resulting expression matrix are displayed for each one.



**Figure 4: Pathway Level Correlation networks from the 15 species.** PLC networks were build to contain ca. 750 vertices. Receptors are displayed in red, Histidine Phosphotransferases in blue and Response Regulators in green. Co-expressed genes shared among guide genes (marker genes) are drawn in orange. Communities are surrounded in orange as identified by the clustering algorithm. Genes are annotated with their respective locus tag and their Arabidopsis ortholog's names as identified by OrthoDB.

**Figure 5: Amount of conserved network edges among species.** Conserved edges between species as identified by pair-wise Pinalog comparison on MSP PLC networks of the different species containing approximatively 1250 vertices.

**Figure 6: Arabidopsis PLC networks including cross-species edge conservation information.** Pinalog based on initial networks containing ca. 750 vertices. Receptors are displayed in red, Histidine Phosphotransferases in blue, Response Regulators in green and marker genes are drawn in orange. Communities are surrounded in orange as identified by the clustering algorithm. Edge thickness is proportional to inter-species co-occurrence. A. Initial Arabidopsis PLC network with co-occurrence weighted edges B. Conserved Arabidopsis PLC network: Based on the initial network but only vertices implied in edges found in Arabidopsis and at least one other species were retained for network construction.

**Figure 7: Putative relationships between Gene Ontology terms and MSP related genes.** A selection of GO terms linked through marker genes to MSP-related guide genes for the different species. Representation size of the different MSP steps in the pathway (receptor, HPT, RR) is proportional to the number of genes found associated to a given GO term (see Supplemental data). As used as a reference in this study, Arabidopsis gene names are indicated for each functional cluster when relevant.



## **Supplemental data**

**Supplemental data 1: Relationship between identified MSP-related genes based Pfam domains and genome size for each species.**

**Supplemental data 2: List of MSP-related genes in each species.**

Following genes are used as guide genes for PLC network construction in this study. Genes were selected on two criteria: presence of Pfam domains related to MSP pathway and clustering with an Arabidopsis guide gene using OrthoDB.

**Supplemental data 3: Phylogenetic trees constructed from identified MSP receptors (A), transmitters (B) and response regulators (C) for each species.** Gene names are annotated with their corresponding Arabidopsis gene according to OrthoDB clustering.

**Supplemental data 4: PLC network characteristics for different thresholded networks.** Network transitiviy is defined as the relative number of triangles in the graph compared to the total number of connected triples of nodes. Network degree of a vertex is the number of edges incident to the vertex.

**Supplemental 5: Network community composition.** (file: Species\_nettwok\_comm\_name) Communities for each species' PLC network as identified by the greedy optimization of modularity algorithm.

## **References**

Bartrina I., Otto E., Strnad M., Werner T., Schmülling T. 2011. Cytokinin Regulates the Activity of Reproductive Meristems, Flower Organ Size, Ovule Formation, and Thus Seed Yield in



\textless i\textgreater Arabidopsis thaliana\textless /i\textgreater. *The Plant Cell* 23:69–80. DOI: 10.1105/tpc.110.079079.

Bateman A., Martin MJ., O'Donovan C., Magrane M., Alpi E., Antunes R., Bely B., Bingley M., Bonilla C., Britto R., Bursteinas B., Bye-AJee H., Cowley A., Da Silva A., De Giorgi M., Dogan T., Fazzini F., Castro LG., Figueira L., Garmiri P., Georghiou G., Gonzalez D., Hatton-Ellis E., Li W., Liu W., Lopez R., Luo J., Lussi Y., MacDougall A., Nightingale A., Palka B., Pichler K., Poggioli D., Pundir S., Pureza L., Qi G., Rosanoff S., Saidi R., Sawford T., Shypitsyna A., Speretta E., Turner E., Tyagi N., Volynkin V., Wardell T., Warner K., Watkins X., Zaru R., Zellner H., Xenarios I., Bougueleret L., Bridge A., Poux S., Redaschi N., Aimo L., ArgoudPuy G., Auchincloss A., Axelsen K., Bansal P., Baratin D., Blatter MC., Boeckmann B., Bolleman J., Boutet E., Breuza L., Casal-Casas C., De Castro E., Coudert E., Cuche B., Doche M., Dornevil D., Duvaud S., Estreicher A., Famiglietti L., Feuermaier M., Gasteiger E., Gehant S., Gerritsen V., Gos A., Gruaz-Gumowski N., Hinz U., Hulo C., Jungo F., Keller G., Lara V., Lemercier P., Lieberherr D., Lombardot T., Martin X., Masson P., Morgat A., Neto T., Nouspikel N., Paesano S., Pedruzzi I., Pilbaut S., Pozzato M., Pruess M., Rivoire C., Roechert B., Schneider M., Sigrist C., Sonesson K., Staehli S., Stutz A., Sundaram S., Tognolli M., Verbregue L., Veuthey AL., Wu CH., Arighi CN., Arminski L., Chen C., Chen Y., Garavelli JS., Huang H., Laiho K., McGarvey P., Natale DA., Ross K., Vinayaka CR., Wang Q., Wang Y., Yeh LS., Zhang J. 2017. UniProt: The universal protein knowledgebase. *Nucleic Acids Research* 45:D158-D169. DOI: 10.1093/nar/gkw1099.

Bergmann S., Ihmels J., Barkai N. 2004. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biology* 2. DOI: 10.1371/journal.pbio.0020009.



- Binder BM., Kim HJ., Mathews DE., Hutchison CE., Kieber JJ., Schaller GE. 2018. A role for two-component signaling elements in the Arabidopsis growth recovery response to ethylene. *Plant Direct* 2:e00058. DOI: 10.1002/pld3.58.
- Bolger AM., Lohse M., Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.
- Brenner WG., Romanov GA., Köllmer I., Bürkle L., Schmülling T. 2005. Immediate-early and delayed cytokinin response genes of *Arabidopsis thaliana* identified by genome-wide expression profiling reveal novel cytokinin-sensitive processes and suggest cytokinin action through transcriptional cascades. *Plant Journal* 44:314–333. DOI: 10.1111/j.1365-313X.2005.02530.x.
- Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009. BLAST+: architecture and applications. *BMC bioinformatics* 10:421. DOI: 10.1186/1471-2105-10-421.
- Cho Y-H., Yoo S-D. 2006. ETHYLENE RESPONSE 1 Histidine Kinase Activity of *Arabidopsis* Promotes Plant Growth. *Plant Physiology* 143:612–616. DOI: 10.1104/pp.106.091504.
- Choi J., Choi D., Lee S., Ryu C-M., Hwang I. 2011. Cytokinins and plant immunity: old foes or new friends? *Trends in plant science* 16:388–94. DOI: 10.1016/j.tplants.2011.03.003.
- Clauset A., Newman MEJ., Moore C. 2004. Finding community structure in very large networks. 066111:1–6. DOI: 10.1103/PhysRevE.70.066111.
- Csárdi G., Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Systems* 1695:1695.
- Desper R., Gascuel O. 2002. Fast and accurate phylogeny minimum-evolution principle. *J Comput Biol* 9:687–705. DOI: 10.1089/106652702761034136.



- Etheridge N., Hall BP., Schaller GE. 2006. Progress report: Ethylene signaling and responses. *Planta* 223:387-391. DOI: 10.1007/s00425-005-0163-2.
- Ficklin SP., Feltus FA. 2011. Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice. *Plant Physiology* 156:1244-1256. DOI: 10.1104/pp.111.173047.
- Finn RD., Coggill P., Eberhardt RY., Eddy SR., Mistry J., Mitchell AL., Potter SC., Punta M., Qureshi M., Sangrador-Vegas A., Salazar GA., Tate J., Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 44:D279-D285. DOI: 10.1093/nar/gkv1344.
- Franklin KA. 2003. Phytochromes B, D, and E Act Redundantly to Control Multiple Physiological Responses in Arabidopsis. *Plant Physiology* 131:1340-1346. DOI: 10.1104/pp.102.015487.
- Grefen C., Harter K. 2004. Plant two-component systems: Principles, functions, complexity and cross talk. *Planta* 219:733-742. DOI: 10.1007/s00425-004-1316-4.
- Hansen BO., Vaid N., Musialak-Lange M., Janowski M., Mutwil M. 2014. Elucidating gene function and function evolution through comparison of co-expression networks of plants. *Frontiers in Plant Science* 5. DOI: 10.3389/fpls.2014.00394.
- Hutchison CE., Li J., Argueso C., Gonzalez M., Lee E., Lewis MW., Maxwell BB., Perdue TD., Schaller GE., Alonso JM., Ecker JR., Kieber JJ. 2006. The Arabidopsis Histidine Phosphotransfer Proteins Are Redundant Positive Regulators of Cytokinin Signaling. *the Plant Cell Online* 18:3073-3087. DOI: 10.1105/tpc.106.045674.
- Hwang I., Sheen J. 2001. Two-component circuitry in Arabidopsis cytokinin signal transduction. *Nature* 413:383-389. DOI: 10.1038/35096500.



- Hwang I., Sheen J., Müller B. 2012. Cytokinin Signaling Networks. *Annual Review of Plant Biology* 63:353-380. DOI: 10.1146/annurev-arplant-042811-105503.
- Ishida K., Yamashino T., Yokoyama A., Mizuno T. 2008. Three type-B response regulators, ARR1, ARR10 and ARR12, play essential but redundant roles in cytokinin signal transduction throughout the life cycle of *Arabidopsis thaliana*. *Plant and Cell Physiology* 49:47-57. DOI: 10.1093/pcp/pcm165.
- Jukes TH., Cantor CR. 1969. CHAPTER 24 - Evolution of Protein Molecules. In: Munro HN ed. *Mammalian Protein Metabolism*. Academic Press, 21-132. DOI: 10.1016/B978-1-4832-3211-9.50009-7.
- Jung JH., Domijan M., Klose C., Biswas S., Ezer D., Gao M., Khattak AK., Box MS., Charoensawan V., Cortijo S., Kumar M., Grant A., Locke JCW., Schäfer E., Jaeger KE., Wigge PA. 2016. Phytochromes function as thermosensors in *Arabidopsis*. *Science* 354:886-889. DOI: 10.1126/science.aaf6005.
- Kiba T., Aoki K., Sakakibara H., Mizuno T. 2004. *Arabidopsis* response regulator, ARR22, ectopic expression of which results in phenotypes similar to the wol cytokinin-receptor mutant. *Plant & cell physiology* 45:1063-77. DOI: 10.1093/pcp/pch128.
- Kieber JJ., Schaller GE. 2014. Cytokinins. *The Arabidopsis book / American Society of Plant Biologists* 12:e0168. DOI: 10.1199/tab.0168.
- Larkindale J. 2002. Protection against Heat Stress-Induced Oxidative Damage in *Arabidopsis* Involves Calcium, Abscisic Acid, Ethylene, and Salicylic Acid. *Plant Physiology* 128:682-695. DOI: 10.1104/pp.128.2.682.
- Li S., Fu Q., Chen L., Huang W., Yu D. 2011. *Arabidopsis thaliana* WRKY25, WRKY26, and WRKY33 coordinate induction of plant thermotolerance. *Planta* 233:1237-1252. DOI: 10.1007/s00425-011-1375-2.
- Liesecke F., Daudu D., Dugé de Bernonville R., Besseau S., Clastre M., Courdavault V., de Craene J-O., Crèche J., Giglioli-Guivarc'h N., Glévarec



- G., Pichon O., Dugé de Bernonville T. 2018. Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Scientific Reports* 8:10885. DOI: 10.1038/s41598-018-29077-3.
- Lu Y., Huggins P., Bar-Joseph Z. 2009. Cross species analysis of microarray expression data. *Bioinformatics* 25:1476–1483. DOI: 10.1093/bioinformatics/btp247.
- Medeiros DB., Daloso DM., Fernie AR., Nikoloski Z., Araújo WL. 2015. Utilizing systems biology to unravel stomatal function and the hierarchies underpinning its control. *Plant, Cell and Environment* 38:1457–1470. DOI: 10.1111/pce.12517.
- Mewis I., Tokuhisa JG., Schultz JC., Appel HM., Ulrichs C., Gershenzon J. 2006. Gene expression and glucosinolate accumulation in *Arabidopsis thaliana* in response to generalist and specialist herbivores of different feeding guilds and the role of defense signaling pathways. *Phytochemistry* 67:2450–2462. DOI: 10.1016/j.phytochem.2006.09.004.
- Mira-Rodado V., Sweere U., Grefen C., Kunkel T., Fejes E., Nagy F., Schäfer E., Harter K. 2007. Functional cross-talk between two-component and phytochrome B signal transduction in *Arabidopsis*. *Journal of Experimental Botany* 58:2595–2607. DOI: 10.1093/jxb/erm087.
- Mira-Rodado V., Veerabagu M., Witthöft J., Teply J., Harter K., Desikan R. 2012. Identification of two-component system elements downstream of AHK5 in the stomatal closure response of *Arabidopsis thaliana*. *Plant signaling & behavior* 7:1467–76. DOI: 10.4161/psb.21898.
- Mistry J., Finn RD., Eddy SR., Bateman A., Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research* 41:e121–e121. DOI: 10.1093/nar/gkt263.



- Muller B., Sheen J. 2007. Advances in Cytokinin Signaling. *Science* 318:68-69. DOI: 10.1126/science.1145461.
- Mutwil M., Klie S., Tohge T., Giorgi FM., Wilkins O., Campbell MM., Fernie AR., Usadel B., Nikoloski Z., Persson S. 2011. PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species. *The Plant Cell* 23:895-910. DOI: 10.1105/tpc.111.083667.
- Nakamichi N., Kita M., Ito S., Yamashino T., Mizuno T. 2005. PSEUDO-RESPONSE REGULATORS, PRR9, PRR7 and PRR5, Together play essential roles close to the circadian clock of *Arabidopsis thaliana*. *Plant and Cell Physiology* 46:686-698. DOI: 10.1093/pcp/pci086.
- Nishimura C. 2004. Histidine Kinase Homologs That Act as Cytokinin Receptors Possess Overlapping Functions in the Regulation of Shoot and Root Growth in *Arabidopsis*. *the Plant Cell Online* 16:1365-1377. DOI: 10.1105/tpc.021477.
- Papadopoulos JS., Agarwala R. 2007. COBALT: Constraint-based alignment tool for multiple protein sequences. *Bioinformatics* 23:1073-1079. DOI: 10.1093/bioinformatics/btm076.
- Park DH., Somers DE., Kim YS., Choy YH., Lim HK., Soh MS., Kim HJ., Kay SA., Nam HG. 1999. Control of circadian rhythms and photoperiodic flowering by the *Arabidopsis* GIGANTEA gene. *Science* 285:1579-1582. DOI: 10.2307/2898120.
- Patel RV., Nahal HK., Breit R., Provart NJ. 2012. BAR expressolog identification: Expression profile similarity ranking of homologous genes in plant species. *Plant Journal* 71:1038-1050. DOI: 10.1111/j.1365-313X.2012.05055.x.
- Patro R., Duggal G., Love MI., Irizarry RA., Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* 14:417-419. DOI: 10.1038/nmeth.4197.



- Phan HTT., Sternberg MJE. 2012. PINALOG: A novel approach to align protein interaction networks-implications for complex detection and function prediction. *Bioinformatics* 28:1239-1245. DOI: 10.1093/bioinformatics/bts119.
- Pokhilko A., Mas P., Millar AJ. 2013. Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs. *BMC Systems Biology* 7:1-12. DOI: 10.1186/1752-0509-7-23.
- R Core Team 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rai MI., Wang X., Thibault DM., Kim HJ., Bombyk MM., Binder BM., Shakeel SN., Schaller GE. 2015. The ARGOS gene family functions in a negative feedback loop to desensitize plants to ethylene. *BMC Plant Biology* 15:1-14. DOI: 10.1186/s12870-015-0554-x.
- Raines T., Shanks C., Cheng CY., McPherson D., Argueso CT., Kim HJ., Franco-Zorrilla JM., López-Vidriero I., Solano R., Vaňková R., Schaller GE., Kieber JJ. 2016. The cytokinin response factors modulate root and shoot growth and promote leaf senescence in Arabidopsis. *Plant Journal* 85:134-147. DOI: 10.1111/tpj.13097.
- Riefler M. 2006. Arabidopsis Cytokinin Receptor Mutants Reveal Functions in Shoot Growth, Leaf Senescence, Seed Size, Germination, Root Development, and Cytokinin Metabolism. *the Plant Cell Online* 18:40-54. DOI: 10.1105/tpc.105.037796.
- Ruprecht C., Mendrinna A., Tohge T., Sampathkumar A., Klie S., Fernie AR., Nikoloski Z., Persson S., Mutwil M. 2016. FamNet: A framework to identify multiplied modules driving pathway diversification in plants. *Plant Physiology* 170:pp.01281.2015. DOI: 10.1104/pp.15.01281.
- Sakai H., Honma T., Aoyama T., Sato S., Kato T., Tabata S., Oka A. 2001. ARR1, a Transcription Factor for Genes Immediately responsive to Cytokinins. *Science* 294:1519-1521. DOI: 10.1126/science.1065201.



- Salomé PA., McClung CR., Salome PA., McClung CR. 2005. PSEUDO-RESPONSE REGULATOR 7 and 9 are partially redundant genes essential for the temperature responsiveness of the *Arabidopsis* circadian clock. *Plant Cell* 17:791–803. DOI: 10.1105/tpc.104.029504.In.
- Saul MC., Blatti C., Yang W., Bukhari SA., Shpigler HY., Troy JM., Seward CH., Sloofman LG., Chandrasekaran S., Bell AM., Stubbs LJ., Robinson GE., Zhao SD., Sinha S. 2017. Cross-species systems analyses reveal a conserved brain transcriptional response to social challenge. *bioRxiv*. DOI: 10.1101/219444.
- Schaller GE., Kieber JJ., Shiu S-H. 2008. Two-Component Signaling Elements and Histidyl-Aspartyl Phosphorelays <sup>†</sup>. *The Arabidopsis Book* 6:e0112. DOI: 10.1199/tab.0112.
- Schliep KP. 2011. phangorn: Phylogenetic analysis in R. *Bioinformatics* 27:592–593. DOI: 10.1093/bioinformatics/btq706.
- Singh A., Kushwaha HR., Soni P., Gupta H., Singla-Pareek SL., Pareek A. 2015. Tissue specific and abiotic stress regulated transcription of histidine kinases in plants is also influenced by diurnal rhythm. *Frontiers in Plant Science* 6:1–14. DOI: 10.3389/fpls.2015.00711.
- Song J., Liu Q., Hu B., Wu W. 2016. Comparative transcriptome profiling of *Arabidopsis* Col-0 in responses to heat stress under different light conditions. *Plant Growth Regulation* 79:209–218. DOI: 10.1007/s10725-015-0126-y.
- Stock AM., Robinson VL., Goudreau PN. 2000. Two-Component Signal Transduction. *Reactions* 69:183–215. DOI: 10.1146/annurev.biochem.69.1.183.
- Stuart JM., Segal E., Koller D., Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302:249–255. DOI: 10.1126/science.1087447.



Sun L., Zhang Q., Wu J., Zhang L., Jiao X., Zhang S., Zhang Z., Sun D., Lu T., Sun Y. 2014. Two Rice Authentic Histidine Phosphotransfer Proteins, OsAHP1 and OsAHP2, Mediate Cytokinin Signaling and Stress Responses in Rice. *Plant Physiology* 165:335–345. DOI: 10.1104/pp.113.232629.

Sweere U., Eichenberg K., Lohrmann J., Mira-Rodado V., Bäurle I., Kudla J., Nagy F., Schäfer E., Harter K. 2001. Interaction of the response regulator ARR4 with phytochrome B in modulating red light signaling. *Science* 294:1108–1111. DOI: 10.1126/science.1065022.

Tirosh I., Bilu Y., Barkai N. 2007. Comparative biology: beyond sequence analysis. *Current Opinion in Biotechnology* 18:371–377. DOI: 10.1016/j.copbio.2007.07.003.

To JPC., Kieber JJ. 2008. Cytokinin signaling: two-components and more. *Trends in Plant Science* 13:85–92. DOI: 10.1016/j.tplants.2007.11.005.

Urao T., Miyata S., Yamaguchi-Shinozaki K., Shinozaki K. 2000. Possible His to Asp phosphorelay signaling in anArabidopsis two-component system. *FEBS Letters* 478:227–232. DOI: 10.1016/s0014-5793(00)01860-3.

Urao T., Yakubov B., Yamaguchi-Shinozaki K., Shinozaki K. 1998. Stress-responsive expression of genes for two-component response regulator-like proteins in Arabidopsis thaliana. *FEBS Letters* 427:175–178. DOI: 10.1016/S0014-5793(98)00418-9.

Wagner A. 2008. Gene duplications, robustness and evolutionary innovations. *BioEssays* 30:367–373. DOI: 10.1002/bies.20728.

Waterhouse RM., Tegenfeldt F., Li J., Zdobnov EM., Kriventseva EV. 2013. OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research* 41:358–365. DOI: 10.1093/nar/gks1116.

Wei H., Persson S., Mehta T., Srinivasasainagendra V., Chen L., Page GP., Somerville C., Loraine A. 2006. Transcriptional Coordination of the



Metabolic Network in Arabidopsis. *Plant Physiology* 142:762-774. DOI: 10.1104/pp.106.080358.

Wolfe CJ., Kohane IS., Butte AJ. 2005. Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC Bioinformatics* 6:1-10. DOI: 10.1186/1471-2105-6-227.

Zdarska M., Dobisová T., Gelová Z., Pernisová M., Dabrevolski S., Hejátko J. 2015. Illuminating light, cytokinin, and ethylene signalling crosstalk in plant development: Fig. 1. *Journal of Experimental Botany* 66:4913-4931. DOI: 10.1093/jxb/erv261.



## Discussion générale



## **Discussion**

### **La mise en place de réseaux de co-expression, un outil précieux pour l'élucidation de voies cellulaires**

Le nombre d'analyses de processus biologiques par des approches *in silico* de réseaux de co-expression a augmenté de manière drastique ces dernières années (Tian et al., 2017; Wisecaver et al., 2017; Ruprecht et al., 2017; Proost & Mutwil, 2018; Yu et al., 2018). Ces nombreuses études portent soit sur des applications directes vouées à la compréhension du fonctionnement des organismes considérés (Wisecaver et al., 2017; Ruprecht et al., 2017), soit sur des aspects logiciels ou d'optimisation du processus de construction et d'analyse (Tian et al., 2017; Yu et al., 2018). Plusieurs bases de données de co-expression dédiées à la lignée verte ont été mises en place récemment (CoexpNetViz, (Tzfadia et al., 2015); PhytoNet, (Ferrari et al., 2018), CoxPathDB pour la tomate (Narise et al., 2017)) et offrent la possibilité à la communauté scientifique de conduire assez aisément ce type d'analyse. Le développement rapide de ces approches démontre qu'elles constituent un outil précieux pour une meilleure compréhension de processus biologiques et évolutifs. La croissance de ce type de bases de données est alimentée par l'augmentation continue du nombre de données disponibles publiquement. Cependant ces bases de données sont limitées à une ou quelques espèces modèles et n'offrent pas de contrôle sur la manière de calculer la distance entre gènes ni sur les jeux de données sous-jacents.

Motivés par la volonté de mettre en place une méthodologie impliquant la construction de réseaux de co-expression sur des espèces non-modèles étudiées à l'EA2106, telles que le pommier ou la Pervenche de Madagascar, et de maîtriser tout le processus de construction, nous avons initié une étude exhaustive de ces réseaux tout en considérant leur applicabilité à des problématiques biologiques. Parmi celles-ci, l'identification de gènes codant des enzymes impliquées dans les voies de biosynthèse du métabolisme spécialisé reste un enjeu majeur. La mise en évidence d'associations transcriptionnelles spécifiques entre gènes intervenant dans une même voie de signalisation a également été entreprise. Dans ces deux cas, l'objectif est de trouver les gènes les mieux co-exprimés avec ceux connus pour intervenir dans le processus d'intérêt (métabolisme ou signalisation). Par conséquent, l'analyse exhaustive des co-



expressions exige que le réseau soit construit à l'échelle du génome, de manière à ne pas manquer d'associations potentielles.

Construire des réseaux sur des matrices d'expression possédant un grand nombre de gènes ( $>30,000$ ) entraîne des difficultés techniques, liés notamment aux ressources de calcul, quant à la manière d'inférer les relations entre gènes. Parmi les nombreuses manière d'estimer ces relations (López-Kleine, Leal & López, 2013), certaines ne sont justement pas appropriées à de telles tailles de jeux de données. C'est le cas notamment de méthodes supervisées généralement utilisées pour l'inférence de réseaux de régulation. Les méthodes non supervisées sont à l'inverse adaptées au calcul de distances entre un grand nombre de gènes. Parmi ces méthodes, l'utilisation de coefficients de corrélation (CC, Pearson (PCC) ou Spearman (SCC)) est la plus intuitive et la plus économique en terme de ressources informatiques, elle est donc très largement utilisés pour les études de co-expression. De plus, la pertinence de ces coefficients a été démontrée (Ballouz, Verleyen & Gillis, 2015). Nous avons donc choisi le PCC comme distance de base pour la construction des réseaux de co-expression à partir de larges matrices en appliquant une méthode d'ordonnancement supplémentaire de manière à intégrer le comportement des autres gènes. Au travers de notre première évaluation, nous avons montré que des PCC ordonnés réciproquement (PCC-HRR) capturent de manière très efficace des associations connues et sont donc potentiellement capables de générer des réseaux robustes. Il s'agit de la première évaluation globale de ce genre de PCC face à d'autres méthodes classiquement utilisées, telles que l'information mutuelle ou les corrélations partielles, même si Obayashi & Kinoshita avaient déjà démontré une supériorité de rangs mutualisés par rapport à des valeurs absolues de CC (Obayashi & Kinoshita, 2009). Les HRR ont été introduits par Mutwil et al pour la construction de réseaux de co-expression sur *Arabidopsis* (Mutwil et al., 2010). Dans cette étude, le réseau PCC-HRR a été comparé à un réseau de type Gaussian Graphical Model (GGM) sur cette même plante, toutefois sans réelle confrontation de leurs performances.

Les résultats de notre première étude sont importants à plusieurs titres : (i) un classement des CC augmente la performance du réseau global et des PLC, (ii) un programme parallélisable pour calculer rapidement des CC ordonnés à l'échelle du génome a été développé, (iii) des distances plus complexes à mettre en place n'augmentent pas la performance des réseaux, (iv) les méthodes de normalisation de données RNA-seq n'augmentent pas la performance des réseaux, (v) la possibilité de combiner des réseaux issus de données microarrays et RNA-seq



et (vi) le développement d'une feuille de route pour la récupération et le traitement de données transcriptomiques.

De cette première étude découlent également des réseaux de co-expression ciblés sur la voie de signalisation des CK, ainsi que sur les voies métaboliques des carbohydrates, des acides gras, des phenylpropanoides et des terpènes. Les données générées sont disponibles à la communauté scientifique et pourront être ré-utilisées par d'autres équipes.

Certains gènes, non utilisés en appâts mais apparaissant comme éléments clés dans les voies, ont pu être identifiés dans chacune de ces voies. Il apparaît dans la littérature que les optimisations/mises au point de réseaux sont fréquemment réalisées sur les voies des métabolismes primaire et secondaire (Ma, Gong & Bohnert, 2007), Gaussian Graphical Model, Cell Wall Biosynthesis ; (Mutwil et al., 2011b), PCC-HRR + HCCA, Photosynthèse & Flavonoïdes) et plus rarement sur les voies de signalisation. En plus de présenter une coordination transcriptionnelle plus stable, il est possible que les validations fonctionnelles sur ce genre de métabolisme soient plus évidentes à réaliser que sur des voies exhibant beaucoup de redondances fonctionnelles. Concernant la voie CK, une expression simultanée des récepteurs dans une condition peut être possible, mais cette expression ne tient pas compte de leur spécificité de substrat. En effet, il a été montré que le domaine CHASE, la partie senseur du récepteur, ne fixe pas toutes les CK avec les mêmes affinités selon le récepteur (Daudu et al., 2017). Cette notion de spécificité complexifie davantage l'analyse de la voie CK par des analyses transcriptomiques uniquement.

Une fois la distance établissant la co-expression entre gènes fixée (PCC-HRR), nous avons travaillé sur le jeu de données en lui-même. Voulant mettre en place des études de réseaux de co-expression sur des espèces non-modèles, et pour lesquelles la quantité de données transcriptomiques est bien moins importante que pour *A. thaliana* par exemple, nous nous sommes intéressés à la relation entre performance du réseau et nombre d'échantillons dans la matrice d'expression. L'intérêt de cette étude était d'examiner si des réseaux construits sur de petits jeux de données ont tout de même une certaine puissance et à l'inverse, de déterminer si des jeux de données trop volumineux peuvent avoir une performance altérée par effet de dilution de variations transcriptionnelles plus transitoires. L'impact d'un sous-échantillonnage sur la construction de réseaux de co-expression a été étudié à plusieurs reprises mais ces évaluations restent fragmentaires (Reverter & Chan, 2008; Cosgrove, Gardner & Kolaczyk,



2010; Altay, 2012; Gibson et al., 2013; Ballouz, Verleyen & Gillis, 2015). Une information communément mise en évidence par ces travaux est la saturation assez rapide de la performance : en effet, au-delà d'un nombre seuil d'échantillons, la performance n'augmente. Complémentairement à la notion de sous-échantillonnage, la possibilité d'agréger des réseaux semble être une manière adéquate de créer un réseau consensus présentant plus d'associations biologiquement pertinentes qu'un réseau isolé (Lee et al., 2004; Ballouz, Verleyen & Gillis, 2015). L'évaluation de l'impact de la taille de matrice en terme de nombre de conditions expérimentales considérées sur un réseau final s'étend donc aussi (i) aux méthodes de sous-échantillonnage, (ii) aux méthodes d'agrégation et (iii) à la combinaison d'agrégats provenant de données microarrays et RNA-seq. Nous avons mené ces évaluations sur des données microarray et RNA-seq, et dans une optique d'applicabilité étendue, trois espèces ont été étudiées : *A. thaliana*, *S. lycopersicum* et *Z. mays*.

Les résultats importants de cette évaluation sont : (i) une saturation de la performance des réseaux pour des tailles  $>200$  échantillons, (ii) les matrices larges permettent la création de réseaux avec une performance satisfaisante, (iii) créer des réseaux individuels à partir de matrices d'expression provenant d'une partition par k-moyennes des conditions d'une matrice initiale et les agréger selon le nombre d'occurrences des liens offre une performance augmentée, (iv) la possibilité de trouver un réseau consensus résultant de l'agrégation de petits réseaux sur des données RNA-seq et microarray, (v) l'observation d'une corrélation négative, entre corrélations, échantillonnage et performance. Le partitionnement par k-moyennes d'une matrice d'expression large (microarray, *Arabidopsis*) avait déjà montré que les réseaux individuels construits avec cette approche permettaient de capturer une plus grande diversité de processus biologiques (Feltus et al., 2013). Dans cette dernière étude, les performances du jeu de données global ainsi qu'un potentiel agrégat n'avaient pas été comparées.

Nous avons effectué ces comparaisons dans notre deuxième évaluation et réaffirmé la pertinence du partitionnement par k-moyennes. D'un point de vue théorique, le fait qu'un agrégat ait une performance supérieure à un réseau unique résultat d'une matrice combinant un maximum d'échantillons pourrait s'expliquer de la manière suivante : le calcul des corrélations sur un jeu de données unique et large fournit des corrélations intégrant un vaste



panel de conditions expérimentales, et pourrait ainsi être susceptible d'échouer dans la détection d'associations transitoires, masquées par des événements transcriptionnels plus constants. A l'inverse, calculer ces corrélations sur des matrices plus petites et ciblées en terme de processus étudiés, permettrait de mettre en évidence ce type d'associations. Leur combinaison dans un agrégat résulterait en un réseau consensus mettant en évidence ce genre d'associations transitoires. Cependant, nous avons montré qu'au moment de l'agrégation, combiner des liens, selon leur nombre d'occurrence dans les réseaux individuels plutôt que sur leur valeur de HRR, générait des agrégats plus robustes considérant la capture de termes GO. Ceci impliquerait que les associations les plus significatives (HRR bas) ne sont pas forcément les plus représentées (retrouvées dans peu de sous-réseaux, ce qui est lié à la nature des données dans la matrice d'expression) alors que celles étant les plus cohérentes avec des annotations fonctionnelles sont justement les plus fréquentes. A l'avenir, cette méthodologie pourrait permettre des avancées notables la compréhension de voies biologiques. L'exemple développé sur la voie JA montre clairement la puissance de ce type d'analyse, certains facteurs de transcription ayant pu être retrouvés communément co-exprimés entre les trois espèces avec les gènes de la voie de biosynthèse. De plus, cette stratégie pourrait idéalement être adaptée dans les bases de données de co-expression, telles que PlaNet et ATTED-II, qui utilisent des corrélations calculées sur des matrices larges.

Tirant parti des nombreuses données transcriptomiques disponibles pour des espèces modèles et non-modèles, et de la robustesse des réseaux de co-expressions générées par PCC-HRR, nous avons initié une comparaison de réseaux de co-expression entre différentes espèces végétales pour l'étude du fonctionnement des voies de signalisation de type MSP (incluant la voie répondant aux cytokinines (CK)). Ce genre de comparaisons multi-espèces a été initié depuis plusieurs années entre l'Homme et la Souris (Tsaparas et al., 2006) ou l'Homme et le Chimpanzé (Oldham, Horvath & Geschwind, 2006), mais aussi multi-mammifères (COEXPRESdb, (Obayashi et al., 2013)). Des comparaisons entre espèces végétales ont été menées, notamment sur la base de données PlaNet ((Mutwil et al., 2011b), considérant sept espèces différentes) mais aussi dans d'autres études ((Ficklin & Feltus, 2011); Maïs vs Riz). La difficulté majeure de ce type d'analyse est de faire un lien entre relations d'orthologie et de co-expression de type PLC. Deux orthologues directs (meilleurs hits en Double Blast Réciproque) ne présentent pas nécessairement une homologie fonctionnelle. Notre objectif,



dans cette comparaison multi-espèces, est de mettre en évidence des associations transcriptionnelles conservées entre réseaux de co-expressions d'espèces différentes. Une telle conservation pourrait être synonyme de robustesse. Nous avons retenu 15 espèces représentant une large diversité dans la lignée verte en fonction de la quantité de données RNA-seq disponibles et de l'accès à un transcriptome de référence pour l'estimation de l'abondance des transcrits. La base de données Ensembl Plants a été d'une grande aide pour ce dernier point. Des PLC ciblées sur les gènes codant des éléments des voies de signalisation de type MSP ont été créées pour chacun des espèces et confrontées par un algorithme d'alignement de réseaux combinant les relations de co-expression et d'orthologie. De cette manière, nous avons pu mettre en évidence certaines associations très conservées entre l'expression des transcrits de la voie MSP. Comme mentionné plus haut, l'approche PLC semble moins puissante sur cette voie de signalisation en comparaison avec les voies du métabolisme. Identifier des relations conservées entre espèces végétales divergentes renforce d'autant plus la validité de l'approche.

Concernant la voie CK, il apparaît qu'elle fonctionne de manière modulaire au niveau transcriptionnel. L'ensemble des acteurs n'étant pas tous exprimés au même moment, certains seraient donc plus aptes que d'autres à interagir entre eux. Ce type de données pourraient être incorporées à des réseaux PPI, tel que celui créé par Dortay et al par l'approche de double hybrides en levure (Dortay et al., 2008), la considération conjointe d'informations de similarité de séquence, de conservation de profils d'expression et d'interactions protéines-protéines, ne pouvant que renforcer la robustesse des résultats obtenus. Confronter co-expression et orthologie appelle forcément à considérer un aspect évolutif sur la fonction des gènes. Au même titre que des paralogues au sein d'une espèce, même présentent un taux de similarité très important mais des fonctions potentiellement divergentes, deux orthologues dans deux espèces différentes peuvent ne pas exercer la même fonction, ou tout du moins ne pas avoir les mêmes interacteurs. La notion d'expressologue, c'est-à-dire de gènes de différentes espèces partageant un profil d'expression similaire, prend sa source dans cette observation. Les études à venir qui prendront en compte davantage d'espèces devraient permettre de mieux comprendre ces aspects évolutifs. Ceci devrait être possible avec le développement de nouveaux algorithmes pour l'alignement et la comparaison inter-espèce de réseaux biologiques. Effectivement, nous avons été confrontés à la difficulté de comparer



plusieurs espèces en même temps. Les relations d'orthologie sont complexes (1 pour 1, 1 pour plusieurs, plusieurs pour 1 et plusieurs pour plusieurs (typiquement le cas des RR)) et rendent délicate la comparaison de réseaux de co-expression. Pour ces derniers, l'utilisation des transcriptomes de référence garantit une certaine qualité sur la représentation des transcrits. Nous nous sommes limités, pour l'instant, à des comparaisons deux à deux grâce au programme Pinalog (Phan & Sternberg, 2012), qui offre à la fois, un temps de calcul adapté et, une sortie qui peut être traitée facilement sous R.

### **Valider les associations détectées dans les réseaux**

Les associations détectées ou les nouveaux gènes identifiés comme impliqués dans le processus d'intérêt *in silico* doivent être validés expérimentalement. Cette validation peut se faire par invalidation de gènes ou caractérisation *in vitro* après production de protéines recombinantes. La caractérisation de mutants, notamment chez les espèces modèles comme *A. thaliana*, est également une manière puissante de valider ces associations (par exemple : (Mutwil et al., 2011a; Ruiz-Sola et al., 2016; Hansen et al., 2018)).

Toutefois, dans le cas où les associations détectées sont trop nombreuses, ou lorsque les modèles expérimentaux sont plus lourds à mettre en place (typiquement sur des espèces non-modèles avec des croissances plus lentes ou bien des réticences à la transformation), un renforcement des associations candidates *in silico* peut se révéler fortement utile. En plus de la force statistique des associations détectées (*p*-value de la corrélation, valeur du rang, pertinence de la communauté,...), nous supposons qu'une détection simultanée dans des réseaux de nature différente confère un autre niveau de robustesse. Cette propriété notamment utilisée pour la création d'EnsembleNet (Hansen et al., 2018), sera discutée un peu plus loin. Une première manière d'ajouter ce niveau de robustesse est de comparer les réseaux obtenus à partir de données différentes, comme celles issues de RNA-seq et de microrrays. La structure même de ces deux types de données rend leur combinaison délicate. De plus comme abordé dans notre deuxième partie, rajouter de nouveaux échantillons sur une matrice d'expression déjà large ne garantit pas une meilleure performance. Plutôt que de chercher à homogénéiser les données de l'une et de l'autre technologie, par exemple avec certaines méthodes de normalisation (Giorgi, Del Fabbro & Licausi, 2013), une perspective intéressante est plutôt de les considérer de manière complémentaire en exploitant leurs divergences, (i) techniques, car les deux approches fonctionnent de manière totalement différentes et (ii) thématiques, car



bien qu'une certaine partie des échantillons soit potentiellement commune, le contenu des deux matrices doit différer de manière importante, au niveau de la représentation de conditions expérimentales. Par conséquent, la structure des réseaux résultant diffère de manière assez importante comme nous avons pu le montrer dans les parties 1 et 2 de ce travail. Cependant, lors de l'étude de voies spécifiques, un certain nombre d'associations ont été tout de même retrouvées conservées entre les deux types de réseaux. Cette co-occurrence pourrait être dûe au hasard, à une propagation d'une erreur technique ou de calcul avec les deux jeux de données ou bien avoir une réelle pertinence biologique. L'évaluation statistique de la validité n'est pas évidente car elle implique une intersection entre deux jeux de données de dimensions différentes. Une analyse par génération aléatoire de réseaux sera à envisager pour estimer la probabilité de trouver une telle co-occurrence. Toutefois, en se reposant sur les domaines fonctionnels et annotations de protéines, il semblerait que ces associations décrivent une réalité biologique.

Par exemple dans le cas de la voie des phénylpropanoïdes, le réseau de co-occurrence entre microarray et RNA-seq présentait une communauté contenant des gènes guides associés au premier bloc de la voie (phénylalanine vers p-coumaroyl-CoA) mais aussi des gènes associés au métabolisme du shikimate qui est une connexion évidente avec ce même bloc. La co-occurrence concernant la voie des CK reste cependant plus complexe à interpréter au niveau biologique, montrant encore une fois la complexité de mettre en évidence des associations transcriptionnelles au niveau de cette voie de signalisation. Il est possible que l'étude de ce genre de voie nécessite de travailler à des seuils de corrélation moins stringents que ceux applicables aux voies métaboliques, en raison de leur intégration cellulaire et physiologique plus étendue. En effet, la voie de signalisation pourrait être considérée comme en amont d'un grand nombre de processus d'induction ou de répression (éventuellement condition-spécifique) d'expression de gènes, complexifiant les relations transcriptionnelles à mettre en évidence.

L'étude de la co-occurrence entre réseaux agrégés sera à ce titre prometteuse pour l'étude de la voie CK. La co-occurrence entre microarray et RNA-seq présentait assez peu d'associations mais nous avons observé une association très intéressante entre une polyamine oxidase (AT4G29720) et plusieurs RR de type A, l'ensemble de gènes appartenant à une même communauté au sein du réseau. Il a été très récemment montré que cette polyamine oxidase



intervient dans le contrôle de la différentiation du xylème, processus connu pour être régulé par les CK (**Figure 19**; (Alabdallah et al., 2017) Alabdallah et al 2017). Ainsi, des associations biologiques spécifiques peuvent être mises en évidence de cette manière, et de nouvelles associations mériteraient d'être considérées, en abaissant le seuil de stringence.

La seconde manière de renforcer le poids des associations détectées est de comparer les réseaux construits sur plusieurs espèces. Comme décrit ci-dessus, la génomique comparative est un outil puissant car la mise en évidence d'associations, ou de fonctions conservées, tend à servir de validation. De plus, ce genre d'analyse permet de mettre en lumière comment des modules transcriptionnels ont évolué. Selon Ruprecht et al 2017, au sein d'une même grande lignée taxonomique (par exemple les plantes vasculaires), les gènes plus anciens forment des modules globalement conservés dans les réseaux de co-expression (Ruprecht et al., 2017). De plus, il est possible de mettre en évidence les événements de duplication ou de spéciation associés à l'évolution des gènes. Dans notre cas, nous cherchions à montrer que des associations transcriptionnelles spécifiques entre gènes guides de la voie CK étaient conservées dans plusieurs espèces. Dans les premiers réseaux obtenus sur cette voie chez *Arabidopsis*, ces associations avaient été mises en évidence mais leur validation fonctionnelle reste complexe. En effet, les CK régulent des processus difficilement quantifiables *in planta* (contrôle de la sénescence, de l'activité du méristème apical, différentiation du xylème, ...) et les mutants simples ne présentent généralement pas de phénotype particulier (Hwang, Sheen & Müller, 2012). Par l'approche multi-espèces, certaines de ces associations semblent être conservées et mériteraient d'être testées prioritairement. Dans la recherche d'une méthode de priorisation de gènes candidats, l'approche multi-espèces pourrait donc être une manière adéquate de rationaliser les validations.

## Aspect critiques

### *Construction*

L'utilisation des PCC-HRR semble convenir quelque soit le type de données d'expression initiales. Nos évaluations sur des espèces avec des données relativement larges amènent cependant d'autres questions. En particulier, il conviendrait à présent de tester, les différentes distances en fonction de la taille des matrices initiales, car il est possible que certaines manières de calculer ces distances puissent montrer une meilleure performance lorsque moins



de conditions sont retenues (soit aléatoirement, soit par leur ressemblance biologique). De plus, il sera pertinent d'inclure d'autres espèces non-modèles, disposant de moins de données, pour renforcer ces évaluations et l'applicabilité de la méthode. Notre étude sur la conservation de réseaux de co-expression ciblant la voie CK chez différentes espèces a révélé que la qualité du réseau variait entre espèces. Cette variation pourrait s'expliquer par plusieurs facteurs : (i) le nombre de gènes dans le génome et la ploïdie, (ii) la nature des conditions expérimentales disponibles et (iii) un fonctionnement moins conservé entre espèces éloignées taxonomiquement.

Pour certaines espèces, beaucoup de transcrits sont prédis pour un même gène, que ce soit pour des espèces polyploïdes comme le blé, ou diploïdes comme le soja. Les modèles de gènes incluant une importante variation de transcrits sont généralement construits à partir de données RNA-seq, permettant de confirmer leur existence ou leur simple identification. Nous ne savons pas à ce stade comment une telle redondance impacte la valeur des HRR. Il est à noter, cependant, que la construction des réseaux de co-expression pourrait justement être un moyen de hiérarchiser ces différents variants ou bien différents allèles d'un même gène, autrement que sur leur simple abondance absolue et leur pourcentage de représentativité par rapport aux autres.

Concernant les conditions expérimentales, nous avons choisi de considérer un maximum d'échantillons disponibles pour chacune des espèces car d'après notre seconde évaluation, les réseaux résultants devraient être robustes, même si une optimisation serait possible en agrégeant des réseaux construits à partir de conditions sous-échantillonées du jeu de données initial par k-moyennes comme décrit dans la partie 2. Cette approche sera très prochainement appliquée à l'analyse comparative des réseaux décrite dans la partie 3 pour compléter les résultats obtenus. Enfin, la comparaison de réseaux, et plus particulièrement de PLC entre espèces génétiquement éloignées, soulève toujours le problème de l'identification correcte des orthologues. Dans le cas d'une voie de signalisation, comme celle répondant aux CK, une importante variation est observée dans la structure de la voie pour les espèces appartenant aux plantes inférieures (*P. patens* et *C. reinhardtii*). Ceci est aussi valable pour les voies métaboliques, où certaines branches (notamment dans le cas du métabolisme spécialisé) sont restreintes à une famille, un genre voire une espèce. De plus, la structure même des organismes (unicellulaire vs pluricellulaire, invascularisé vs vascularisé) pose des questions



sur la manière dont la voie CK est mobilisée pour transmettre un signal. Cette évolution propre des voies métaboliques et de signalisation pourrait expliquer, en partie, les différentes qualités de PLC obtenues dans la partie 3. De plus, cela montre que la construction de PLC multi-espèces nécessite un important travail en amont pour la préparation des sets de gènes guides. Dans notre cas, nous avons étendu la PLC aux autres histidine kinases (impliquées dans la signalisation éthylène et lumière), afin de ne pas manquer d'éventuelles multifonctionnalités et interactions entre les voies.

Par ailleurs, le choix d'un seuil pour considérer comme significatives les associations entre gènes est une question complexe qui peut être abordée de plusieurs manières. La plus simple, consiste à fixer une valeur empirique arbitraire, notamment lorsque des CCs sont utilisés. Le CC peut aussi être élevé à une puissance beta, tel que suggéré dans WGCNA, pour augmenter la stringence sur la détection des associations. Pour rationaliser le seuil, certains algorithmes peuvent être utilisés, tels que celui inspiré par la théorie des matrices aléatoires (RMT ; (Luo et al., 2007; Gibson et al., 2013)). Des approches visant à optimiser des paramètres topologiques, tels que l'invariance d'échelle peuvent aussi être envisagées (Couto, Comin & da Fontoura Costa, 2017). Dans notre cas, nous avons suivi la qualité de construction de réseaux sur une gamme de plusieurs seuils. Les meilleures PLC retenues avaient, par exemple, une transitivité et une modularité élevées. Les seuils ont donc été instinctivement retenus selon ces critères pour l'analyse des PLC. Ces paramètres topologiques sont généralement associés à une structure modulaire importante avec des groupes de gènes, plus connectés entre eux qu'aux autres gènes du réseau, facilitant la détection des communautés et l'interprétation biologique.

### *Evaluation*

Nous avons pu démontrer que la construction des réseaux par PCC-HRR fonctionne de manière satisfaisante avec les données microarrays et RNA-seq. Dans le cas d'une application à une ou des espèces non-modèles, il est attendu que les réseaux ainsi construits soient robustes. Cependant, une évaluation par des associations connues expérimentalement ou transférées *in silico*, constituant un set de référence, est toujours préférable. De tels sets sont généralement indisponibles pour des espèces non-modèles, et dans ce cas l'approche consiste à transférer *in silico* des annotations fonctionnelles de type GO par homologie. L'évaluation

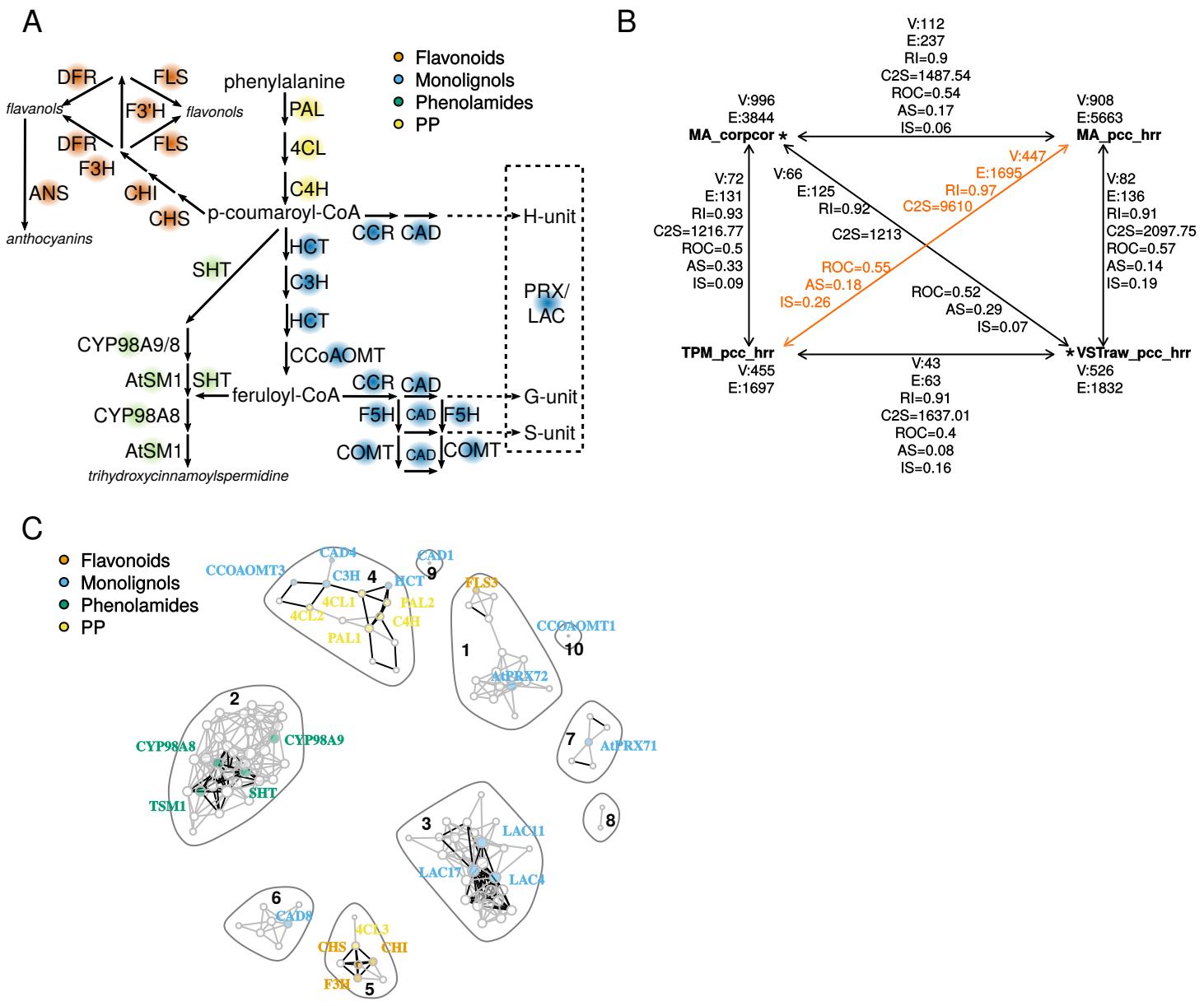


des réseaux peut se faire par des calculs de surreprésentation de termes fonctionnels, ou bien par le calcul d'une AUROC renseignant sur la prédictivité d'un réseau, face à ses associations de référence. Alors qu'un nombre de termes significativement enrichi sera meilleur d'autant plus qu'il sera élevé (mais jusqu'à quelle valeur ?), l'AUROC est une valeur absolue comprise entre 0,5 et 1, cette dernière valeur indiquant une prédictivité parfaite.

Au-delà de ces aspects techniques se pose la question de la qualité du set de référence. Des associations inférées *in silico* uniquement, peuvent être partielles ou mal adaptées, et donc leur utilisation conduire à une baisse apparente de la performance du réseau. La performance limitée des réseaux de *S. lycopersicum*, vue en deuxième partie, indique que beaucoup de gènes récupérés n'ont pas de termes GO attribués, suggérant une faiblesse dans l'annotation disponible sur Agrigo, pour cette espèce. La comparaison de réseaux de co-expression entre espèces pourrait cependant corriger cela. La preuve est montrée par l'exemple développé sur la voie JA. La PLC, construite sur un agrégat obtenu par co-occurrence, et en utilisant des gènes guides indiqués dans la base de données Plant Metabolic Networks (PMN) révèle la présence d'autres gènes très certainement impliqués dans la synthèse de JA, telle qu'une oxophytodienic acid réductase différente de celle indiquée dans PMN chez *S. lycopersicum*. D'autre part, nous avons également observé que l'évaluation des PLC sur les voies métaboliques, par l'enrichissement en termes GO, n'était pas toujours le meilleur critère de qualité. Mesurer la pertinence de la séparation des gènes guides dans les communautés s'est également avéré très utile pour identifier les distances de calcul les plus robustes. Dans le cas de voies fonctionnant par branches spécifiques, telle que celle des phénylpropanoïdes, des blocs dont la composition en gènes guides s'apparentent à ces branches spécifiques ont pu être correctement séparés dans la PLC. Ceci fonctionne d'autant mieux que les gènes présentent une spécificité d'expression remarquable, tels que les gènes impliqués dans la synthèse des phénolamides, au cours de la maturation du pollen.

### *Comparaison de réseaux*

Un aspect, qui a également été problématique, concerne la comparaison de réseaux et plus généralement leur alignement. La comparaison de deux réseaux d'une même espèce peut s'effectuer très simplement au niveau des liens conservés. Cependant, il est possible de considérer le voisinage à une distance donnée de chaque gène et de regarder si des liens



**Figure 26: Alignement de PLC microarrays et RNA-seq ciblant le métabolisme des phénylpropanoïdes chez *Arabidopsis thaliana*.** A, organisation de la voie en 4 blocs, le précurseur étant la voie PP (Précurseur Phénylpropanoïdes). B, comparaisons deux à deux de 4 PLC issues de jeux de données différents: microarray avec corrélations partielles (MA\_corpcor) ou avec PCC-HRR (MA\_PCC\_HRR) et RNA-seq exprimés en Transcrits Par Millions avec PCC-HRR (TPM\_pcc\_hrr) ou exprimés en comptes brutes normalisés par stabilisation de variance (VSTraw\_pcc\_hrr). Ces 4 PLC étaient les plus performantes parmi l'ensemble de celles testées dans la partie 1 de ce travail. Pour chaque PLC et comparaison de PLC deux à deux sont indiqués: V, le nombre de noeuds; E, le nombre de liens; RI, Rand Index pour la distribution des gènes guides en communautés selon leur blocs théoriques; C2S, test de Chi<sup>2</sup> pour la distribution des gènes guides en communautés selon leur blocs théoriques; ROC, performance de capture de termes GO; AS, score d'alignement; ICS, similarité par structures conservées. Ces deux derniers indices ont été calculés grâce au paquet "netcom" de R, s'inspirant de l'algorithme présenté dans GHOST (Patro & Kingsford, 2012). La comparaison en orange reflète le plus de similarités entre les deux PLC. C, PLC d'intersection entre RNA-seq et microrarrays.

adjacents sont aussi conservés. Cette procédure est appelée alignement de réseau, à l'image des alignements entre séquences nucléiques ou protéiques, dans lesquels des interruptions et des erreurs peuvent être tolérées. Dans le cas d'un alignement de réseaux, l'idée est de mesurer une valeur de similarité entre deux réseaux, par rapport à la conservation de structures topologiques ou d'enchaînements de nœuds. Les méthodologies d'alignement ne sont pas très nombreuses (Clark & Kalita, 2014) et leur utilisation s'est avérée parfois complexe. Ainsi nos comparaisons de réseaux RNA-seq et microarray se sont limitées à la conservation simple des liens. En tenant compte du voisinage des gènes, une plus grande similarité aurait pu être observée. Dans le cas de deux gènes guides connectés à un ensemble de gènes marqueurs différent fortement, entre réseaux RNA-seq et microarray, une comparaison simple des liens montrerait une faible similarité, alors que biologiquement, ce rapprochement des deux gènes guides pourrait refléter la différence des conditions expérimentales testées, et donc l'association de ces gènes avec différentes fonctions selon les conditions considérées. Cette différence, serait donc représentative d'une hétérogénéité des conditionnons expérimentales dans les jeux de données, plutôt que de différences inhérentes aux deux technologies.

La mise en place d'algorithmes d'alignement de réseaux plus performants serait un réel avantage pour les études à venir. Un des outils disponibles s'appelle GHOST (Patro & Kingsford, 2012). Cet algorithme s'appuie sur des signatures spectrales appliquées à des chemins de type *seed-and-extend* pour comparer deux réseaux. Une similarité est calculée d'après la conservation de structures dans les deux réseaux. En utilisant cette approche, nous avons montré que les PLC obtenues par microarrays et RNA-seq étaient plus comparables lorsque les distances sont calculées avec les PCC-HRR (**Figure 26**).

La comparaison devient d'autant plus complexe lorsque les deux réseaux à comparer proviennent d'espèces différentes. Cette comparaison nécessite une phase d'identification préalable des orthologues, d'une espèce pour l'autre, généralement à l'aide d'un BLAST bidirectionnel (espèce A contre B, espèce B contre A). Les relations d'orthologie sont complexes : un gène pour un gène, un pour plusieurs, plusieurs pour un, plusieurs pour plusieurs (typiquement le cas des acteurs des voies MSP). Ces relations pourraient être visualisées elles-mêmes sous forme de réseau. Pour déterminer si un lien présent dans un

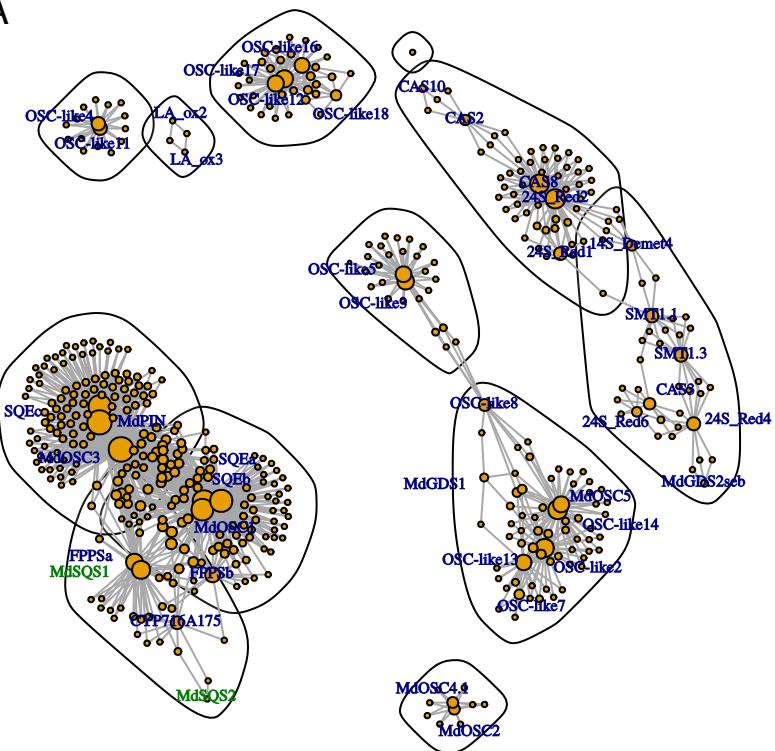


réseau d'une espèce est conservé dans l'autre, il faut pouvoir tenir compte de cette complexité. Le problème s'accroît si plusieurs espèces doivent être comparées. Il existe quelques outils disponibles pour traiter ces points mais ils se limitent généralement à deux espèces. Certains d'entre eux présentent une sortie qui est difficilement traitable pour les analyses en aval. Une alternative consiste à remplacer les gènes espèces-spécifiques par leurs domaines fonctionnels (Pfam par exemple) tel qu'utilisé dans la base de données PlaNet. Il est important de garder en tête que ces comparaisons d'expression inter-espèces sont complémentaires aux recherches de similarité de séquences et peuvent aider à comprendre l'évolution des organismes (Ruprecht et al., 2017).

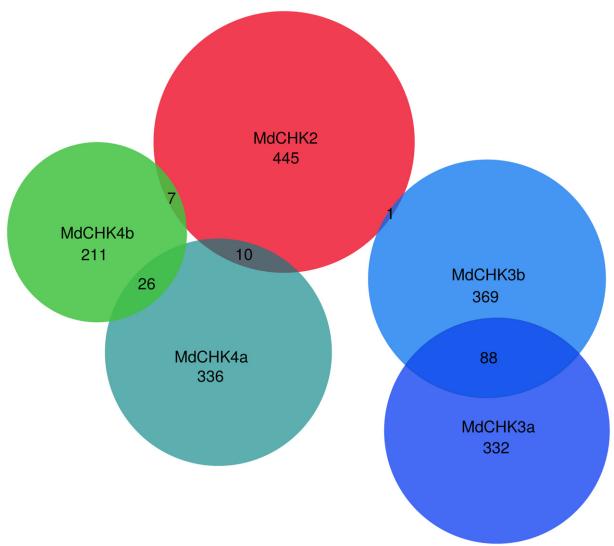
### **Application/transfert**

Lorsqu'une étude de co-expression large est menée, le premier obstacle est la taille des données en elles-mêmes. Le calcul de distances entre un grand nombre de gènes nécessite d'importantes ressources de calcul, de mémoire ainsi que de temps et peut être irréalisable sur un ordinateur personnel. Si ceci est vrai pour les CC, la difficulté est encore plus importante pour d'autres types de distances. Le calcul des MI s'est avéré particulièrement délicat et, dans le cadre de notre travail, seul le paquet 'Parmigene' a été efficace pour les obtenir en multicoeurs. Les calculs conduits avec d'autres paquets, tels que 'PCIT' n'ont pas abouti, même sur le cluster de calcul Artemis. Le problème a été similaire pour les PC. L'approche par calcul des coefficients b d'une régression linéaire multiple à laquelle une restriction Lasso (quand il y a plus de gènes que de conditions) a été appliquée, n'a pas pu être mise en place sur les matrices complètes. Il s'est toutefois avéré, que calculer l'inverse d'une matrice de covariance, obtenue à l'aide d'estimateurs était relativement efficace. Cependant, la performance des réseaux obtenus dans ces cas-là restait inférieure au CC ordonnés dans nos conditions. Pour le calcul des CC et leur ordonnancement réciproque, nous avons pu mettre au point un programme écrit en C permettant de répartir les nombreux calculs sur différents processeurs. Dans ce cas, chaque processeur traite une sous-matrice  $m$  gènes x  $p$  conditions, extraite d'une matrice initiale  $n$  x  $p$ , où  $m < n$  et est un multiple du nombre de processeurs. Lors de l'étape de détermination des HRR, la difficulté est d'établir une communication entre processeurs, de manière à obtenir les rangs qui peuvent être traités sur un autre processeur. La technologie MPI (Message Passing Interface) a été utilisée dans ce but.

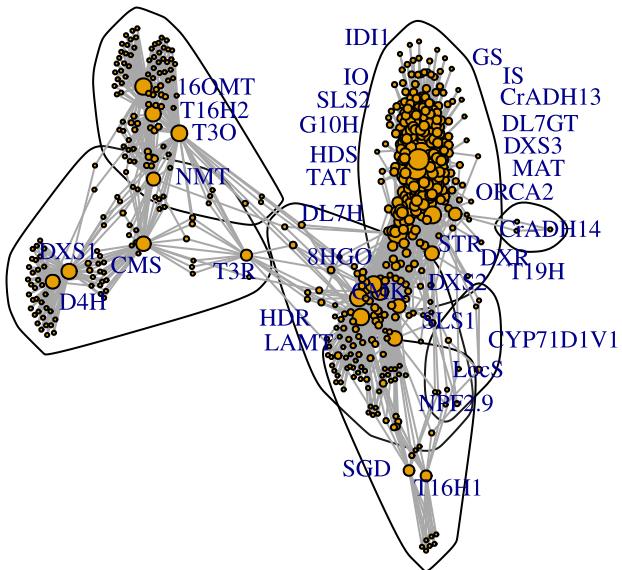
A



B



C



**Figure 27: Application des réseaux de co-expression à des différentes thématiques.** A, réseau de co-expression de type PLC chez le Pommier pour la synthèse des triterpènes et des phytostérols (Navarro Gallon et al 2017). Les deux isoformes de SQS sont indiquées en vert. SQS2 est peu connecté aux autres gènes du réseau. La taille des noeuds est proportionnelle à leur degré. Les polygones délimités en noir définissent des communautés déterminées par un algorithme de type "fast greedy" qui cherche à optimiser la modularité. B, nombre de gènes communément co-exprimés entre les différents récepteurs aux CK du Pommier (CHK). Le faible nombre de gènes dans les intersections révèle la spécificité d'expression et probablement d'implication dans des processus spécifiques. C, réseau de co-expression de type PLC chez *Catharanthus roseus* pour la synthèse des alcaloïdes indoles monoterpéniques. Les communautés à gauche regroupent notamment le bloc associé à synthèse de la vindoline (T16H2, 16OMT, T3O, T3R, NMT et D4H). Voir la Figure XX du contexte pour les voies complètes.

**Table VI: Temps indicatifs pour le calcul de différentes distances, sur les matrices d'expression d'*Arabidopsis thaliana* obtenues par microarrays ou RNA-seq.**

Distance	PC	MI	PCC	SCC	PCC-HRR	SCC-HRR
Programme	'corpcor' R	'Parmigene' R	Parallélisation en C			
microarrays	13h	2 jours	2h	2h	2h	2h
RNA-seq	1h30	2 jours	3h	3h	3h	3h

Les calculs ont tous été réalisés sur une machine possédant 20 coeurs (processeurs Xeon) et 64 Go de RAM.

Comme indiqué dans la **Table VI**, les temps de calculs obtenus par notre programme sont relativement courts. Un temps de calcul plus court est important car il permet de traiter un grand nombre de matrices d'expression plus rapidement. Cela serait notamment le cas pour la construction d'agrégats de réseaux de co-expression obtenus à partir d'une ou plusieurs matrices sous-échantillonnées comme décrit dans la partie 2.

La mise en place de cet outil a clairement permis d'apporter une méthode efficace et transférable à d'autres thématiques travaillées au laboratoire. En effet, des études de co-expression ont pu être menées sur la voie des triterpènes et des CK chez le pommier, ainsi que sur le métabolisme des alcaloïdes indoles monoterpéniques (AIM) chez la Pervenche de Madagascar (**Figure 27**). Par ces approches de co-expression, nous avons pu analyser les environnements transcriptionnels de deux isoformes de squalène synthases (SQS) du pommier ((Navarro Gallón et al., 2017), **article en annexe**). La SQS catalyse la formation de squalène, un composé-clé dans le métabolisme des triterpènes. Nous avons mis en évidence, l'implication différentielle de ces deux isoformes, l'une, SQS2 étant inductible et participant à la réponse de la plante à des agressions extérieures. En effet, contrairement aux autres gènes de la voie, SQS2 n'est que très faiblement connecté aux autres gènes du réseau. Ceci suggère un fonctionnement basal de cette voie, impliquant SQS1 qui ne montre pas de spécificité d'expression, et un fonctionnement inductible, impliquant SQS2. Dans ce cas, la transcription accrue du gène codant SQS2 permettrait d'alimenter le flux en squalène en cas de conditions physiologiques particulières.

Concernant la voie des CK chez le pommier, nous avons appliqué notre méthodologie de récupération de données RNA-seq et de construction de réseau de co-expression pour établir



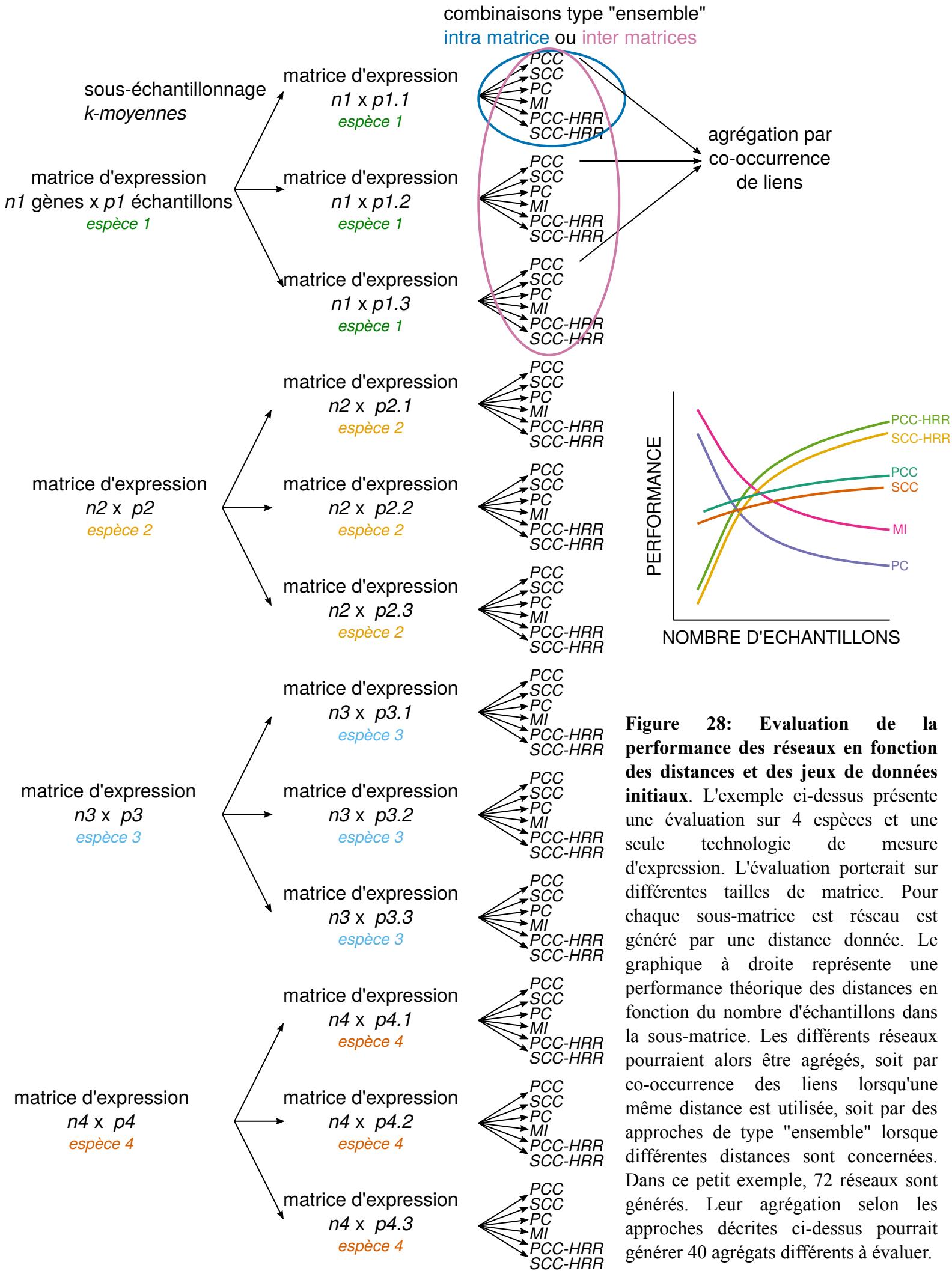
les paysages transcriptionnels des récepteurs HK au CK chez cette plante ((Daudu et al., 2017), **article en annexe**). Le Pommier possède effectivement 5 formes de CHK, MdCHK2, MdCHK3.1, MdCHK3.2, MdCHK4.1 et MdCHK4.2, semblables au trois récepteurs CK d'*Arabidopsis*. Un très faible de nombre co-exprimés ont été retrouvés entre ces 5 formes, montrant clairement une spécificité d'expression et suggérant que les récepteurs sont impliqués dans des processus physiologiques particuliers. Ceci renforce les observations expérimentales de spécificités différentielles d'association, de ces 5 récepteurs avec des différentes formes de CK. De plus, comme démontré en partie I sur *Arabidopsis*, des associations transcriptionnelles spécifiques ressortent de la PLC avec les gènes de cette voie renforçant davantage cette notion de chemins spécifiques entre acteurs de la voie (**Figure 19**).

Enfin, bien que moins de conditions expérimentales étaient disponibles, un réseau de co-expression a été construit pour *Catharanthus roseus* afin d'élucider une partie des étapes de synthèse des AIM. Bon nombre d'étapes de cette voie ont été découvertes par des approches expérimentales, et dernièrement plusieurs ont été mises en évidence par des approches bio-informatiques. En utilisant un ensemble de conditions incluant une élicitation biotique par des larves de *Manduca sexta* ((Dugé de Bernonville et al., 2017), **article en annexe**), une PLC a été construite. De cette manière, une communauté de gènes rassemblant des étapes clés situées après la strictosidine (le premier MIA précurseur), dans la voie de biosynthèse, a pu être mise en évidence, permettant l'identification de plusieurs nouvelles étapes dont une catalysée par la Tabersonine synthase (Caputi et al., 2018). La tabersonine est le précurseur de la vindoline, composé capable de se coupler avec la catharanthine, une molécule très proche de la tabersonine, pour former l'anhydrovinblastine puis la vinblastine, un puissant anticancéreux. Ainsi, de telles percées dans l'identification des gènes ouvre de nouvelles perspectives aux technologies de biologie synthétique pour produire ces composés à fort intérêt thérapeutique, à moindre coût.

## Perspectives

### *Evaluations supplémentaires*

Dans notre processus d'évaluation des méthodologies d'inférence de réseaux de co-expression, une combinaison supplémentaire devra être intégrée. Il s'agit de la combinaison



**Figure 28: Evaluation de la performance des réseaux en fonction des distances et des jeux de données initiaux.** L'exemple ci-dessus présente une évaluation sur 4 espèces et une seule technologie de mesure d'expression. L'évaluation porterait sur différentes tailles de matrice. Pour chaque sous-matrice est réseau est généré par une distance donnée. Le graphique à droite représente une performance théorique des distances en fonction du nombre d'échantillons dans la sous-matrice. Les différents réseaux pourraient alors être agrégés, soit par co-occurrence des liens lorsqu'une même distance est utilisée, soit par des approches de type "ensemble" lorsque différentes distances sont concernées. Dans ce petit exemple, 72 réseaux sont générés. Leur agrégation selon les approches décrites ci-dessus pourrait générer 40 agrégats différents à évaluer.

taille de matrice x distance x espèce (**Figure 28**). En effet, notre évaluation des distances s'est portée sur les matrices d'Arabidopsis contenant plus de 1000 échantillons. Il serait pertinent de vérifier que les PCC-HRR sont tout aussi performants sur des matrices plus petites comparé à d'autres mesures, car d'autres distances type PC et MI pourraient présenter une performance accrue sur une matrice moins complexe. Cette vérification sera d'autant plus nécessaire si l'agrégation de réseaux issus de matrices plus petites s'avère être une technique de choix. Dans ce cas là, il faut s'assurer que la distance estimée entre les gènes sur ces matrices sous-échantillonnées reste toujours pertinent avec les PCC-HRR. En envisageant une évaluation similaire à celle faite pour la taille de matrice (3 espèces, 2 technologies), le nombre de calculs à tester va cependant être conséquent. L'avantage de nos résultats présentés en partie 2, est qu'ils suggèrent que la sélection aléatoire des échantillons n'améliore pas la performance sur les agrégats de réseaux (alors que les réseau individuels eux semblent l'être). Le sous-échantillonnage aléatoire dans ce cas n'aurait donc plus à être évalué. Le nombre de matrices sous-échantillonnées par k-moyennes peut varier entre une vingtaine et une cinquantaine, selon l'espèce et la technologie, et pourrait nécessiter l'analyse de plus de 1000 matrices, pour une méthode de sous-échantillonnage seulement. La manière de visualiser l'ensemble de ces données risque également d'être délicate, et sera l'occasion de développer de nouveaux moyens, pour en tirer un maximum d'informations.

L'effet « espèce » sur la construction de réseaux pourra être davantage approfondi. Comme vu dans la partie 3, la qualité des PLC est variable selon l'espèce considérée (voir plus haut). Il reste à déterminer, si cette variation est uniquement liée à la taille de la matrice, aux conditions expérimentales représentées ou à une complexité supérieure de l'organisme considéré. Une évaluation tailles x distances x espèces, aussi colossale soit-elle, pourrait permettre de répondre à cette question. L'objectif serait de pouvoir définir une distance optimale, en fonction des conditions expérimentales impliquées, dans le cas où, où des méthodes autres que les PCC-HRR capturent des associations significatives plus efficacement, sur des tailles d'échantillons plus faibles. De plus, il serait pertinent de mesurer un indice sur les échantillons présents dans la matrice qui permettrait de prédire la performance du réseau en résultant. Dans la partie 2, nous avons vu que des échantillons trop fortement corrélés (tels que dans les matrices sous-échantillonnées par k-moyennes) étaient associés à des réseaux moins performants que lorsque les échantillons étaient sélectionnés au



hasard. Dans le cas d'espèces non-modèles, pour lesquelles peu ( $<100$  conditions) sont disponibles, l'approche par clustering des conditions, par k-moyennes et agrégation doit être plus difficilement réalisable. Sur ce type de matrices, mesurer un tel indice pourrait permettre d'estimer si la capture des associations correspond plutôt à une capture « moyenne » (échantillons peu corrélés, comme pour l'échantillonnage aléatoire), ou à une capture spécifique (échantillons fortement corrélés). En effet, le fait que les agrégats issus de réseaux obtenus par échantillonnage type k-moyennes aient une performance supérieure à ceux obtenus par échantillonnage aléatoire, suggère qu'une plus grande diversité d'associations est représentée dans l'agrégat.

### *Dissémination*

Une des perspectives de notre travail va être la création d'un ensemble de programmes utilisant en entrée une matrice d'expression pour inférer plusieurs réseaux, et capable après sous-échantillonnage (par k-moyennes) de les agréger afin d'obtenir un réseau optimal. Notre approche qui reste conceptuellement simple pourrait être diffusée de manière plus large via un tel pipeline, permettant une étude de co-expression robuste, sur n'importe quel organisme pour lequel des données d'expression seraient disponibles. Avec la mise en place d'espaces de calculs par Google ou Amazon, il est maintenant très simple d'avoir une machine de calcul virtuelle à très faible coût, ce qui rendra encore plus simple ces analyses de co-expression impliquant de larges données. Une telle mise en place pourrait se faire par le langage Python qui se développe de plus en plus dans la communauté bio-informatique. A l'inverse des bases de données pré-calculées disponibles en ligne, notre approche serait plutôt dédiée à une prise en main complète de la démarche de construction du réseau de co-expression, à l'image de ce qui est proposé dans le paquet 'WGCNA'. Une amélioration potentielle serait aussi de pouvoir intégrer de manière régulière les nouveaux échantillons déposés dans le SRA ou l'ENA. Ces intégrations ne devraient *a priori* pas avoir d'effet pertinent sur des réseaux utilisant la matrice globale, mais cela pourrait jouer sur des réseaux inférés après sous-échantillonnage. De plus dans l'objectif de rendre accessible notre méthodologie aux biologistes, nous pourrions envisager d'intégrer dans notre programme un outil permettant de récupérer et de préparer des données disponibles dans les bases de données.

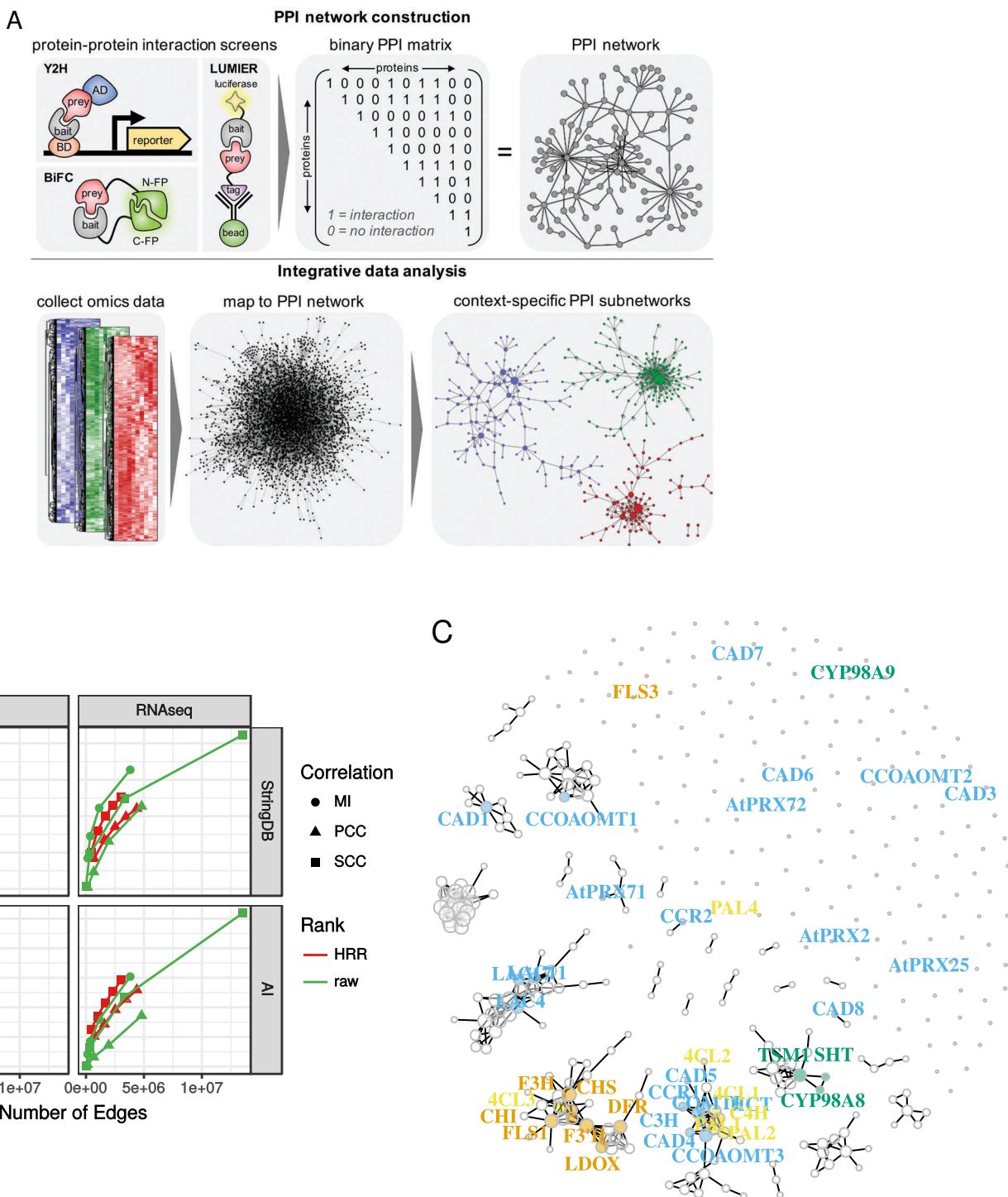


### *Ensemble networking*

Une des perspectives à envisager serait l'utilisation des approches « ensemble » (Li, Pearl & Jackson, 2015). L'étude conduite par Hansen et al (Hansen et al., 2018) combine des données de réseaux de différents types, impliquant de la co-expression de gène ou des PPI, pour identifier les gènes voisins d'un ensemble de gènes guides les mieux représentés dans l'ensemble de ces sources. Les auteurs ont développé une méthodologie appelée Neighbour Counting Ensemble, qui utilise le nombre d'occurrences de termes GO autour des gènes guides, dans les différents réseaux. Dans notre cas, une combinaison de réseaux issus de différents calculs pourrait être pertinente. En effet, il est envisageable que chacune des 6 distances testées dans la partie 1, soit en mesure de mettre en évidence certains types d'associations (de Siqueira Santos et al., 2013). Dans l'absolu, c'est peut-être une combinaison exhaustive de différentes distances, différentes tailles et différentes espèces qui pourra révéler un réseau optimal. Ces méthodologies « ensemble » font appel à des algorithmes plus complexes, que la manière dont nous avons réalisé nos agrégats de réseaux basée sur le nombre de co-occurrence des liens ou sur leur significativité. La comparaison de la performance de toutes ces méthodologies d'agrégation sera informative.

### *Inférences multi-échelles*

Comme décrit dans le paragraphe précédent, l'inférence de réseaux utilisant différents types de données, moyennant une méthodologie efficace d'intégration, permet d'augmenter leur performance. Ceci est le cas, par exemple, des données présentées dans les bases Aranet et StringDb. Ces bases fournissent des réseaux dont les liens peuvent être de différentes natures : interactions protéine-protéines, co-expression, co-occurrence textuelle etc. Le quantité et le type de données potentiellement intégrables est grande comme décrit en introduction. Dans le cas de voies de signalisation, pour lesquelles les réseaux de co-expression semblent présenter quelques limitations, cumuler les données d'expression avec des quantifications de protéines, par exemple pourrait potentiellement renforcer leur performance. L'intégration de données -omiques peut se faire de différentes manières, mais les aspects méthodologiques restent un champ d'investigation d'actualité (Rai, Saito & Yamazaki, 2017). Une approche assez simple consiste à aligner des réseaux -omiques à un réseau de type PPI construit à l'aide de tests haut



**Figure 29: Intégration de données d'interactions protéine-protéines (PPI) dans la construction de données biologiques.** A, méthodologie d'acquisition haut débit de données PPI par double hybride (Y2H), complémentation de fluorescence bimoléculaire (BiFC) et interactions chez les mammifères basées sur la luminescence (LUMIER) pour la construction d'un réseau PPI. Sur ce réseau PPI peuvent être indiquées d'autres associations inférées par d'autres types de données (figure tirée de Robinson & Nielsen, 2016). B, nombre de PPI correspondant à des associations transcriptionnelles dans les jeux d'expression d'*Arabidopsis thaliana* en fonction du nombre de liens contenus dans le réseau. 5 distances sont évaluées: information mutuelle (MI), PCC, SCC, PCC-HRR et SCC-HRR. L'évaluation est faite contre la base générique StringDb ou bien contre Arabidopsis Interactome (AI). C, réseau de co-occurrence PPI (tirées des meilleures connexions de StringDb) et co-expression pour la PLC ciblée sur les phénylpropanoïdes.

débit d’interactions protéines-protéines (Robinson & Nielsen, 2016)(**Figure 29**). Les analyses multivariées sont particulièrement appropriées à l’intégration multi-omique, comme l’analyse canonique des correspondances (CCA) qui permet de confronter deux matrices représentant deux jeux d’observations faites sur les mêmes individus. Des paquets R tels que ‘mixOmics’ ont été développés pour permettre ces intégrations multi-dimensionnelles (Rohart et al., 2017). La difficulté reste d’obtenir des jeux de données complets, de préférence présentant des observations obtenues sur les mêmes individus. La CCA robuste (rCCA) permet cependant de prendre en compte la variabilité introduite si des individus différents, bien que traités similairement, sont utilisés. Bien que non montré dans notre évaluation en partie 1, nous avons comparés réseaux PPI et co-expression chez *Arabidopsis thaliana* (**Figure 29**). Le nombre de liens dans le réseau de co-expression correspondant à des PPI décrits dans la base STRING, augmente avec la diminution de la stringence du seuil de significativité des associations. Nous avons pu montrer que les PCC-HRR maximisaient bien la capture de ces PPI. La co-occurrence des liens de co-expressions avec les PPI présents dans StringDb au niveau de la PLC ciblée sur les phénylpropanoïdes, constitue un exemple remarquable de conservation entre les deux types de réseaux.



# Conclusion



D'une manière générale, avec l'émergence du *big data*, la manipulation de données larges devient un requis incontournable en recherche biologique. Ces données peuvent être massives par le nombre de conditions expérimentales, d'individus, ou par la combinaison de données issues de différentes études -omiques. Elles permettent une vue large de systèmes vivants, qui comme abordé en introduction, ne peuvent se résumer à l'interaction de quelques molécules, et permettent ainsi un aperçu et une étude plus globale de leur fonctionnement. A titre d'exemple de ce que peut représenter cette quantité colossale de données et les études à très grande échelle associées, une publication très récente, utilise les données de 1 191 588 individus et étudie les redondances de facteurs génétiques impliqués dans diverses pathologies, illustrant encore une fois l'inter-connectivité existante au sein des organismes vivants (Anttila et al., 2018).

Propulsé par l'acquisition massive de données RNA-seq, le traitement de données transcriptomiques s'est considérablement développé ces dernières années dans le but d'étudier des processus biologiques diversifiés. La visualisation de données transcriptomiques par des réseaux de co-expression a l'avantage de pouvoir mettre en évidence des associations entre transcrits, qui peuvent refléter le fonctionnement global d'un processus biologique, et potentiellement prédire une fonction pour un gène donné selon le principe de « culpabilité par association ». Au cours de ce travail, nous avons mis en lumière les points critiques rencontrés lors de la construction des réseaux de co-expression, et entrepris d'optimiser certains d'entre eux, dans l'objectif de développer une méthodologie applicable à des plantes non-modèles telles que celles travaillées au sein de l'EA2106. L'intérêt de cette approche est d'exploiter des ressources transcriptomiques disponibles publiquement ou acquises *de novo*, pour étudier un processus biologique donné et identifier des associations candidates qui seront validées de manière fonctionnelle. Tout travail bio-informatique doit donc tenir compte de cet aspect et du temps qui sera consacré par d'autres chercheurs à cette validation expérimentale. Ceci implique de disposer de méthodes d'analyse robustes et fiables, limitant au maximum l'intégration de faux positifs.

Un des obstacles initiaux a été la taille des données, dépassant largement celles que les biologistes ont généralement l'habitude de manipuler. Nous avons pu toutefois mener à bien la construction et l'analyse de réseaux avec des ressources informatiques qui sont



relativement faciles d'accès. Le second obstacle est la complexité d'analyse qui nécessite une connaissance approfondie d'un langage de programmation. Dans notre cas, la puissance du langage R et les nombreux paquets disponibles (en particulier *via* la plateforme dédiée aux paquets d'analyses bio-informatiques Bioconductor), nous a permis de réaliser l'ensemble des opérations nécessaires à l'analyse des réseaux. Il s'agit d'un langage relativement simple qui pourrait convenir, si intégré dans une suite de programmes correctement documentés, aux biologistes désirant initier des études de co-expression. L'approche par PLC que nous avons optimisé en appliquant les PCC-HRR semble particulièrement adapté à l'étude de voies métaboliques en extrayant une information biologique précise d'un réseau très large. Si des améliorations sont toujours à envisager, l'approche reste très prometteuse dès à présent pour un large panel d'applications.



## Bibliographie



- Alabdallah O., Ahou A., Mancuso N., Pompili V., Macone A., Pashkoulov D., Stano P., Cona A., Angelini R., Tavladoraki P. 2017. The *Arabidopsis* polyamine oxidase/dehydrogenase 5 interferes with cytokinin and auxin signaling pathways to control xylem differentiation. *Journal of experimental botany* 68:997–1012.
- Altay G. 2012. Empirically determining the sample size for large-scale gene network inference algorithms. *IET systems biology* 6:35–43.
- Anders S., Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* 11:R106. DOI: 10.1186/gb-2010-11-10-r106.
- Anttila V., Bulik-Sullivan B., Finucane HK., Walters RK., Bras J., et al. 2018. Analysis of shared heritability in common disorders of the brain. *Science* 360.
- Aoki M., Mieda M., Ikeda T., Hamada Y., Nakamura H., Okamoto H. 2007. R-spondin3 is required for mouse placental development. *Developmental Biology* 301:218–226.
- Aoki Y., Okamura Y., Ohta H., Kinoshita K., Obayashi T. 2016. ALCOdb: Gene Coexpression Database for Microalgae. *Plant and Cell Physiology* 57:e3–e3.
- Ballouz S., Verleyen W., Gillis J. 2015. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* 31:2123–2130.
- Ballouz S., Weber M., Pavlidis P., Gillis J. 2016. EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics* 33:612–614.
- Barabási A-L., Albert R. 1999. Emergence of scaling in random networks. *science* 286:509–512.
- Barabási A-L., Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5:101–113.
- Bolstad BM., Irizarry R., Åstrand M., Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193.
- Botía JA., Vandrovčová J., Forabosco P., Guelfi S., D'Sa K., The United Kingdom Brain Expression Consortium, Hardy J., Lewis CM., Ryten M., Weale ME. 2017. An



additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Systems Biology* 11:47.

Bray NL., Pimentel H., Melsted P., Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34:525.

Brown DM., Zeef LAH., Ellis J., Goodacre R., Turner SR. 2005. Identification of Novel Genes in Arabidopsis Involved in Secondary Cell Wall Formation Using Expression Profiling and Reverse Genetics. *The Plant Cell* 17:2281.

Buness A., Huber W., Steiner K., Sultmann H., Poustka A. 2005. arrayMagic: two-colour cDNA microarray quality control and preprocessing. *Bioinformatics (Oxford, England)* 21:554–556.

Caputi L., Franke J., Farrow SC., Chung K., Payne RM., Nguyen T-D., Dang T-T., Carqueijeiro IST., Koudounas K., de Bernonville TD., others 2018. Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science*:eaat4100.

Chai LE., Loh SK., Low ST., Mohamad MS., Deris S., Zakaria Z. 2014. A review on the computational approaches for gene regulatory network construction. *Computers in biology and medicine* 48:55–65.

Chen JJ. 2007. Key aspects of analyzing microarray gene-expression data. *Pharmacogenomics* 8:473–482.

Clark C., Kalita J. 2014. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics* 30:2351–2359. DOI: 10.1093/bioinformatics/btu307.

Clauset A., Newman ME., Moore C. 2004. Finding community structure in very large networks. *Physical review E* 70:066111.

Consortium IHGS. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Consortium T 1000 GP. 2015. A global reference for human genetic variation. *Nature* 526:68–74.



- Cosgrove EJ., Gardner TS., Kolaczyk ED. 2010. On the choice and number of microarrays for transcriptional regulatory network inference. *BMC bioinformatics* 11:454.
- Couto CMV., Comin CH., da Fontoura Costa L. 2017. Effects of threshold on the topology of gene co-expression networks. *Molecular BioSystems* 13:2024–2035.
- Csardi G., Nepusz T. 2006. The igraph software paquet for complex network research. *InterJournal, Complex Systems* 1695:1–9.
- Daudu D., Allion E., Liesecke F., Papon N., Courdavault V., Dugé de Bernonville T., Mélin C., Oudin A., Clastre M., Lanoue A., Courtois M., Pichon O., Giron D., Carpin S., Giglioli-Guivarc'h N., Crèche J., Besseau S., Glévarec G. 2017. CHASE-Containing Histidine Kinase Receptors in Apple Tree: From a Common Receptor Structure to Divergent Cytokinin Binding Properties and Specific Functions. *Frontiers in Plant Science* 8:1614.
- De La Fuente A., Bing N., Hoeschele I., Mendes P. 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20:3565–3574.
- DeLuca DS., Levin JZ., Sivachenko A., Fennell T., Nazaire M-D., Williams C., Reich M., Winckler W., Getz G. 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 28:1530–1532.
- Des Marais DL., Guerrero RF., Lasky JR., Scarpino SV. 2017. Topological features of a gene co-expression network predict patterns of natural diversity in environmental response. *Proceedings of the Royal Society B: Biological Sciences* 284:20170914.
- D'haeseleer P., Liang S., Somogyi R. 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16:707–726.
- Dobin A., Davis CA., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.



- Dong X., Jiang Z., Peng Y-L., Zhang Z. 2015. Revealing Shared and Distinct Gene Network Organization in Arabidopsis Immune Responses by Integrative Analysis. *Plant Physiology* 167:1186.
- Dortay H., Gruhn N., Pfeifer A., Schwerdtner M., Schmu T., Heyl A. 2008. Toward an Interaction Map of the Two-Component Signaling Pathway of *Arabidopsis thaliana* research articles. *Journal of proteome research* 7:3649–3660.
- Dugé de Bernonville T., Carqueijeiro I., Lanoue A., Lafontaine F., Sánchez Bel P., Liesecke F., Musset K., Oudin A., Glévarec G., Pichon O., Besseau S., Clastre M., St-Pierre B., Flors V., Maury S., Huguet E., O'Connor SE., Courdavault V. 2017. Folivory elicits a strong defense reaction in *Catharanthus roseus*: metabolomic and transcriptomic analyses reveal distinct local and systemic responses. *Scientific Reports* 7:40453.
- Fabregat A., Korninger F., Viteri G., Sidiropoulos K., Marin-Garcia P., Ping P., Wu G., Stein L., D'Eustachio P., Hermjakob H. 2018. Reactome graph database: Efficient access to complex pathway data. *PLOS Computational Biology* 14:1–13.
- Feltus FA., Ficklin SP., Gibson SM., Smith MC. 2013. Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an *Arabidopsis* case study. *BMC systems biology* 7:44. DOI: 10.1186/1752-0509-7-44.
- Ferrari C., Proost S., Ruprecht C., Mutwil M. 2018. PhytoNet: comparative co-expression network analyses across phytoplankton and land plants. *Nucleic acids research*.
- Ficklin SP., Feltus FA. 2011. Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice. *Plant Physiology* 156:1244–1256.
- Finn RD., Attwood TK., Babbitt PC., Bateman A., Bork P., Bridge AJ., Chang H-Y., Dosztányi Z., El-Gebali S., Fraser M., Gough J., Haft D., Holliday GL., Huang H., Huang X., Letunic I., Lopez R., Lu S., Marchler-Bauer A., Mi H., Mistry J., Natale DA., Necci M., Nuka G., Orengo CA., Park Y., Pesceat S., Piovesan D., Potter SC., Rawlings ND., Redaschi N., Richardson L., Rivoire C., Sangrador-Vegas A., Sigrist C., Sillitoe I., Smithers B., Squizzato S., Sutton G., Thanki N., Thomas PD., Tosatto



- SCE., Wu CH., Xenarios I., Yeh L-S., Young S-Y., Mitchell AL. 2017. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research* 45:D190–D199.
- Freeman TC., Goldovsky L., Brosch M., van Dongen S., Mazière P., Grocock RJ., Freilich S., Thornton J., Enright AJ. 2007. Construction, Visualisation, and Clustering of Transcription Networks from Microarray Expression Data. *PLoS Computational Biology* 3:e206.
- Friedman N. 2004. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* 303:799. 8.
- Gautier L., Cope L., Bolstad BM., Irizarry RA. 2004. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307–315.
- Gibson SM., Ficklin SP., Isaacson S., Luo F., Feltus FA., Smith MC. 2013. Massive-scale gene co-expression network construction and robustness testing using random matrix theory. *PloS one* 8:e55871.
- Gillis J., Pavlidis P. 2011. The impact of multifunctional genes on "guilt by association" analysis. *PloS one* 6:e17258.
- Giorgi FM., Del Fabbro C., Licausi F. 2013. Comparative study of RNA-seq-and microarray-derived coexpression networks in *Arabidopsis thaliana*. *Bioinformatics* 29:717–724.
- Girvan M., Newman MEJ. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99:7821.
- Gomez-Cabrero D., Abugessaisa I., Maier D., Teschendorff A., Merkenschlager M., Gisel A., Ballestar E., Bongcam-Rudloff E., Conesa A., Tegnér J. 2014. Data integration in the era of omics: current and future challenges. *BMC Systems Biology* 8:I1.
- Goodwin S., McPherson JD., McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17:333.
- Guerin C., Joet T., Serret J., Lashermes P., Vaissayre V., Agbessi MD., Beule T., Severac D., Amblard P., Tregear J., others 2016. Gene coexpression network analysis of oil biosynthesis in an interspecific backcross of oil palm. *The Plant Journal* 87:423–441.



- Hansen BO., Meyer EH., Ferrari C., Vaid N., Movahedi S., Vandepoele K., Nikoloski Z., Mutwil M. 2018. Ensemble gene function prediction database reveals genes important for complex I formation in *Arabidopsis thaliana*. *New Phytologist* 217:1521–1534.
- Hartwell LH., Hopfield JJ., Leibler S., Murray AW. 1999. From molecular to modular cell biology. *Nature* 402:C47.
- Herschkowitz JI., Simin K., Weigman VJ., Mikaelian I., Usary J., Hu Z., Rasmussen KE., Jones LP., Assefnia S., Chandrasekharan S., Backlund MG., Yin Y., Khramtsov AI., Bastein R., Quackenbush J., Glazer RI., Brown PH., Green JE., Kopelovich L., Furth PA., Palazzo JP., Olopade OI., Bernard PS., Churchill GA., Van Dyke T., Perou CM. 2007. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology* 8:R76.
- Hirai MY., Sugiyama K., Sawada Y., Tohge T., Obayashi T., Suzuki A., Araki R., Sakurai N., Suzuki H., Aoki K., Goda H., Nishizawa OI., Shibata D., Saito K. 2007. Omics-based identification of <em>Arabidopsis</em> Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proceedings of the National Academy of Sciences* 104:6478.
- Hirschhorn JN., Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6:95–108.
- Hitzemann R., Bottomly D., Darakjian P., Walter N., Iancu O., Searles R., Wilmot B., McWeeney S. 2013. Genes, behavior and next-generation RNA sequencing. *Genes, brain, and behavior* 12:1–12.
- Horan K., Jang C., Bailey-Serres J., Mittler R., Shelton C., Harper JF., Zhu J-K., Cushman JC., Gollery M., Girke T. 2008. Annotating Genes of Known and Unknown Function by Large-Scale Coexpression Analysis. *Plant Physiology* 147:41–57.
- Horvath S., Dong J. 2008. Geometric interpretation of gene coexpression network analysis. *PLoS computational biology* 4:e1000117.
- Hughes TR., Marton MJ., Jones AR., Roberts CJ., Stoughton R., Armour CD., Bennett HA., Coffey E., Dai H., He YD., Kidd MJ., King AM., Meyer MR., Slade D., Lum PY.,



- Stepaniants SB., Shoemaker DD., Gachotte D., Chakraburty K., Simon J., Bard M., Friend SH. 2000. Functional Discovery via a Compendium of Expression Profiles. *Cell* 102:109–126.
- Hummon NP., Dereian P. 1989. Connectivity in a citation network: The development of DNA theory. *Social Networks* 11:39–63.
- Hwang I., Sheen J., Müller B. 2012. Cytokinin Signaling Networks. *Annual Review of Plant Biology* 63:353–380. DOI: 10.1146/annurev-arplant-042811-105503.
- Imperatorskaia akademia nauk (Russia) 1726. *Commentarii Academiae scientiarum imperialis Petropolitanae*. Petropolis, Typis Academiae.
- Irvine KM. 2018. BILL SHIPLEY. Cause and Correlation in Biology: A User’s Guide to Path Analysis, Structural Equations and Causal Inference with R, 2nd ed. United Kingdom: Cambridge University Press. *Biometrics* 74:779–780.
- Jeong H., Mason SP., Barabási A-L., Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* 411:41.
- Jordan JD., Landau EM., Iyengar R. 2000. Signaling Networks: The Origins of Cellular Multitasking. *Cell* 103:193–200.
- Joshi-Tope G., Gillespie M., Vastrik I., D’Eustachio P., Schmidt E., de Bono B., Jassal B., Gopinath GR., Wu GR., Matthews L., Lewis S., Birney E., Stein L. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* 33:D428–D432.
- Joyce AR., Palsson BØ. 2006. The model organism as a system: integrating “omics” data sets. *Nature Reviews Molecular Cell Biology* 7:198–210.
- Kaiser W., Huguet E., Casas J., Commin C., Giron D. 2010. Plant green-island phenotype induced by leaf-miners is mediated by bacterial symbionts. *Proceedings of the Royal Society B: Biological Sciences*.
- Kanehisa M., Sato Y., Kawashima M., Furumichi M., Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* 44:D457–D462.
- Karlebach G., Shamir R. 2008. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* 9:770.



- Kauffmann A., Gentleman R., Huber W. 2009. arrayQualityMetrics - A bioconductor paquet for quality assessment of microarray data. *Bioinformatics* 25:415–416.
- Keller EF. 2005. The century beyond the gene. *Journal of Biosciences* 30:3–10.
- Kliebenstein DJ., Kroymann J., Brown P., Figuth A., Pedersen D., Gershenson J., Mitchell-Olds T. 2001. Genetic Control of Natural Variation in Arabidopsis Glucosinolate Accumulation. *Plant Physiology* 126:811–825.
- Kogenaru S., Yan Q., Guo Y., Wang N. 2012. RNA-seq and microarray complement each other in transcriptome profiling. *BMC Genomics* 13:629–629. .
- Koo AJK., Chung HS., Kobayashi Y., Howe GA. 2006. Identification of a Peroxisomal Acyl-activating Enzyme Involved in the Biosynthesis of Jasmonic Acid in Arabidopsis. *Journal of Biological Chemistry* 281:33511–33520.
- Krause J., Lusseau D., James R. 2009. Animal social networks: an introduction. *Behavioral Ecology and Sociobiology* 63:967–973.
- Kumari S., Nie J., Chen H-S., Ma H., Stewart R., Li X., Lu M-Z., Taylor WM., Wei H. 2012. Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery. *PLOS ONE* 7:1–17.
- Lamesch P., Berardini TZ., Li D., Swarbreck D., Wilks C., Sasidharan R., Muller R., Dreher K., Alexander DL., Garcia-Hernandez M., Karthikeyan AS., Lee CH., Nelson WD., Ploetz L., Singh S., Wensel A., Huala E. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* 40:D1202–D1210.
- Langfelder P., Horvath S. 2008. WGCNA: an R paquet for weighted correlation network analysis. *BMC Bioinformatics* 9:559.
- Langfelder P., Luo R., Oldham MC., Horvath S. 2011. Is my network module preserved and reproducible? *PLoS Computational Biology* 7.
- Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9:357–359.



- Larrañaga P., Calvo B., Santana R., Bielza C., Galdiano J., Inza I., Lozano JA., Armañanzas R., Santafé G., Pérez A., Robles V. 2006. Machine learning in bioinformatics. *Briefings in Bioinformatics* 7:86–112.
- Leal LG., López C., López-Kleine L. 2014. Construction and comparison of gene co-expression networks shows complex plant immune responses. *PeerJ* 2:e610.
- Lee HK., Hsu AK., Sajdak J., Qin J., Pavlidis P. 2004. Coexpression analysis of human genes across many microarray data sets. *Genome research* 14:1085–1094.
- Leek JT., Scharpf RB., Bravo HC., Simcha D., Langmead B., Johnson WE., Geman D., Baggerly K., Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* 11:733–739.
- Levin JZ., Yassour M., Adiconis X., Nusbaum C., Thompson DA., Friedman N., Gnirke A., Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods* 7:709–715.
- Li B., Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323–323.
- Li J., Lv R., Yang Z., Yang S., Mo H., Huang X. 2006. Construction of Small World Networks Based on K-Means Clustering Analysis. In: Alexandrov VN, van Albada GD, Sloot PMA, Dongarra J eds. *Computational Science – ICCS 2006*. Berlin, Heidelberg: Springer Berlin Heidelberg, 997–1000.
- Li Y., Pearl SA., Jackson SA. 2015. Gene networks in plant biology: approaches in reconstruction and analysis. *Trends in plant science* 20:664–675.
- Li B., Ruotti V., Stewart RM., Thomson JA., Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26:493–500.
- Liesecke F., Daudu D., Dugé de Bernonville R., Besseau S., Clastre M., Courdavault V., de Craene J-O., Crèche J., Giglioli-Guivarc'h N., Glévarec G., Pichon O., Dugé de Bernonville T. 2018. Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Scientific Reports* 8:10885. DOI: 10.1038/s41598-018-29077-3.



- Lim WK., Wang K., Lefebvre C., Califano A. 2007. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* 23:i282–i288.
- Lindlof A., Lubovac Z. 2005. Simulations of simple artificial genetic networks reveal features in the use of Relevance Networks. *In silico biology* 5:239–249.
- Liu R., Cheng Y., Yu J., Lv Q-L., Zhou H-H. 2015. Identification and validation of gene module associated with lung cancer through coexpression network analysis. *Gene* 563:56–62.
- López-Kleine L., Leal L., López C. 2013. Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. *Briefings in Functional Genomics* 12:457–467
- Love MI., Huber W., Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 15:550.
- Luo F., Yang Y., Zhong J., Gao H., Khan L., Thompson DK., Zhou J. 2007. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC bioinformatics* 8:299.
- Lynch M., Katju V. 2004. The altered evolutionary trajectories of gene duplicates. *Trends in Genetics* 20:544–549.
- Ma S., Gong Q., Bohnert HJ. 2007. An Arabidopsis gene network based on the graphical Gaussian model. *Genome research* 17:000–000.
- Ma C., Wang X. 2012. Application of the Gini Correlation Coefficient to Infer Regulatory Relationships in Transcriptome Analysis. *Plant Physiology* 160:192.
- Ma C., Zhang HH., Wang X. 2014. Machine learning for Big Data analytics in plants. *Trends in Plant Science* 19:798–808.
- MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 281–297.



- Mantione KJ., Kream RM., Kuzelova H., Ptacek R., Raboch J., Samuel JM., Stefano GB. 2014. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. *Medical Science Monitor Basic Research* 20:138–141.
- Mao L., Van Hemert JL., Dash S., Dickerson JA. 2009. Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics* 10:346–346.
- Margolin AA., Nemenman I., Basso K., Wiggins C., Stolovitzky G., Favera RD., Califano A. 2006. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* 7:S7–S7. DOI: 10.1186/1471-2105-7-S1-S7.
- Mortazavi A., Williams BA., McCue K., Schaeffer L., Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5:621.
- Mutwil M., Klie S., Tohge T., Giorgi FM., Wilkins O., Campbell MM., Fernie AR., Usadel B., Nikoloski Z., Persson S. 2011a. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell*:tpc–111.
- Mutwil M., Klie S., Tohge T., Giorgi FM., Wilkins O., Campbell MM., Fernie AR., Usadel B., Nikoloski Z., Persson S. 2011b. PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species. *The Plant Cell* 23:895–910.
- Mutwil M., Usadel B., Schutte M., Loraine A., Ebenhoh O., Persson S. 2010. Assembly of an Interactive Correlation Network for the Arabidopsis Genome Using a Novel Heuristic Clustering Algorithm. *Plant Physiology* 152:29–43.
- Nagarajan N., Pop M. 2013. Sequence assembly demystified. *Nature Reviews Genetics* 14:157.
- Narise T., Sakurai N., Obayashi T., Ohta H., Shibata D. 2017. Co-expressed Pathways DataBase for Tomato: a database to predict pathways relevant to a query gene. *BMC genomics* 18:437.



- Navarro Gallón SM., Elejalde-Palmett C., Daudu D., Liesecke F., Jullien F., Papon N., Dugé de Bernonville T., Courdavault V., Lanoue A., Oudin A., Glévarec G., Pichon O., Clastre M., St-Pierre B., Atehortùa L., Yoshikawa N., Giglioli-Guivarc'h N., Besseau S. 2017. Virus-induced gene silencing of the two squalene synthase isoforms of apple tree (*Malus × domestica* L.) negatively impacts phytosterol biosynthesis, plastid pigmentation and leaf growth. *Planta* 246:45–60.
- Nazarov PV., Muller A., Kaoma T., Nicot N., Maximo C., Birembaut P., Tran NL., Dittmar G., Vallar L. 2017. RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. *BMC Genomics* 18:443.
- Nerur S., Sikora R., Mangalaraj G., Balijepally V. 2005. Assessing the Relative Influence of Journals in a Citation Network. *Commun. ACM* 48:71–74.
- Newman MEJ. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103:8577–8582.
- Obayashi T., Aoki Y., Tadaka S., Kagaya Y., Kinoshita K. 2018. ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index. *Plant and Cell Physiology* 59:e3–e3.
- Obayashi T., Kinoshita K. 2009. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA research* 16:249–260.
- Obayashi T., Okamura Y., Ito S., Tadaka S., Aoki Y., Shirota M., Kinoshita K. 2014. ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. *Plant & cell physiology* 55:e6.
- Obayashi T., Okamura Y., Ito S., Tadaka S., Motoike IN., Kinoshita K. 2013. COXPRESdb: A database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Research* 41:1014–1020.
- Oldham MC., Horvath S., Geschwind DH. 2006. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences* 103:17973–17978.



- Ozsolak F., Milos PM. 2011. RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics* 12:87–98.
- Palsson BØ. 2015. Systems Biology by Bernhard Ø. Palsson. Available at [/core/books/systems-biology/7F8445BC87019806B3625DFC4B5C27D4](http://core/books/systems-biology/7F8445BC87019806B3625DFC4B5C27D4) (accessed August 31, 2018). DOI: 10.1017/CBO9781139854610.
- Patro R., Duggal G., Love MI., Irizarry RA., Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* 14:417–419.
- Patro R., Kingsford C. 2012. Global network alignment using multiscale spectral signatures. *Bioinformatics* 28:3105–3114.
- Pawson T., Saxton TM. 1999. Signaling Networks—Do All Roads Lead to the Same Genes? *Cell* 97:675–678.
- Persson S., Wei H., Milne J., Page GP., Somerville CR. 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences of the United States of America* 102:8633.
- Pertea M., Kim D., Pertea G., Leek JT., Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie, and Ballgown. *Nature protocols* 11:1650–1667.
- Phan HTT., Sternberg MJE. 2012. PINALOG: A novel approach to align protein interaction networks-implications for complex detection and function prediction. *Bioinformatics* 28:1239–1245.
- Priest HD., Fox SE., Rowley ER., Murray JR., Michael TP., Mockler TC. 2014. Analysis of Global Gene Expression in *Brachypodium distachyon* Reveals Extensive Network Plasticity in Response to Abiotic Stress. *PLOS ONE* 9:e87499..
- Priness I., Maimon O., Ben-Gal I. 2007. Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics* 8:111.
- Proost S., Mutwil M. 2018. CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic acids research*. 46(W1):W133-W140.



- Quackenbush J. 2002. Microarray data normalization and transformation. *Nature Genetics* 32:496.
- Rai A., Saito K., Yamazaki M. 2017. Integrated omics analysis of specialized metabolism in medicinal plants. *The Plant Journal* 90:764–787.
- Reshef DN., Reshef YA., Finucane HK., Grossman SR., McVean G., Turnbaugh PJ., Lander ES., Mitzenmacher M., Sabeti PC. 2011. Detecting Novel Associations in Large Data Sets. *Science* 334:1518–1524.
- Reverter A., Chan EKF. 2008. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24:2491–2497.
- Ritchie ME., Dunning MJ., Smith ML., Shi W., Lynch AG. 2011. BeadArray Expression Analysis Using Bioconductor. *PLoS Computational Biology* 7:e1002276.
- Robinson JL., Nielsen J. 2016. Integrative analysis of human omics data using biomolecular networks. *Mol. BioSyst.* 12:2953–2964.
- Rohart F., Gautier B., Singh A., Le Cao K-A. 2017. mixOmics: An R paquet for omics feature selection and multiple data integration. *PLoS computational biology* 13:e1005752.
- Ruiz-Sola MÁ., Coman D., Beck G., Barja MV., Colinas M., Graf A., Welsch R., Rütimann P., Bühlmann P., Bigler L., others 2016. Arabidopsis GERANYLGERANYL DIPHOSPHATE SYNTHASE 11 is a hub isozyme required for the production of most photosynthesis-related isoprenoids. *New Phytologist* 209:252–264.
- Ruprecht C., Proost S., Hernandez-Coronado M., Ortiz-Ramirez C., Lang D., Rensing SA., Becker JD., Vandepoele K., Mutwil M. 2017. Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *The Plant Journal* 90:447–465.
- Schäfer J., Strimmer K. 2005. Learning Large-Scale Graphical Gaussian Models from Genomic Data. In: *AIP Conference Proceedings*. AIP, 263–276.
- Schena M., Shalon D., Davis RW., Brown PO. 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* 270:467.



- Shibata N., Kajikawa Y., Takeda Y., Sakata I., Matsushima K. 2011. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting and Social Change* 78:274–282.
- Sims D., Sudbery I., Ilott NE., Heger A., Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 15:121.
- de Siqueira Santos S., Takahashi DY., Nakata A., Fujita A. 2013. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in bioinformatics* 15:906–918.
- Sîrbu A., Kerr G., Crane M., Ruskin HJ. 2012. RNA-Seq vs Dual- and Single-Channel Microarray Data: Sensitivity Analysis for Differential Expression and Clustering. *PLoS ONE* 7:e50986.
- Smyth GK., Speed T. 2003. Normalization of cDNA microarray data. *Methods* 31:265–273.
- Song L., Langfelder P., Horvath S. 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* 13:328.
- Steuer R., Kurths J., Daub CO., Weise J., Selbig J. 2002. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* 18:S231–S240.
- Szklarczyk D., Morris JH., Cook H., Kuhn M., Wyder S., Simonovic M., Santos A., Doncheva NT., Roth A., Bork P., others 2017. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research* 45:D362–D368.
- Tian T., Liu Y., Yan H., You Q., Yi X., Du Z., Xu W., Su Z. 2017. agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic acids research* 45:W122–W129.
- Tsaparas P., Mariño-Ramírez L., Bodenreider O., Koonin EV., Jordan IK. 2006. Global similarity and local divergence in human and mouse gene co-expression networks. *BMC evolutionary biology* 6:70.



- Tzfadia O., Diels T., De Meyer S., Vandepoele K., Aharoni A., Van de Peer Y. 2015. CoExpNetViz: comparative co-expression networks construction and visualization tool. *Frontiers in plant science* 6.
- Usadel B., Obayashi T., Mutwil M., Giorgi FM., Bassel GW., Tanimoto M., Chow A., Steinhauser D., Persson S., Provart NJ. 2009. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment* 32:1633–1651.
- van Dam S., Craig T., de Magalhães JP. 2015. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Research* 43:D1124–D1132.
- Venter JC., Adams MD., Myers EW., Li PW., Mural RJ., Sutton GG., Smith HO., et al. 2001. The Sequence of the Human Genome. *Science* 291:1304–1351.
- Verleyen W., Ballouz S., Gillis J. 2015. Measuring the wisdom of the crowds in network-based gene function inference. *Bioinformatics* 31:745–752.
- Vijesh N., Chakrabarti SK., Sreekumar J. 2013. Modeling of gene regulatory networks: A review. *Journal of Biomedical Science and Engineering* 06:223.
- Wagner CS., Horlings E., Whetsell TA., Mattsson P., Nordqvist K. 2015. Do Nobel Laureates Create Prize-Winning Networks? An Analysis of Collaborative Research in Physiology or Medicine. *PLOS ONE* 10:e0134164.
- Wagner GP., Kin K., Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* 131:281–285.
- Wang Z., Gerstein M., Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* 10:57–63.
- Wang P., Qi H., Song S., Li S., Huang N., Han W., Ma D. 2015. ImmuCo: a database of gene co-expression in immune cells. *Nucleic Acids Research* 43:D1133–D1139.
- Wang L., Wang S., Li W. 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28:2184–2185.



- Wei H., Persson S., Mehta T., Srinivasasainagendra V., Chen L., Page GP., Somerville C., Loraine A. 2006. Transcriptional Coordination of the Metabolic Network in Arabidopsis. *Plant Physiology* 142:762–774.
- Wisecaver JH., Borowsky AT., Tzin V., Jander G., Kliebenstein DJ., Rokas A. 2017. A global co-expression network approach for connecting genes to specialized metabolic pathways in plants. *The Plant Cell Online*:tpc–00009.
- Wolfe CJ., Kohane IS., Butte AJ. 2005. Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC Bioinformatics*:10.
- Woo Y., Affourtit J., Daigle S., Viale A., Johnson K., Naggert J., Churchill G. 2004. A Comparison of cDNA, Oligonucleotide, and Affymetrix GeneChip Gene Expression Microarray Platforms. *Journal of Biomolecular Techniques : JBT* 15:276–284.
- Yang S., Kim CY., Hwang S., Kim E., Kim H., Shim H., Lee I. 2017. COEXPEDIA: exploring biomedical hypotheses via co-expressions associated with medical subject headings (MeSH). *Nucleic Acids Research* 45:D389–D396.
- Yim WC., Yu Y., Song K., Jang CS., Lee B-M. 2013. PLANEX: the plant co-expression database. *BMC Plant Biology* 13:83–83.
- Yu H., Jiao B., Lu L., Wang P., Chen S., Liang C., Liu W. 2018. NetMiner-an ensemble pipeline for building genome-wide and high-quality gene co-expression network using massive-scale RNA-seq samples. *PloS one* 13:e0192613.
- Zhang H., Bernonville TD de., Body M., Glevarec G., Reichelt M., Unsicker S., Bruneau M., Renou J-P., Huguet E., Dubreuil G., Giron D. 2016. Leaf-mining by Phyllonorycter blancardella reprograms the host-leaf transcriptome to modulate phytohormones associated with nutrient mobilization and plant defense. *Journal of Insect Physiology* 84:114–127.



## Annexes

### BBV Interactome

From	To	Research	Teaching	R+T	Interaction
Nathalie Guivarch	Olivier Pichon	1	1	1	1
Nathalie Guivarch	Eric Ducos	1	1	1	1
Nathalie Guivarch	Vincent Courdavault	1	1	1	1
Nathalie Guivarch	Arnaud Lanoue	1	0	0	1
Nathalie Guivarch	Kostas Koudounas	1	0	0	1
Nathalie Guivarch	Ines Carqueijeiro	1	0	0	1
Nathalie Guivarch	Gaelle Glevarec	0	0	0	0
Nathalie Guivarch	Joel Crèche	0	0	0	0
Nathalie Guivarch	Benoit Saint Pierre	0	1	0	1
Nathalie Guivarch	Martine Courtois	0	0	0	0
Nathalie Guivarch	François Friocourt	0	0	0	0
Nathalie Guivarch	Céline Melin	1	0	0	1
Olivier Pichon	Eric Ducos	1	1	1	1
Olivier Pichon	Vincent Courdavault	0	1	0	1
Olivier Pichon	Arnaud Lanoue	0	0	0	0
Olivier Pichon	Kostas Koudounas	0	0	0	0
Olivier Pichon	Pamela Cruz	0	0	0	0
Olivier Pichon	Sébastien Besseau	0	1	0	1
Olivier Pichon	Kevin Billet	0	0	0	0
Olivier Pichon	Thibault Munsch	0	0	0	0
Olivier Pichon	Martine Courtois	0	0	0	0

Name	Science	Theme
Nathalie Guivarch	1	All
Olivier Pichon	1	PP
Eric Ducos	1	PP
Vincent Courdavault	1	MIA
Arnaud Lanoue	0	Still
Kostas Koudounas	1	MIA
Ines Carqueijeiro	1	MIA

**Extrait des tables pour la construction des réseaux du laboratoire BBV**





ORIGINAL ARTICLE

# Virus-induced gene silencing of the two squalene synthase isoforms of apple tree (*Malus × domestica* L.) negatively impacts phytosterol biosynthesis, plastid pigmentation and leaf growth

Sandra M. Navarro Gallón<sup>1,2</sup> · Carolina Elejalde-Palmett<sup>1</sup> · Dimitri Daudu<sup>1</sup> ·  
Franziska Liesecke<sup>1</sup> · Frédéric Jullien<sup>3</sup> · Nicolas Papon<sup>4</sup> · Thomas Dugé de Bernonville<sup>1</sup> ·  
Vincent Courdavault<sup>1</sup> · Arnaud Lanoue<sup>1</sup> · Audrey Oudin<sup>1</sup> · Gaëlle Glévarec<sup>1</sup> ·  
Olivier Pichon<sup>1</sup> · Marc Clastre<sup>1</sup> · Benoit St-Pierre<sup>1</sup> · Lucia Atehortúa<sup>2</sup> ·  
Nobuyuki Yoshikawa<sup>5</sup> · Nathalie Giglioli-Guivarc'h<sup>1</sup> · Sébastien Besseau<sup>1</sup>

Received: 20 January 2017 / Accepted: 17 March 2017 / Published online: 27 March 2017  
© Springer-Verlag Berlin Heidelberg 2017

## Abstract

**Main conclusion** The use of a VIGS approach to silence the newly characterized apple tree SQS isoforms points out the biological function of phytosterols in plastid pigmentation and leaf development.

Triterpenoids are beneficial health compounds highly accumulated in apple; however, their metabolic regulation is poorly understood. Squalene synthase (SQS) is a key branch point enzyme involved in both phytosterol and triterpene biosynthesis. In this study, two SQS isoforms were identified in apple tree genome. Both isoforms are located at the endoplasmic reticulum surface and were demonstrated to be functional SQS enzymes using an in vitro activity assay.

**Electronic supplementary material** The online version of this article (doi:[10.1007/s00425-017-2681-0](https://doi.org/10.1007/s00425-017-2681-0)) contains supplementary material, which is available to authorized users.

Sandra M. Navarro Gallón and Carolina Elejalde-Palmett contributed equally to this work.

✉ Sébastien Besseau  
sebastien.besseau@univ-tours.fr

<sup>1</sup> EA2106 Biomolécules et Biotechnologies Végétales, Université François Rabelais de Tours, Tours, France

<sup>2</sup> Laboratorio de Biotecnología, Sede de Investigacion Universitaria, Universidad de Antioquia, Medellin, Colombia

<sup>3</sup> EA3061 Laboratoire de Biotechnologies Végétales appliquées aux plantes aromatiques et médicinales, Université Jean Monnet de Saint Etienne, Saint Etienne, France

<sup>4</sup> EA3142 Groupe d'Etude des Interactions Hôte-Pathogène, Université d'Angers, Angers, France

<sup>5</sup> Plant Pathology Laboratory, Iwate University, Morioka, Japan

MdSQS1 and MdSQS2 display specificities in their expression profiles with respect to plant organs and environmental constraints. This indicates a possible preferential involvement of each isoform in phytosterol and/or triterpene metabolic pathways as further argued using RNAseq metatranscriptomic analyses. Finally, a virus-induced gene silencing (VIGS) approach was used to silence MdSQS1 and MdSQS2. The concomitant down-regulation of both MdSQS isoforms strongly affected phytosterol synthesis without alteration in triterpene accumulation, since triterpene-specific oxidosqualene synthases were found to be up-regulated to compensate metabolic flux reduction. Phytosterol deficiencies in silenced plants clearly disturbed chloroplast pigmentation and led to abnormal development impacting leaf division rather than elongation or differentiation. In conclusion, beyond the characterization of two SQS isoforms in apple tree, this work brings clues for a specific involvement of each isoform in phytosterol and triterpene pathways and emphasizes the biological function of phytosterols in development and chloroplast integrity. Our report also opens the door to metabolism studies in *Malus domestica* using the apple latent spherical virus-based VIGS method.

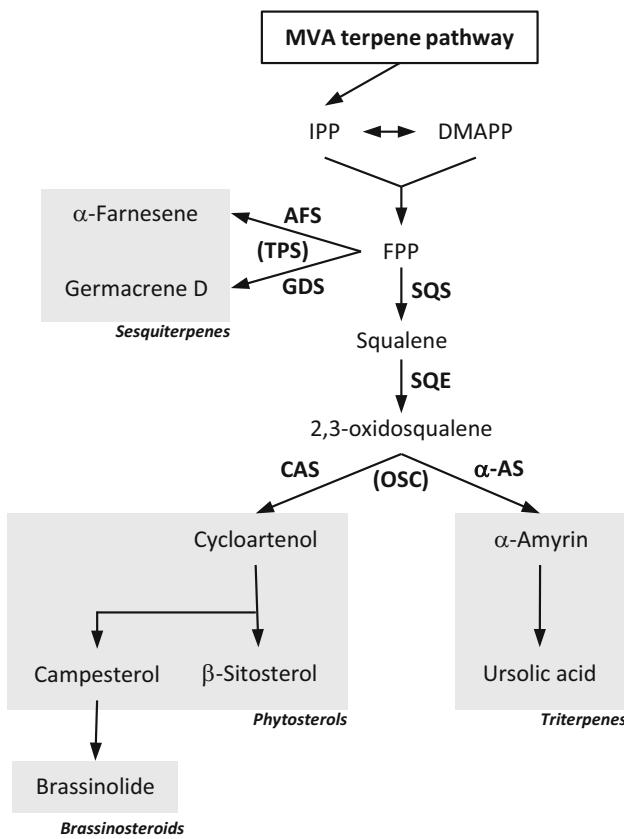
**Keywords** Apple latent spherical virus (ALSV) · Apple tree · Phytosterol · Squalene synthase · Triterpene · VIGS

## Abbreviations

ALSV	Apple latent spherical virus
CAS	Cycloartenol synthase
SQS	Squalene synthase
FPP	Farnesyl diphosphate
OSC	Oxidosqualene cyclase
VIGS	Virus-induced gene silencing

## Introduction

Squalene synthase (SQS) catalyzes the first enzymatic step of triterpenoid biosynthesis which are lipophilic C<sub>30</sub>-derivative terpenes, ubiquitously produced in animals, fungus and plants. SQS condenses two farnesyl diphosphates (FPP) to produce squalene. This reaction is followed by an oxidative step catalyzed by the squalene epoxidase (SQE) yielding 2,3 oxidosqualene. This reactive intermediate is further diversely cyclized by a large family of oxidosqualene cyclases (OSC) (Fig. 1). Depending on specific carbocation rearrangements catalyzed by OSC, wide series of triterpenoid skeletons can be generated. In plants, the most widespread triterpenoid forms are phytosterols (i.e., tetracyclic triterpenoids) and triterpenes (i.e., pentacyclic triterpenoids). Sterols are produced in all eukaryotic organisms and are currently considered as essential compounds. Their synthesis involved OSC that



**Fig. 1** Simplified mevalonate-derived terpene biosynthesis pathway leading to sesquiterpenes, triterpenes, phytosterols and brassinosteroids. Main compounds found to accumulate in apple tree are indicated for each terpene family. Enzymes found in several plants to be involved in primary steps are mentioned. *IPP* isopentenyl pyrophosphate, *DMAPP* dimethylallyl pyrophosphate, *FPP* farnesyl pyrophosphate, *SQS* squalene synthase, *SQE* squalene epoxidase, *AFS* α-farnesene synthase, *GDS* germacrene D synthase, *CAS* cycloartenol synthase, *α-AS* α-amyrin synthase, *TPS*s terpene synthases (e.g., *GDS* and *AFS*), *OSC* oxidosqualene cyclases (e.g., *CAS* and *α-AS*)

cyclize 2,3 oxidosqualene in a chair-boat-chair conformation with a protosteryl cation intermediate (Phillips et al. 2006). In mammals and fungi, lanosterol synthase (LAS) (Shi et al. 1994; Baker et al. 1995) leads to lanosterol, further converted into cholesterol and ergosterol as main compounds, respectively, whereas in plants, cycloartenol synthase (CAS) (Corey et al. 1993) is the entry point for the synthesis of a mixture of sterols, the so-called phytosterols (Fig. 1), composed of β-sitosterol, stigmasterol and campesterol as predominant forms (Benveniste 2002). However, the existence of an additional minor branch to produce cholesterol in some plants can be noticed (Kolesnikova et al. 2006; Suzuki et al. 2006; Sonawane et al. 2016). Sterols are important membrane structural components, impacting fluidity and impermeability (Hartmann 1998). Beside this function, sterols participate in plant development and defense regulation (Lindsey et al. 2003) notably through brassinosteroids which are plant-specific steroid hormones produced from campesterol (Wang et al. 2014) (Fig. 1). Triterpenes are a large class of plant natural products. Their synthesis also involves distinct OSC cyclizing 2,3 oxidosqualene in a chair-chair-chair conformation with a dammarenyl cation intermediate (Phillips et al. 2006). β-Amyrin synthase (β-AS) (Kushiro et al. 1998), α-amyrin synthase (α-AS) (Brendolise et al. 2011) and lupeol synthase (LUP) (Herrera et al. 1998) are the main representative triterpene synthases leading, respectively, to the common oleanane, ursane and lupane-type triterpenes (Fig. 1). These compounds belong to specialized membranes and are usually reported in higher amounts in cuticular waxes (Bringe et al. 2006; Jäger et al. 2009; Wang et al. 2011; Racovita and Jetter 2016), providing protective function against insect and microbial attacks (González-Coloma et al. 2011). Finally, triterpenes can be glycosylated and referred to as saponins (Vincken et al. 2007).

SQS are endoplasmic reticulum (ER) bound enzymes with a cytosolic catalytic domain catalyzing the synthesis of squalene in a two consecutive steps reaction. First, presqualene diphosphate is generated by the condensation of two FPP, followed by a reduction reaction implying NADPH as co-factor. SQS genes were first cloned and studied in yeast (Jennings et al. 1991) and mammals (McKenzie et al. 1992; Summers et al. 1993). Thereafter, the enzyme was characterized in plants revealing the conservation of the three catalytic domains (A, B and C-domains) previously identified in rat by mutagenesis studies (Gu et al. 1998) and in human through crystal structure determination (Pandit et al. 2000). However, SQS greatly differs between kingdoms in their C-terminal end which contains the transmembrane domain allowing ER anchoring. In higher plants, SQS cDNA was cloned and studied in plethora of plants; however, information on gene

copy numbers is patchy. To date, *SQS* gene was reported either as single copy gene like in mammals and yeast or as a small gene family, as well. A single *SQS* gene was identified in genomes of rice (Hata et al. 1997), *Taxus cuspidata* (Huang et al. 2007) and *Euphorbia tirucallii* (Uchida et al. 2009). Two genes were found in tobacco (Devarenne et al. 1998), *Glycyrrhiza glabra* (Hayashi et al. 1999) *Medicago truncatula* (Suzuki et al. 2002) and *Arabidopsis thaliana* albeit one seems to be a pseudogene (Busquets et al. 2008). Finally, in *Panax ginseng* up to three *SQS* copies were functionally characterized (Kim et al. 2011). *SQS* constitutes a rate-limiting enzyme for triterpenoid synthesis as demonstrated by sterols and triterpene saponins accumulation in *SQS*-overexpressed lines of *P. ginseng* and *Eleutherococcus senticosus* (Lee et al. 2004; Seo et al. 2005). Therefore, regulation of *SQS* ensures a fine-tuned control of metabolic fluxes through the triterpenoid metabolism; however, the function of *SQS* for squalene distribution in subsequent branches of the pathway remains unclear. For instance, a fungal elicitor treatment on parsley and tobacco cell cultures led to a concomitant decrease in *SQS* activity and sterol biosynthesis, allowing FPP substrates to be rerouted to sesquiterpene synthesis as plant defense response metabolites (Vögeli and Chappell 1988; Haudenschild and Hartmann 1995; Devarenne et al. 1998). Interestingly, the regulation of *SQS* activity was observed with only minor transcript variations suggesting an important post-translational regulation of the enzyme (Devarenne et al. 1998). Finally, additional regulations occur independently between phytosterol and triterpene metabolisms at several other enzymatic steps along the pathways (Schaller 2004).

Consumption of apple has been proven beneficial to human health. The apple tree polyphenol metabolism mainly participates to this property by ensuring a significant production of dihydrochalcones (Dugé de Bernonville et al. 2010) and flavonoids (Eberhardt et al. 2000). However, triterpenoids produced in apple also present interesting nutritional and pharmacologic virtues. For instance, apple tree triterpene extracts were found to display radical scavenging activities and an antiproliferative potential against tumor cells (D'Abrosca et al. 2005; Ma et al. 2005; He and Rui 2007). The availability of the apple tree genome sequence (Velasco et al. 2010) has progressively promoted the identification of enzymes involved in the biosynthesis of triterpenes and sesquiterpenes. Up to now, ten terpene synthases (TPS, Fig. 1) and five OSC (Fig. 1) were characterized in this species (Pechous and Whitaker 2004; Brendolise et al. 2011; Nieuwenhuizen et al. 2013; Andre et al. 2016). Among them, MdOSC1 and MdOSC3 were shown to be  $\alpha$ -amyrin synthase, allowing the accumulation of ursolic acid as the main apple tree triterpene (Brendolise et al. 2011). In a

continuing effort to elucidate the apple tree sterol and triterpene biosynthetic pathways, we were herein particularly interested in identifying and characterizing *SQS* genes as well as deciphering their physiological roles in planta. Two functional and differentially expressed *SQS* isoforms were identified. By taking advantage of a virus-induced gene silencing (VIGS) method, we demonstrated the crucial role of *SQS* in phytosterol synthesis and their importance in leaf development and plastid pigmentation in apple tree.

## Materials and methods

### Plant material and growth conditions

Seeds of *Malus domestica* cultivar ‘Golden Delicious’ (provided by INRA d’Angers, France) were vernalized 3 months in wet vermiculite at 4 °C. Plants were grown in a greenhouse at 21 °C with 50% relative humidity and under a light/dark cycle of 14/10 h. Apple tree materials were collected on 2-month-old plants except for flowers collected on mature trees in orchards. Cold and osmotic stresses were performed on 5-week-old plants. For cold treatment, plants were grown at 4 °C with light cycle. To induce osmotic stress, soil was watered and saturated with 15% PEG-6000.

### Sequence alignment and phylogeny analysis

Previously characterized At*SQS1*, At*SQS2* (*A. thaliana*), Mo*SQS* (*Magnolia officinalis*), Nt*SQS* (*Nicotiana tabacum*), Pg*SQS1* (*P. ginseng*) and Lj*SQS* (*Lotus japonicus*) sequences were used as queries for BLASTN searches in the apple tree genome (Velasco et al. 2010) using the phytozome database (Goodstein et al. 2012). Corresponding protein sequences were aligned with ClustalX (v2.0.11) and the resulting alignment was used to construct a neighbor-joining phylogenetic tree with bootstrap non-parametric resampling procedure.

### Isolation of *SQS* genes

Md*SQS1* and Md*SQS2* ORFs were amplified with Phusion high-fidelity DNA polymerase (Fermentas) from leaf cDNA using the same couple of primers (MdSQSFwd and MdSQSRev; Suppl. Table S1). PCR products were further cloned in pGEM-T easy vector (Promega). Sequencing of positive clones discriminated Md*SQS1* and Md*SQS2* sequences which were submitted in NCBI database (accession numbers KC895979 and KC895980). *Saccharomyces cerevisiae ERG9* (Sc*ERG9*) was cloned in pGEM-T easy using yeast gDNA and specific primers

(ScERG9Fwd and ScERG9Rev; Suppl. Table S1). Finally, chimeric *MdSQS1<sup>ScTM</sup>* and *MdSQS2<sup>ScTM</sup>*, in which the C-terminal region of apple tree SQS (residues 316–412 and 316–410, respectively, for *MdSQS1* and *MdSQS2*) was replaced by the equivalent region of ScERG9 (residues 326–444), were generated by PCR overlap extension method (Ho et al. 1989) and cloned in pGEM-T easy vector. PCRs were performed using external primer (*MdSQSFwd* or *ScERG9Rev*) and overlap internal primer (*MdSQSwoTMRev* or *ScERG9TMFwd*; Suppl. Table S1).

### Yeast complementation experiments

For yeast complementation assays, *MdSQS1*, *MdSQS2*, *MdSQS1<sup>ScTM</sup>*, *MdSQS2<sup>ScTM</sup>* and *ScERG9* were amplified from sequences sub-cloned in pGEM-T easy vector, using specific primers containing BamHI and XhoI restriction sites (Suppl. Table S1) and cloned in pYES2.0 yeast expression vector under the inducible *GAL1* gene promoter. Plasmids were integrated in the *S. cerevisiae erg9* mutant strain 2C1 (provided by Marc Fischer) following a standard electroporation procedure (Becker and Guarente 1991). Cells were selected on CSM-URA plates (0.8% yeast nitrogen base w/o amino acids, 2% glucose and dropout mix-URA) complemented with 80 µg/ml ergosterol (in tergitol NP-40/ethanol, 1:1, v/v). After 5 days at 30 °C, transformed yeasts were plated on YPGal medium (2% bacto peptone, 1% yeast extract and 2% galactose) to induce recombinant protein expression in strain 2C1 (Kribii et al. 1997).

### Enzyme assay

*Saccharomyces cerevisiae erg9* mutant strain 2C1 harboring expressing vectors were grown in liquid medium complemented with 4 µg/ml ergosterol, 2 days at 30 °C in CSM-URA and followed by an additional 12 h culture at 30 °C in YPGal medium to induce recombinant protein expression. Cells were harvested and mechanically lysed with acid-washed glass beads in 50 mM Tris–HCl buffer, pH 7.5, containing 1 mM EDTA, 600 mM sorbitol, 1 mM DTT, 1 mM PMSF. Crude extracts were directly used for enzymatic assays. The SQS activity was measured in 100 µl volume reaction containing 50 mM Tris–HCl buffer (pH 7.5), 5 mM MgCl<sub>2</sub>, 2 mM β-mercaptoethanol, 50 µM FPP and 3 mM NADPH. After 30 min at 35 °C, reaction products were extracted three times with two volumes of hexane and concentrated under nitrogen flux. Samples were analyzed by TLC on silica gel 60, developed in hexane/chloroform (9:1, v/v) and colored with iodine vapor. Authentic squalene was used as standard (Sigma-Aldrich).

### Subcellular localization studies

The full-length coding sequences of *MdSQS1* and *MdSQS2* were amplified with specific primers containing SpeI restriction sites (Suppl. Table S1) and cloned in the pSCA-YFP vector in frame with the 3'-end of the YFP coding sequence. *Catharanthus roseus* C20D cells were transiently transformed with the resulting pSCA-YFP constructs in combination with the endoplasmic reticulum (ER)-CFP marker (Nelson et al. 2007), using particle bombardment as described by (Guirimand et al. 2009). Briefly, *C. roseus* plated cells were bombarded with DNA-coated gold particles (1 µm) and 1100 psi rupture disc at a stopping-screen-to-target distance of 6 cm, using the Bio-Rad PDS1000/He system. Cells were cultivated for 16–38 h before being harvested and observed. The sub-cellular localization was determined using an Olympus BX-51 epifluorescence microscope equipped with an Olympus DP-71 digital camera and a combination of YFP and CFP filters. The pattern of localization presented in this work is representative of circa 50 observed cells.

### Gene expression analysis

Extractions of total RNA were performed using the NucleoSpin RNA kit (Macherey-Nagel), with an improved extraction protocol (Elejalde-Palmett et al. 2015). First-strand cDNA were synthesized from 1 µg of total RNA using oligo(dT)18 and RevertAid Reverse Transcriptase (Fermentas) according to the manufacturer's instructions. Quantitative PCR analysis were performed on a CFX96 Touch Real-Time PCR System (Bio-Rad). Each PCR of 15 µl contains equal amount of cDNA, 1× DyNAmo SYBR Green mix (Fermentas) and specific primers for each genes, except for *MdSQS1* and *MdSQS2* transcripts which were commonly amplified (Suppl. Table S1). Amplification was initiated by a denaturation step at 95 °C for 10 min followed by 40 cycles at 95 °C for 15 s and 60 °C for 30 s. Melting curves were used to determine the specificity of reactions. Calibration curves were made for each couple of primers to ensure PCR efficiency and performed absolute quantification when needed. Otherwise, relative quantification of gene expression was calculated according to the ΔΔC<sub>t</sub> method. *MdEF1alpha* was used as reference gene. Each assay was performed in duplicate and expression measurements were achieved at least twice with independent biological repetitions. Since *MdSQS1* and *MdSQS2* transcripts cannot be distinguished with the primers used for qPCR, expression ratio of each transcript among total *SQS* transcripts amplified was measured (Suppl. Fig. S1). For this purpose, cDNA were pre-amplified using Phusion DNA polymerase and *MdSQS* primers during six cycles. Pre-amplicons were divided into four equivalent tubes and subjected to digestion with *Dra*I, *Pst*I, both or none (fast digest enzymes;

Fermentas). Further qPCR was done with MdSQS primers in classical conditions. *Dra*I restriction site was found in MdSQS1 amplicon but not for MdSQS2, and inversely for *Psi*I. Consequently, *SQS* transcript ratios were calculated with *SQS* copy numbers obtained in single digested pre-amplicons compared to undigested and normalized with MdEF1alpha.

### RNA-seq data analysis

A total of 417 raw RNAseq data for *M. domestica* were downloaded from EBI as fastq files (Suppl. Table S2). Gene expression was quantified by pseudo-aligning reads on *M. domestica* reference transcripts v3.0 with Salmon v0.7.2 (Patro et al. 2016) using the variational Bayesian Expectation–Maximization algorithm and sequence bias correction to improve quantification. A total of 310 samples where more than 50% of the reads pseudo-aligned were kept. Linear correlations between transcript expression levels (TPM, transcript per million) were calculated with the Pearson Correlation Coefficient, which were subsequently used to determine the Highest Reciprocal Ranks (HRR) between each pair of transcripts (Mutwil et al. 2011). This ranking procedure is expected to be more integrative than the two-dimension PCC as it takes into account the PCC of all other genes for each pair. Transcripts were considered to be co-expressed with MdSQS1 or MdSQS2 when the HRR was below 2,000 (Suppl. Table S3).

Annotation of *M. domestica* transcripts was performed as described in Zhang et al. (2015) with the Trinotate pipeline. Candidate triterpene and phytosterol-related genes were obtained from the Gene Ontology (GO) annotation associated with Uniprot entries using “triterpenoid” and “steroid biosynthetic process” search terms (Suppl. Table S4). Average expression profiles of candidate genes from the two classes were obtained over the 310 samples and visualized within R and the ggplot2 package (R Core Team 2015). To improve readability, expression profiles were smoothed with a generalized additive model including a 99% confidence interval.

### VIGS experiments

Virus-induced gene silencing technology was used to induce *SQS* gene silencing in apple tree with the apple latent spherical virus (ALSV) system (Sasaki et al. 2011). A 201pb fragment of MdSQS1 and *rbcS* coding sequence were amplified (Suppl. Fig. S2 and Suppl. Table S1) and cloned into *Xho*I and *Bam*HI sites of vector pEALSR2L5R5. The resulting plasmids, combined with pEALSR1, were used to produce virus in *Chenopodium quinoa*. Virus RNA were used for apple tree seedlings transformation by biolistic as described by (Sasaki et al. 2011).

### Histochemical analysis

Apple tree leaves were fixed with a solution containing 3.7% formaldehyde, 50% ethanol, and 5% acetic acid in water, dehydrated in an ethanol series, transferred to terbutanol, and gradually embedded in paraplast. Transverse sections (10 µm thick) were made with a Leica RM2125 RTS microtome. Paraplast was removed with xylene; leaf cross sections were then rehydrated, stained with 0.5% astra blue for 5 min, rinsed, dehydrated in ethanol series, and mounted on slides with Eukitt. Observations were done with an Olympus BX51 microscope. Cell size was measured with Cell D software.

### Terpene extraction and analysis

Sesquiterpenes of apple leaves were extracted by hexane (4 ml/gFW) with camphor as internal standard (1 mg/l). Injection volume was 2 µl in split 1:2 mode on an Agilent GC 6850 gas chromatograph coupled with an Agilent 5973 ion trap mass detector. The instrument was equipped with a 30 m × 0.25 mm DB5 apolar capillary column. Temperatures of injector and detector were 250 °C. Helium was used as the carrier gas at a flow rate of 1.0 ml/min. Oven temperature settings were: 60 °C at injection followed by a 3 °C/min temperature ramp from 60 to 150 °C, then 7 °C/min from 150 to 240 °C. Temperature was then kept on hold at 240 °C for 5 min. Molecule identification was performed using Wiley, NIST 05 and Adams mass spectra databases (Adams 2007). Quantification of α-farnesene was performed using a pure α-farnesene standard (Payan Bertrand, Grasse, France).

For sterol analysis, around 50 mg DW samples were ground in nitrogen with mortar and pestle. Lipids were extracted and saponified at 80 °C with 15 ml of 6% KOH in methanol for 3 h. Cholesterol was added as internal standard (1.5 mg/gDW). Then, 5 ml of water is added to the mixture and sterols were extracted with three volumes of hexane. Organic layers were dried using a rotavapor at 30 °C and acylation reaction was performed overnight at room temperature, on the dried residue dissolved in toluene, using pyridine/acetic anhydride (6:4, v/v). Steryl acetates were isolated on a preparative thin-layer chromatography (PLC silica gel 60 0.5 mm, Merck) developed in dichloromethane. Phytosterol analyses were performed on Thermo TRACE GC ultra coupled with a flame ionization detector and equipped with a 30 m × 0.25 mm DB5 apolar capillary column. Injector temperature was at 220 °C and detector at 300 °C. Nitrogen was used as the carrier gas at a flow rate of 1.0 ml/min. Oven temperature settings were: 60 °C at injection, maintained 50 s and followed by a 30 °C/min temperature ramp from 60 to 200 °C, then 15 °C/min from 200 to 300 °C. Temperature

was then kept on hold at 300 °C for 30 min. Campesterol, stigmasterol and β-sitosterol were identified and quantified using pure standards (Extrasynthese). It can be noticed that apple tree naturally produce traces of cholesterol. The use of cholesterol as internal standard does not impact the quantification of main phytosterols, but prevents accurate quantification of endogenous cholesterol.

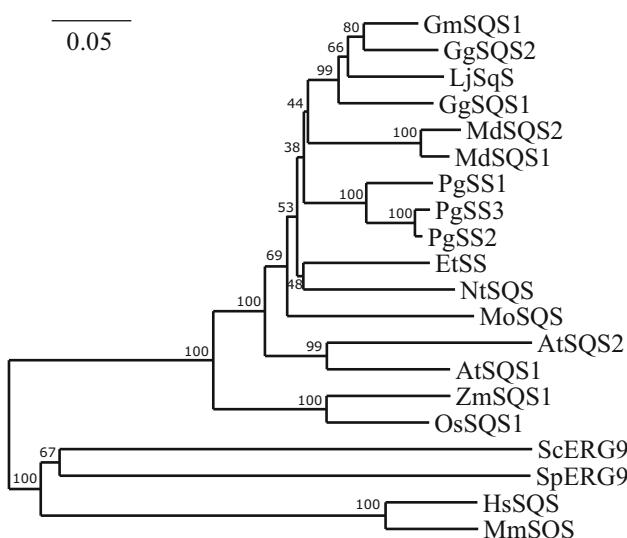
Triterpene content of around 20 mg of freeze-dried samples ground in nitrogen was extracted three times with 1 ml methanol, incubated for 1 h in a sonication bath. Glycyrhetic acid was used as internal standard (0.75 mg/gDW). Supernatants were combined, clarified at 13,000g for 5 min and concentrated in speed-vac. Ursolic acid was quantified as the main apple tree triterpene identified, using an HPLC system composed of a Waters 600 controller, an autosampler (Waters 717 plus) and a UV-visible photodiode array detector (Waters 996) controlled by Empower 2 software. Analyses were performed on a column packed with 3 μm particles (250 × 4 mm, Multospher 120 RP18HP; CS-Service, Langerwehe, Germany) at 24 °C. The separation was isocratically undertaken with a solvent consisting of 0.1% (v/v) aqueous phosphoric acid–acetonitrile (75:25, v/v) at a flow rate of 0.5 ml/min during 60 min. Quantification was performed using an ursolic acid pure standard (Extrasynthese).

## Results

### *Malus domestica* genome harbors a couple of SQS resulting from a recent gene duplication event

Apple tree genome was browsed to identify loci corresponding to putative *SQS*. BLAST searches were performed using sequences of several characterized plant SQS (Fig. 2). Two candidates were found at locus tags MDP0000309694 on chromosome 17 and MDP0000311820 on chromosome 10. Reannotation of the predicted genes was needed to improve sequence retrieval. Our alternative gene predictions were further confirmed by coding sequence (CDS) cloning from cDNA. These sequences of 1239 and 1233 nucleotides are registered in GenBank as *MdSQS1* and *MdSQS2* with the respective accession numbers KC895979 and KC895980. Both genes display 13 exons, and share 96% nucleotide identity and 94.5% at the protein level. As illustrated in the phylogenetic tree generated after alignment of protein sequences (Fig. 2), *MdSQSs* strongly clustered together compared to characterized SQSs of other species, which may indicate a recent gene duplication event in the apple tree genome. This analysis also highlighted that such duplication events are species dependent.

Both *MdSQS1* and *MdSQS2* sequences contain all highly conserved SQS domains (Suppl. Fig. S3): A and B



**Fig. 2** Phylogenetic tree of putative *M. domestica* SQS compared to a selection of previously characterized squalene synthases from various species. The tree was constructed by neighbor-joining distance analysis on protein sequences. Line lengths indicate the relative distances between nodes. *M. domestica* *MdSQS1* (KC895979) and *MdSQS2* (KC895980); *A. thaliana* *AtSQS1* (NP\_195190) and *AtSQS2* (NP\_195191); *M. officinalis* *MoSQS* (KT223496); *N. tabacum* *NtSQS* (AAB08578); *P. ginseng* *PgSS1* (AB010148), *PgSS2* (GQ468527) and *PgSS3* (AB115496); *L. japonicus* *LjSqS* (AB102688); *E. tirucalli* *EtSS* (AB433916); *G. glabra* *GgSQS1* (D86409) and *GgSQS2* (D86410); *Glycine max* *GmSQS1* (NP\_001236365); *Oryza sativa* *OsSQS1* (NM\_001058160); *Zea mays* *ZmSQS1* (NM\_001111369); *Saccharomyces pombe* *SpERG9* (NP\_595363); *S. cerevisiae* *ScERG9* (ACD03847); *Mus musculus* *MmSQS* (BAA06102); *Homo sapiens* *HsSQS* (CAA48896)

domains shared with phytoene synthase (Summers et al. 1993), an SQS-specific C domain which forms a hydrophobic tunnel interacting with FPP and containing Phe and Gln residues required for enzyme activity (Gu et al. 1998), a flap motif that protects the hydrophobic tunnel, two DxxxD motifs for magnesium dependent binding of NADPH co-factor and, finally, a D domain at C-terminal end, which includes a predicted transmembrane helix to anchor the enzyme to the membrane. Unsurprisingly, the D domains of *MdSQS1* and *MdSQS2* are the most divergent part of the protein compared to others SQS. As previously reported, D domains and transmembrane sequences are highly variable among SQSs from plants or animals and are also significantly shorter than in yeast (Suppl. Fig. S3).

### *MdSQS1* and *MdSQS2* are two active isoforms of SQS

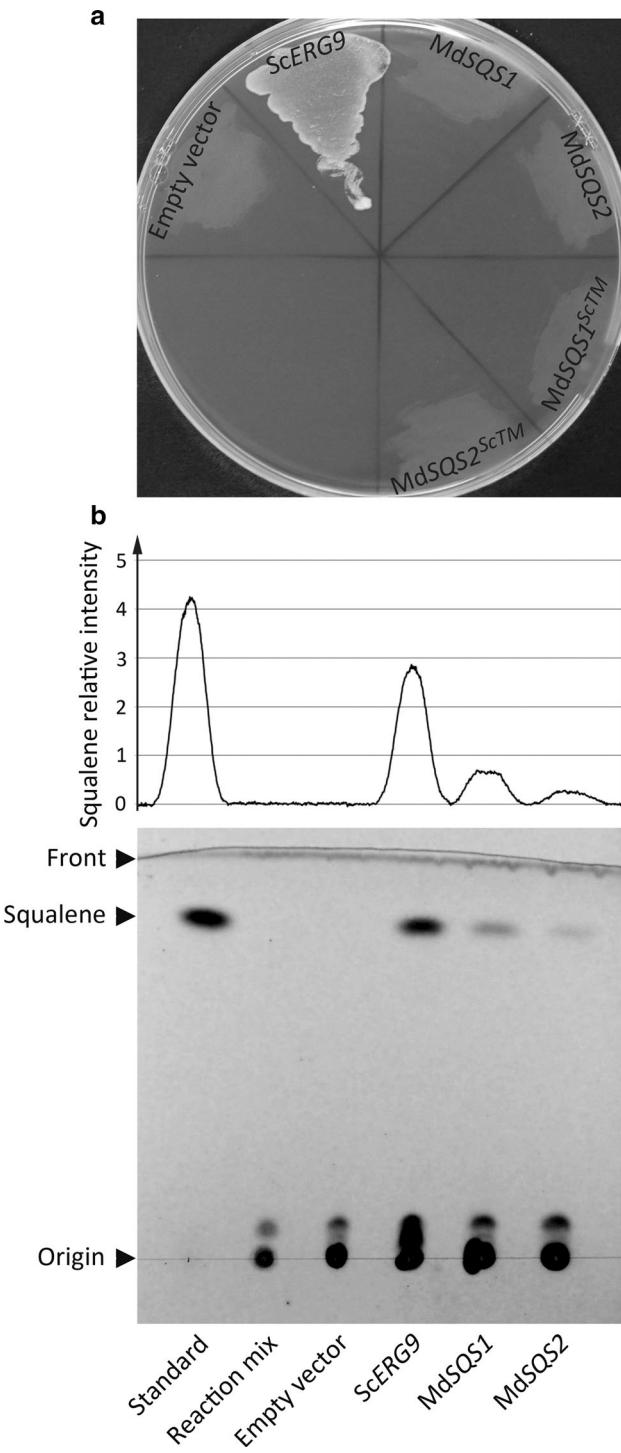
To assess the activity of *MdSQS1* and *MdSQS2*, both sequences were tested for yeast *erg9* mutant complementation (Kim et al. 2011). The *S. cerevisiae erg9* mutant strain (2C1) is deficient in SQS activity resulting in

ergosterol auxotrophy and cell death. To rescue cell growth, yeasts were transformed with the expression vector pYES2 harboring the full-length coding sequence of *MdSQS1* or *MdSQS2*, but also *ScERG9* as a positive control (Fig. 3a). In a first series of experiments, we were unable to complement the *erg9* mutation using either *MdSQS1* or *MdSQS2*. However, the absence of *erg9* complementation was previously observed for few plants and mammals yet active SQS enzymes (Robinson et al. 1993; Kribii et al. 1997). It was proposed that an interaction domain around the SQS transmembrane domain was interacting and channeling with downstream yeast enzymes from the pathway to ensure ergosterol synthesis. Indeed, the SQS transmembrane domain which is quite different between kingdoms as mentioned previously (Suppl. Fig. S3) may impact a metabolon implement. We thus replaced the apple tree SQS transmembrane domain by the equivalent region of *ScERG9*, but unfortunately the expression of the apple tree chimeric proteins (*MdSQS1<sup>ScTM</sup>* and *MdSQS2<sup>ScTM</sup>*) did not rescue the cell growth (Fig. 3a).

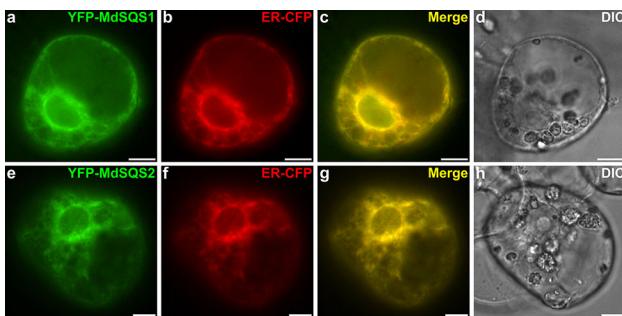
Although an appropriate membrane anchor is needed for successful yeast complementation, it does not impair enzyme activity in vitro (Kribii et al. 1997). Alternatively, yeast *erg9* mutant strains transformed to express *MdSQS1* or *MdSQS2* were selected and grown on ergosterol-complemented medium. By this way, SQS activity was tested in vitro directly on crude extract with farnesyl diphosphate and NADPH. Squalene production was monitored by TLC (Fig. 3b) and further confirmed by GC using authentic standards. While no squalene was detected on TLC for the reaction mix without enzyme and for reaction mix with crude extract of yeast transformed with empty vector, a clear signal was observed with yeast mutant strain complemented with the *ScERG9* expressing plasmid. More importantly, a substantial squalene synthesis was also obtained for crude extracts of yeast expressing *MdSQS1* or *MdSQS2*. However, *M. domestica* SQSs may only lead to low rates of squalene production in yeast, which is probably limiting for a sufficient ergosterol production for yeast growth. Nevertheless, in vitro activities clearly demonstrated the presence of two active SQS isoforms in apple tree.

### Subcellular localization of *MdSQS1* and *MdSQS2*

Squalene synthase is reported to function as ER-anchored enzymes. A transmembrane C-terminal hydrophobic  $\alpha$ -helix was found in the D domain of each characterized SQS but highly diverging in amino acid sequence between species (Suppl. Fig. S3). Therefore, the localization of apple tree SQS isoforms was investigated using an N-terminal yellow fluorescent protein (YFP) fusion protein



**Fig. 3** Characterization of *MdSQS1* and *MdSQS2* activities. **a** Functional complementation of *S. cerevisiae erg9* mutant strain 2C1, lacking SQS activity, by *ScERG9*, *MdSQS1*, *MDSQS2* or chimeric transmembrane variants *MdSQS1<sup>ScTM</sup>* and *MdSQS2<sup>ScTM</sup>* on YPGal medium and incubated for 5 days at 30 °C. **b** TLC analysis of the reaction products of *Erg9*, *MdSQS1* and *MdSQS2*. Farnesyl diphosphate and NADPH were incubated with the different yeast crude extracts. Authentic squalene was used as standard. *Upper part* shows densitometry-based relative quantification of squalene



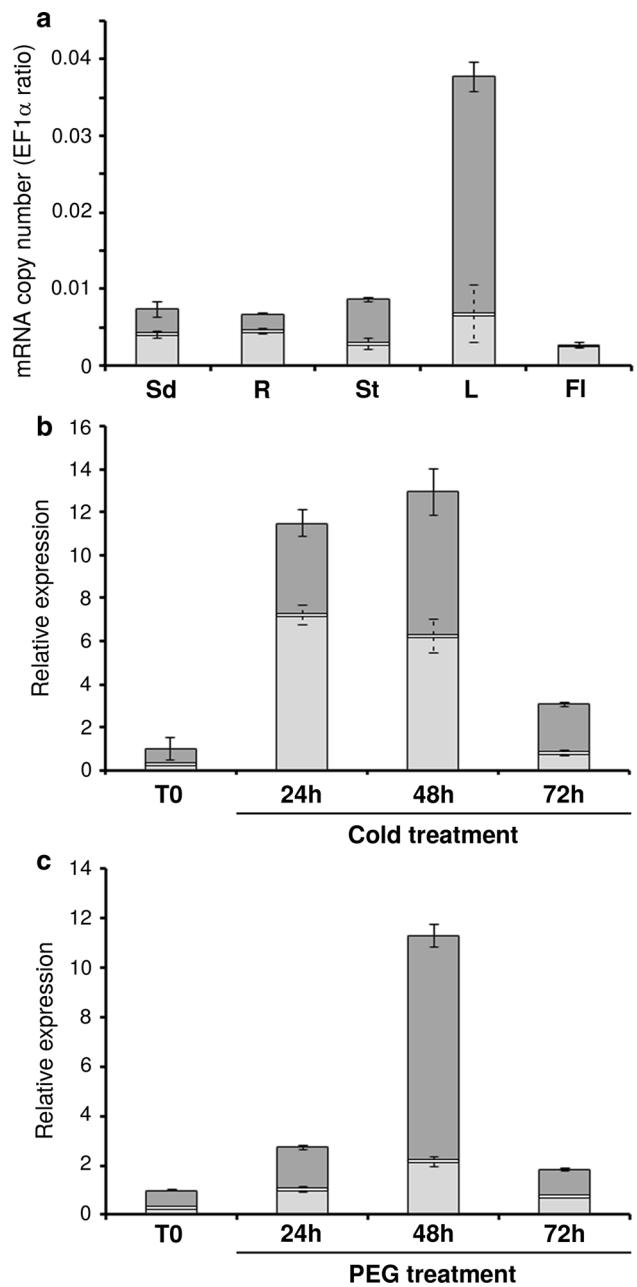
**Fig. 4** Subcellular localization of MdSQS1 and MdSQS2. *C. roseus* cells were transiently transformed with YFP-MdSQS1 (**a**) or YFP-MdSQS2 (**e**) expressing vector, in addition to a plasmid expressing an ER-CFP marker (**b**, **f**). Co-localization of the two fluorescence signals appears on the merged images (**c**, **g**). Cell morphology (**d**, **h**) was observed with differential interference contrast (DIC). Bars 10  $\mu$ m

expressed in transiently transformed *C. roseus* cells (Guirimand et al. 2009). The fluorescence signals of YFP-MdSQS1 and YFP-MdSQS2 (Fig. 4a, e) were observed around the nucleus, branching throughout the cell and perfectly superimposed with the signal of the ER-CFP marker (Fig. 4b, c, f, g). Thus, both apple tree SQSs are likely to be anchored to the endoplasmic reticulum membrane as expected and probably facing to the cytosol.

#### MdSQS isoforms display distinct expression patterns

Distribution of *MdSQS1* and *MdSQS2* transcripts among apple tree organs was investigated by quantitative RT-PCR. Total apple tree *SQS* transcripts (*MdSQS1* + *MdSQS2*) and *MdSQS1*/*MdSQS2* transcript ratios among each condition were measured separately and condensed in Fig. 5a. Total *SQS* transcripts accumulated equally in seedlings, roots and stems. In comparison, transcripts accumulated fourfold higher in leaves and threefold lower in flowers. While both *MdSQS1* and *MdSQS2* were expressed in all organs, *MdSQS1*/*MdSQS2* ratios displayed dramatic differences with a predominance of *MdSQS1* in roots and flowers, representing, respectively, 70 and 90% of *SQS* transcripts and inversely a predominance of *MdSQS2* in stems and leaves representing, respectively, 70 and 80% of *SQS* transcripts (Fig. 5a).

To assess potential variations of gene expression through environmental constraint, *SQS* expression level was investigated after abiotic stress (Fig. 5b, c). First, cold treatment led to a 12-fold transient increase of *SQS* transcripts at 24 h, which was maintained at 48 h and decreased up to the basal level at 72 h (Fig. 5b). Whereas *MdSQS2* transcripts were predominant in control leaves compared to *MdSQS1*, both were up-regulated under cold stress and reached almost equivalent transcript ratios.



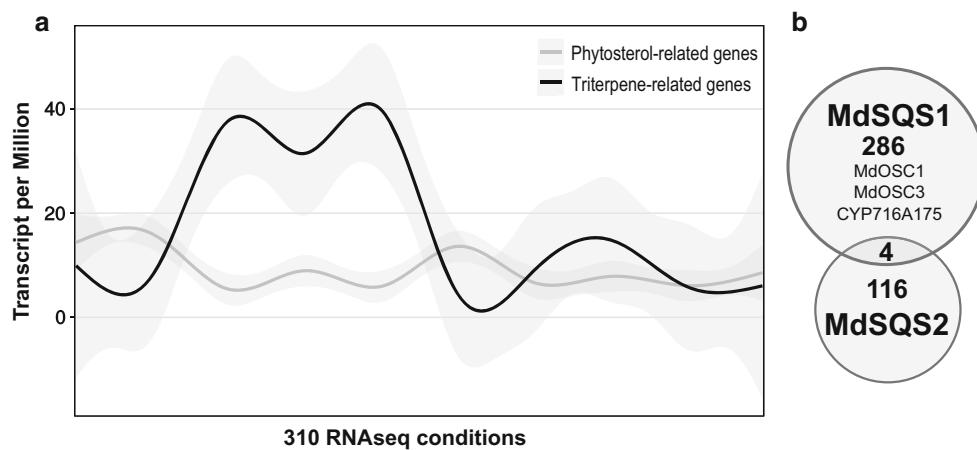
**Fig. 5** Analysis of apple tree *SQS* expression pattern. Combined *MdSQS1* and *MdSQS2* transcript levels (total *SQS* transcripts) were measured by RT-qPCR in various organs (**a**) and in leaves subjected to cold (**b**) or PEG-6000 (**c**) treatment. Transcript copy numbers (**a**) and relative expression (**b**, **c**) were normalized using *MdEF1 $\alpha$* . Data correspond to mean values of three samples, each one composed of plant material from three plants. Ratio between *MdSQS1* (light gray) and *MdSQS2* (dark gray) among total *SQS* transcripts are separated by double line on each histogram. Standard deviations for *MdSQS* ratio are indicated in dot line. *Sd* seedling, *R* root, *St* stem, *L* leaf, *Fl* flower

Therefore, *MdSQS1* had a more pronounced induction factor (30-fold) than *MdSQS2* (fivefold) under cold treatment. Secondly, to induce and control plant water deficit,

we used PEG6000 treatments. Under hydric stress, *SQS* transcripts displayed the same induction profile as under cold stress. Notably, the 12-fold induction factor was conserved, albeit the increase of *SQS* transcripts was delayed to 48 h. Finally, transcript amounts decreased to basal level at 72 h. In contrast to cold stress, *MdSQS1/MdSQS2* ratio of leaves (20/80) remained unchanged all along the experiment. Indeed, *MdSQS1* and *MdSQS2* have an equivalent 12-fold induction factor in response to hydric stress, leading to a predominant expression of *MdSQS2* in these conditions. These results thus suggest that *SQS* expression is induced under abiotic stress to potentially change membrane composition since phytosterol and triterpene contents impact membrane fluidity and stability depending on temperature and osmotic fluctuation.

The differential induction of *SQS* isoform genes upon stresses as well as their organ-specific ratio addresses the question of their specific involvement in distinct downstream pathway branches leading to phytosterol or triterpene synthesis. Therefore, additional expression analyses were performed through RNAseq data mining (Suppl. Table S2). A total of 310 RNAseq samples were analyzed from EBI ENA for *M. domestica*, regrouping a wide range of conditions including various organs (e.g., roots, shoot apex, cotyledons and seeds), developmental processes (e.g., flowers and fruit development) and stresses (e.g., *Venturia inaequalis* infection). Firstly, the global expression profile of selected triterpene and phytosterol-related genes (Suppl. Table S4) was compared (Fig. 6a) to highlight the overall transcriptional tendencies of each metabolism, according to the usual co-expression of genes

involved in a common synthesis pathway. The expression of triterpene-related genes strongly fluctuated depending on plant organs and environmental constraints whereas phytosterol-related gene expression seemed to be more stable in the same conditions. These divergent profiles supported a weak overlap in the transcriptional regulation of each pathway that may facilitate subsequent gene associations. Therefore, lists of genes co-expressed with *MdSQS1* or *MdSQS2* were generated to identify a potential preferred association of each *SQS* isoform to downstream biosynthetic branch (Suppl. Table S3). Genes retrieved from this analysis shared a strict co-expression profile with the target gene among every tested condition and a Venn diagram of co-expressed genes was generated (Fig. 6b). Surprisingly, *MdSQS* isoforms shared only 4 co-expressed genes (HRR <500) confirming strong specificities in their expression profiles and potential distinct roles for both enzymes. In addition, several occurrences of genes involved in triterpene metabolism were found in the gene list co-expressed with *MdSQS1* including the previously characterized  $\alpha$ -amyrin synthases *MdOSC1/MdOSC3* and the C28 triterpene hydroxylase *CYP716A175*, which argued for a specific involvement of *MdSQS1* in triterpene biosynthesis. On the contrary, none of triterpene or phytosterol-related genes was found to be co-expressed with *MdSQS2*, which might reflect the involvement of *MdSQS2* in different metabolism types. On the other hand, it is not excluded that *MdSQS2* might also be associated with several metabolisms, including or not triterpene and phytosterol biosynthesis, hiding specific gene correlations during our analyses.



**Fig. 6** Determination of co-expressed genes with *MdSQS1* and *MdSQS2* using RNA-seq data. **a** Mean expression profiles of putative triterpene and phytosterol-related genes were compared between 310 conditions of RNA-seq runs available for *M. domestica*. Accession numbers of putative triterpene and phytosterol-related genes are available in Suppl. Fig. S6 and the description of the RNAseq accessions used for the construction of the data matrix is available in

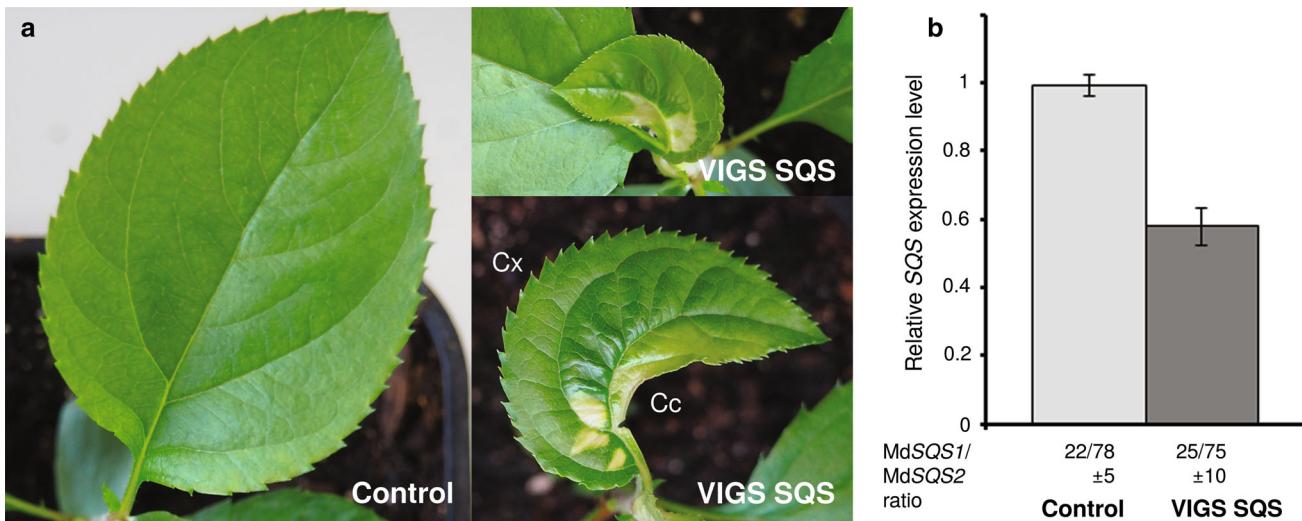
Suppl. Fig. S4. **b** Overlap between best co-expressed genes with each apple tree *SQS* isozyme is visualized on a Venn diagram. Lists were established for an HRR <500. Accession numbers of genes found in the co-expression lists are available in Suppl. Fig. S5. The characterized triterpene-related *MdOSC1*, *MdOSC3* and *CYP716A175* found in *MdSQS1* co-expression list are indicated on the diagram

## MdSQS silencing in apple tree by VIGS leads to abnormal leaves

A virus-induced gene silencing (VIGS) approach was used to investigate the impact of *SQS* down-regulation in apple tree. To significantly abolish the synthesis of squalene in planta, the VIGS assay was designed to target simultaneously *MdSQS1* and *MdSQS2*. To this aim, we generated a modified apple latent spherical virus (ALSV) carrying a 201pb fragment of the *MdSQS1* coding sequence displaying 99% identity with *MdSQS2* and potentially allowing to target both transcripts in planta (Suppl. Fig. S2). The inoculation of this recombinant virus (ALSV-SQS) was performed on cotyledons of apple tree seedlings. VIGS establishment and efficiency were monitored by inoculating control plants with a virus targeting the small subunit of ribulose-1,5-bisphosphate carboxylase (ALSV-*rbcS*) whose silencing generates a visual phenotype of leaf chlorosis (Sasaki et al. 2011). Typically, molecular analyses revealed that the first two true leaves contained ALSV even though the targeted genes were not silenced (Suppl. Fig. S4a and b). Gene silencing was observed in the third leaf with an important heterogeneity as the chlorosis phenotype appeared only in one half of the leaves leading to a 54% decrease in *rbcS* global expression. Finally, a full and stable silencing was measured from the fourth leaf (Suppl. Fig. S4a and b). In plants inoculated with ALSV-SQS, the third true leaves were falciform (shaped like a sickle) rather than oval, partially albino, smaller than in control plants and displaying various sizes (Fig. 6a), while next leaf development was completely aborted. A 42% decrease in

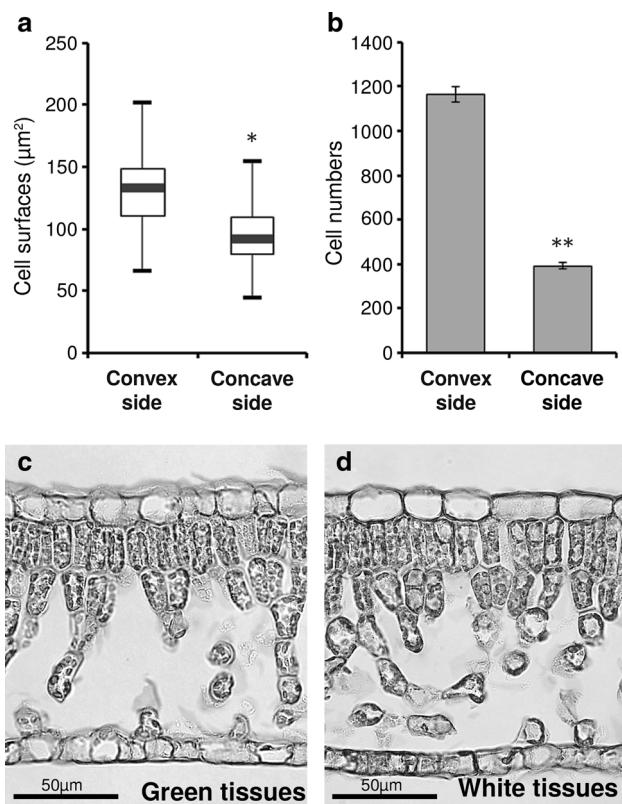
*SQS* expression level was measured in abnormal leaves compared to control plants (Fig. 7b). This *SQS* silencing efficiency is an average and might integrate disparities of VIGS efficiency in the third leaf as observed for *rbcS* (Suppl. Fig. S4b), and explains the heterogeneous phenotype of leaves. Moreover, the absence of fourth leaves potentially revealed the lethality caused by a stronger silencing of *SQS*.

The unexpected falciform and albino phenotypes of the partially silenced *SQS* leaves were further investigated. First, the asymmetric leaf shape in lateral blades (lateral concavity) suggested a difference in tissue development between strong and weakly silenced area. Therefore, surface of palisade parenchyma cells of leaf transversal cross sections was measured on both sides of the central vein (Fig. 8a). Interestingly, cells were slightly but significantly smaller in the concave side of falciform leaves compared to convex side, where cell surface was similar to control leaves (data not shown). However, this weak difference in cell elongation could not explain the pronounced leaf deformation. In addition, the number of cells in the first layer (adjacent to epidermis) of palisade parenchyma was numerated between central vein and leaf margin, on both sides of the *SQS*-silenced leaf (Fig. 8b). We noted an important difference relying on the presence of threefold less cells in the concave side of falciform leaf than in convex side. Therefore, down-regulation of *SQS* expression likely deeply impacts cell division. Secondly, tissues were compared on leaf cross section between green parts and albino spots (Fig. 8c, d). No difference was found in leaf tissue anatomy such as epidermis, palisade or spongy



**Fig. 7** Down-regulation of *SQS* in apple tree strongly affects plant development. *MdSQS1* and *MdSQS2* were simultaneously targeted by VIGS. **a** Typical leaf phenotypes of VIGS SQS plants are shown and compared with control plants, transformed with ALSV empty construct. **b** Relative *SQS* expression (representing both *MdSQS1*

and *MdSQS2*) was measured by RT-qPCR. Ratio between *MdSQS1* and *MdSQS2* among total *SQS* transcripts is indicated under each graphics. *MdEF1α* was used as a reference gene. Data correspond to average values of four samples, each one composed of leaves from ten independent transformed plants. Cx convex, Cc concave

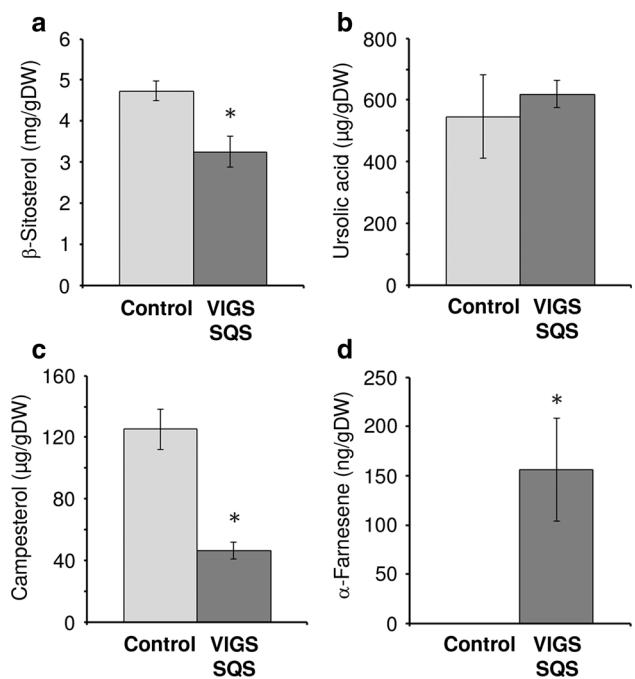


**Fig. 8** *SQS* silencing effects on cell leaf anatomy and structure. Surface (a) and number (b) of palisade cells, observed on cross sections, were compared between convex and concave side of falciform leaves silenced for *SQS*. Plastids are observed on cross section stained with Astra blue. Comparison was made between green tissues (c) and white tissues (d). Asterisk denote statistical significance [ $P < 0.0001$  Student's *t* test,  $n = 100$  (a) and  $n = 6$  (b)]

mesophyll cells indicating that cell differentiation occurred properly all through the leaves, independently of *SQS* expression level. Moreover, abundant plastids were observed in parenchyma cells even in albino tissues, with similar size and quantity compared to green parts. Interestingly, the albino phenotype can be attributed to the depigmentation of chloroplast in *SQS*-silenced leaves rather than a loss of plastids ultrastructure.

#### *SQS*-silencing disturbs terpenoid metabolism

Changes in terpene accumulation were studied in *SQS* knock-down leaves by quantification of the main metabolites produced downstream of *SQS* in the biosynthetic pathway such as ursolic acid and  $\beta$ -sitosterol as the major triterpene and phytosterol in apple tree, respectively (Fig. 1). Ursolic acid levels were similar in leaves of plants silenced for *SQS* and control plants while, in contrast,  $\beta$ -sitosterol contents were significantly lower in *SQS*-silenced leaves (Fig. 9a, b). Campesterol—a phytosterol intermediate in brassinolide biosynthesis—was also quantified.

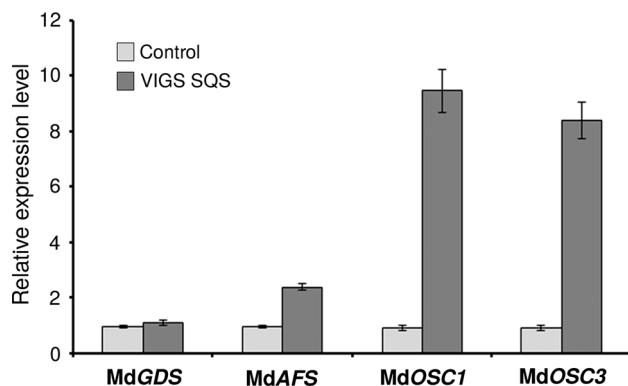


**Fig. 9** Sesquiterpene, triterpene and phytosterol contents in *SQS*-silenced leaves compared to control plants. Main compounds of each triterpenoid family were used including  $\beta$ -sitosterol (a), ursolic acid (b), campesterol (c) and  $\alpha$ -farnesene (d). Standard deviation is representative of 6–9 measurements performed on biological replicates composed of ten leaves. Asterisk denotes statistical significance compared to control ( $P < 0.001$  Student's *t* test)

Campesterol contents decreased as observed for  $\beta$ -sitosterol, but in a stronger manner. *SQS*-silenced leaves only accumulated 37% of control plant campesterol (Fig. 9c). Finally, since sesquiterpene synthases and *SQS* share FPP as substrate, measurements of sesquiterpene contents were also performed. In control plant leaves, sesquiterpene amounts did not reach detection level, whilst  $\alpha$ -farnesene was clearly accumulated in *SQS*-silenced leaves (Fig. 9d). In addition, low amounts of germacrene D were sporadically detected in few *SQS*-silenced plants (data not shown). Taken together, these results clearly showed the importance of *SQS* in phytosterol production and probably on subsequent brassinolide synthesis; however, no impact was found on triterpene synthesis. Finally, down-regulation of *SQS* seemed to increase the FPP pool available in planta for alternative terpene pathways such as sesquiterpene biosynthesis.

#### Reduced triterpenoid accumulation in *SQS*-silenced leaves promotes the expression of terpene biosynthetic enzymes

Since triterpenoid is a fine-tuned metabolism, the effects of *SQS*-silencing and the resulting metabolic perturbations were investigated by analyzing the expression of branch



**Fig. 10** Effect of *SQS* silencing on sesquiterpene synthase and oxidosqualene cyclase expression levels measured by RT-qPCR. Relative expression levels of *MdGDS* (germacrene D synthase), *MdAFS* ( $\alpha$ -farnesene synthase), *MdOSC1* and *MdOSC3* ( $\alpha$ -amyrin synthases) in *SQS*-silenced leaves were compared to control plants and normalized with *MdEF1 $\alpha$* . Data correspond to mean values of four samples, each one composed of leaves from ten independent transformed plants

point oxidosqualene cyclases (OSC) including the two previously characterized apple tree  $\alpha$ -amyrin synthases *MdOSC1* and *MdOSC3* (Fig. 10) (Brendolise et al. 2011). Interestingly, both genes displayed an important upregulation in *SQS*-silenced plants with eightfold more transcripts compared to control plants. The lack in *SQS* expression seems to act positively on expression of OSC dedicated to triterpene synthesis. Sterol-branch specific OSC such as CAS has still not been characterized in apple tree and could not be included in the present analysis. Otherwise, the expression levels of two sesquiterpene synthases (Nieuwenhuizen et al. 2013) were also measured since  $\alpha$ -farnesene and germacrene D transcripts were detected in silenced plants (Fig. 10). A 2.5-fold increase of  $\alpha$ -farnesene synthase (*MdAFS*) and no difference for germacrene D synthase (*MdGDS*) were observed. Taken together, these observations show that reduced triterpenoid accumulation in *SQS*-silenced leaves clearly promotes the expression of triterpene biosynthetic enzymes, while weakly impacts the sesquiterpene biosynthetic enzymes.

## Discussion

Squalene synthase is an essential enzyme for eukaryotic cells, catalyzing FPP condensation to form squalene. In this study, we demonstrated that apple tree contains two functional *SQS* isoforms (Fig. 3). *SQS* gene copy number is heterogeneous in higher plants. Similar to apple tree, four other plant species were reported to contain at least two *SQS* isoforms (Devarenne et al. 1998; Hayashi et al. 1999; Suzuki et al. 2002; Kim et al. 2011). However, some plant species were also found to harbor only one *SQS* gene as in

mammals and yeast (Hata et al. 1997; Huang et al. 2007; Uchida et al. 2009). Therefore, *SQS* duplication events in plants seem to be recent and to have occurred in a species-specific manner. This hypothesis is reinforced by the higher identities between *SQS* isoforms within species than those of *SQS* orthologues (Fig. 2).

Both *MdSQS* isoforms are membrane-bound enzymes, anchored in ER (Fig. 4), which confirms the conservation and functionality of the C-terminal transmembrane domain in the apple tree enzymes. However, full-length coding sequences of *MdSQS1* and *MdSQS2* failed to complement the ergosterol synthesis pathway in the *erg9* yeast mutant. This phenomenon, already observed in several other complementation experiments with plant *SQS*, was previously circumvented by replacement of the plant *SQS* transmembrane domain by the equivalent region from the yeast enzyme (Kribii et al. 1997). However, such approach did not lead to complementation for apple tree *SQS* (Fig. 3). The squalene synthesis activity of *MdSQS1* and *MdSQS2* was finally demonstrated in vitro using yeast crude extracts (Fig. 3). In our conditions, activity of apple tree enzymes was clearly lower than wild-type yeast *SQS* (*Erg9*). Therefore, the absence of yeast complementation even with apple tree *SQS* enzymes modified in their C-terminal domain may be the consequence of a weak intrinsic activity in yeast system. Beyond the confirmation of apple tree *SQS* functionality, these results raise the question of the limits of yeast *erg9* complementation method to identify active *SQS* enzymes since it might easily generate false-negative results.

Phytosterol and triterpene biosynthesis both originate from the common precursor 2,3-oxidosqualene whose synthesis requires *SQS* activity (Fig. 1). However, for both pathways, differences in tissue location exist (Wegel et al. 2009), most probably depending on end-product functions. Genes related to sterol biosynthesis are expressed constitutively throughout the plant since phytosterols are essential for plant growth and development (Clouse 2002; Lindsey et al. 2003). On the contrary, triterpenes act as protective compounds against biotic stresses (González-Coloma et al. 2011), and are differentially accumulated in plant tissues such as in apple skin and leaf cuticle wax (Bringe et al. 2006; Jäger et al. 2009). Therefore, such compartmentalization of specialized metabolism could explain the differential expression profiles observed in our global triterpene and phytosterol-related gene expression analysis (Fig. 6a). Furthermore, the emergence of *SQS* isoforms in apple tree seems to have been followed by a divergence in their expression and a differential involvement in downstream biosynthetic pathways (Figs. 5, 6). *MdSQS2* presented the highest expression level in aerial organs and could be the main isoform involved in both triterpene and phytosterol biosynthesis. Albeit our

bioinformatic analyses did not lead to any association to specific metabolism, one could speculate that *MdSQS2* may integrate triterpene and phytosterol-related gene expression profiles leading to a complex and dissimilar expression pattern which could explain the low number of co-expressed genes for *MdSQS2* and the absence of strict occurrences with one or the other downstream pathways. On the contrary, *MdSQS1* was globally less expressed but restricted to triterpene synthesis (Figs. 5, 6). *MdSQS1* could thus correspond to the recently duplicated SQS isoform, which completes *MdSQS2* functions. However, all these hypothesis are based on qPCR expression studies and gene co-expression network of *MdSQS1* and *MdSQS2* and would require further additional in vivo characterization.

VIGS approaches were successfully developed in apple tree several years ago (Sasaki et al. 2011) but until now, no study has taken advantage of this strategy to disturb metabolic flux and to investigate the function of metabolic enzymes in planta. This was achieved herein for *MdSQS1* and *MdSQS2*; both isoforms were targeted simultaneously due to high sequence identity. Interestingly, although a plethora of *SQS* was cloned among plants, the use of reverse genetic approaches for in planta *SQS* functional characterization especially using gene down-regulation is scarce (Manavalan et al. 2012; Wang et al. 2012; Singh et al. 2015). This is probably due to the essential roles played by sterols for cells, the disruption of their biosynthesis resulting in lethality (Novotny and Karst 1994; Babiychuk et al. 2008; Kim et al. 2010). We demonstrated that the use of VIGS method circumvents this problem, especially in apple tree since various levels of silencing efficiency naturally occur in newly developed leaves (Suppl. Fig. S4). While the most *SQS*-silenced leaves aborted probably due to the lack of sterols, weakly silenced pair of leaves were used for our analysis (Fig. 7). The main counterpart of this incomplete silencing was the production of heterogeneous and partial phenotypes leading to less nuanced metabolite variations (Fig. 9). On the other hand, this study on *MdSQS* highlights that the use of the ALSV-based VIGS method in apple tree constitutes a powerful tool to achieve fast functional genomics on metabolism of interest and allows characterization of essential genes in planta.

*SQS*-silencing in apple tree disturbed terpene metabolism with a main impact on phytosterol accumulation but not on triterpene synthesis (Fig. 9). Indeed, the potential low level of squalene resulting from *SQS*-silencing and consequently reduced amount of 2,3-oxidosqualene seems to positively regulate the expression of triterpene-related *OSC* (Brendolise et al. 2011; Andre et al. 2016) such as *MdOSC1* and *MdOSC3* (Fig. 10). This induced expression might compensate the limiting 2,3-oxidosqualene amount and drains the metabolic flux toward triterpene synthesis instead of phytosterol, thus explaining the absence of

significant alteration in triterpene accumulation. Unfortunately, *CAS* genes were not identified yet in apple tree impeding the study of the impact of *SQS*-silencing on sterol-specific *OSC* expression. However, the important differences in relative amounts of phytosterol compared to triterpene in untreated plants (Fig. 9) argue for a naturally higher limiting flux for sterol biosynthesis.

Apple tree is able to produce a large range of sesquiterpenes including  $\alpha$ -farnesene and germacrene D (Nieuwenhuizen et al. 2013). Sesquiterpenes are volatile compounds involved in plant defense and produced in response to biotic stresses. Sesquiterpene synthases use FPP as a substrate and, therefore, can compete with *SQS* and triterpenoid pathways (Fig. 1). In fact, elicitor treatment on parsley and tobacco cell cultures was found to stimulate sesquiterpene synthesis with an additional inhibition effect on *SQS* activity which is probably needed to get enough metabolic flux available for sesquiterpene fast and efficient production (Vögeli and Chappell 1988; Haudenschild and Hartmann 1995; Deverenne et al. 1998). Therefore, an important change in sesquiterpene synthesis was expected in *SQS*-silenced plants. In spite of a substantial accumulation of sesquiterpenes in silenced plants compared to controls, their total amounts were still poor in comparison of the loss of phytosterols (Fig. 9a, c, d). Natural expression level of sesquiterpene synthases in leaves is limited in the absence of environmental constraint (greenhouse growth) and the weak inductions of *MdAFS* and *MdGDS* in *SQS*-silenced plants (Fig. 10c, d) were probably not sufficient to reroute efficiently metabolic flux through sesquiterpenes.

Since lower phytosterol amount was the main alteration observed in apple tree *SQS*-silenced leaves (Fig. 9), phenotypic alterations could potentially be attributed to sterol deprivation. A plastid depigmentation (Fig. 8) was previously reported in an *Arabidopsis* T-DNA mutant affected in *CAS1* expression (Babiychuk et al. 2008). *CAS1* catalyzes the first enzymatic step specific to phytosterol synthesis, downstream of *SQS* (Fig. 1). Whereas the knock-out allelic mutations *cas1-2* and *cas1-3* did not allow the generation of viable lines, the knock-down allelic mutation *cas1-1* allowed plant development with albino inflorescence shoots suggesting a role of phytosterol in plastid biogenesis. Leaf bleaching was also observed in tobacco down-regulated for *CAS1* using a VIGS approach, confirming the previous result (Gas-Pascual et al. 2014). Even if the chloroplast inner membrane and thylakoids did not contain sterols, they may indirectly disturb chloroplast biogenesis through membrane trafficking with sterol-rich ER. Through our *SQS*-silencing approach, the link between phytosterol depletion and albino phenotype was confirmed by targeting an alternative gene in sterol biosynthesis pathway. In addition to the albino phenotype, falciform leaves were obtained in *SQS*-silenced plants. This falciform phenotype

results from a differential leaf growth on both sides of the central vein (Fig. 8) and may be a consequence from a differential silencing efficiency among the leaf. Developmental alterations are usually observed in brassinosteroid-deficient mutants, characterized by pronounced dwarf phenotype (Schaller 2003). Brassinosteroids are well known to play important roles in the regulation of cell division and elongation (Hu et al. 2000; Miyazawa et al. 2003; Müssig 2005). The strong reduction level of campesterol in *SQS*-silenced leaves (Fig. 9) could explain the observed phenotype since campesterol is the precursor of brassinolide (Fig. 1). However, a direct effect of phytosterol deficiency on cell division cannot be excluded. Indeed, exogenous brassinosteroid treatments were not able to systematically rescue the dwarf phenotype of sterol deficient mutant, indicating a brassinosteroid-independent effect of sterols on development (Schaller 2003).

In summary, the characterization of two *SQS* genes in apple tree brings new insights into the involvement of *SQS* isoforms in phytosterol and triterpene metabolisms in the green lineage. Furthermore, this study opens the way to metabolic disturbance in apple tree using a VIGS approach for functional characterization in planta. Our results provided additional information towards the function of sterols in apple tree but remained limited for triterpenes. At least five different triterpene-specific OSCs were hitherto identified in apple tree (Brendolise et al. 2011; Andre et al. 2016). Direct and specific down-regulations of these *MdOSC* will be needed to give more insight into the complex metabolic flux balance among triterpenoid biosynthesis, to confirm OSC function in planta, and to understand apple tree triterpenes biological function.

**Author contribution statement** SNG, CEP, DD, FJ, NP, VC, AL and SB conducted experiments. FL and TDDB achieved bioinformatics analyses. AO, GG, OP, NY, VC, MC participated in the design of the study and interpretation. NGG, BSP, LA, MC assisted in the supervision of this work. SB conceived; supervised, coordinated the work and wrote the manuscript. VC, NP and TDDB revised the article. All authors read and approved the final manuscript.

**Acknowledgements** The authors wish to thank Dr. Marc Fischer (INRA Colmar) for providing yeast *erg9* strain and Dr. Marie-Noëlle Brisset (INRA Angers) for providing apple tree seeds. We gratefully acknowledge support from the Région Centre-Val de Loire (France) for financing SNG and CEP. DD was financed by a doctoral fellowship from the Région Centre-Val de Loire (France) and the Ministère de l'Enseignement Supérieur et de la Recherche (France).

#### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Adams RP (2007) Identification of essential oil components by gas chromatography/mass spectrometry. Allured Publishing Corporation, Carol Stream
- Andre C, Legay S, Deleruelle A, Nieuwenhuizen N, Punter M, Brendolise C, Cooney JM, Lateur M, Hausman JF, Larondelle Y, Laing W (2016) Multifunctional oxidosqualene cyclases and cytochrome P450 involved in the biosynthesis of apple fruit triterpenic acids. *New Phytol* 211:1279–1294. doi:[10.1111/nph.13996](https://doi.org/10.1111/nph.13996)
- Babiychuk E, Bouvier-Navé P, Compagnon V, Suzuki M, Muranaka T, Van Montagu M, Kushnir S, Schaller H (2008) Allelic mutant series reveal distinct functions for *Arabidopsis* cycloartenol synthase 1 in cell viability and plastid biogenesis. *Proc Natl Acad Sci USA* 105:3163–3168. doi:[10.1073/pnas.0712190105](https://doi.org/10.1073/pnas.0712190105)
- Baker CH, Matsuda SP, Liu DR, Corey EJ (1995) Molecular cloning of the human gene encoding lanosterol synthase from a liver cDNA library. *Biochem Biophys Res Commun* 213:154–160. doi:[10.1006/bbrc.1995.2110](https://doi.org/10.1006/bbrc.1995.2110)
- Becker DM, Guarente L (1991) High-efficiency transformation of yeast by electroporation. *Methods Enzymol* 194:182–187. doi:[10.1016/0076-6879\(91\)94015-5](https://doi.org/10.1016/0076-6879(91)94015-5)
- Benveniste P (2002) Sterol metabolism. *Arabidopsis Book* 1:e0004. doi:[10.1199/tab.0004](https://doi.org/10.1199/tab.0004)
- Brendolise C, Yauk YK, Eberhard ED, Wang M, Chagne D, Andre C, Greenwood D, Beuning LL (2011) An unusual plant triterpene synthase with predominant  $\alpha$ -amyrin-producing activity identified by characterizing oxidosqualene cyclases from *Malus × domestica*. *FEBS J* 278:2485–2499. doi:[10.1111/j.1742-4658.2011.08175.x](https://doi.org/10.1111/j.1742-4658.2011.08175.x)
- Bringe K, Schumacher CF, Schmitz-Eiberger M, Steiner U, Oerke E (2006) Ontogenetic variation in chemical and physical characteristics of adaxial apple leaf surfaces. *Phytochemistry* 67:161–170. doi:[10.1016/j.phytochem.2005.10.018](https://doi.org/10.1016/j.phytochem.2005.10.018)
- Busquets A, Keim V, Closa M, Del Arco A, Boronat A, Arró M, Ferrer A (2008) *Arabidopsis thaliana* contains a single gene encoding squalene synthase. *Plant Mol Biol* 67:25–36. doi:[10.1007/s11103-008-9299-3](https://doi.org/10.1007/s11103-008-9299-3)
- Clouse SD (2002) *Arabidopsis* mutants reveal multiple roles for sterols in plant development. *Plant Cell* 14:1995–2000. doi:[10.1105/tpc.140930](https://doi.org/10.1105/tpc.140930)
- Corey EJ, Matsuda SPT, Bartel B (1993) Isolation of an *Arabidopsis thaliana* gene encoding cycloartenol synthase by functional expression in a yeast mutant lacking lanosterol synthase by the use of a chromatographic screen. *Proc Natl Acad Sci USA* 90:11628–11632
- D'Abrosca B, Fiorentino A, Monaco P, Pacifico S (2005) Radical-scavenging activities of new hydroxylated ursane triterpenes from cv. Annurca apples. *Chem Biodivers* 2:953–958. doi:[10.1002/cbdv.200590072](https://doi.org/10.1002/cbdv.200590072)
- Devarenne TP, Shin DH, Back K, Yin S, Chappell J (1998) Molecular characterization of tobacco squalene synthase and regulation in response to fungal elicitor. *Arch Biochem Biophys* 349:205–215. doi:[10.1006/abbi.1997.0463](https://doi.org/10.1006/abbi.1997.0463)
- Dugé de Bernonville T, Guyot S, Paulin JP, Gaucher M, Loufrani L, Henrion D, Derbré S, Guilet D, Richomme P, Dat JF, Brisset MN (2010) Dihydrochalcones: implication in resistance to oxidative stress and bioactivities against advanced glycation end-products and vasoconstriction. *Phytochemistry* 71:443–452. doi:[10.1016/j.phytochem.2009.11.004](https://doi.org/10.1016/j.phytochem.2009.11.004)
- Eberhardt MV, Lee CY, Liu RH (2000) Antioxidant activity of fresh apples. *Nature* 405:903–904. doi:[10.1038/35016148](https://doi.org/10.1038/35016148)

- Elejalde-Palmett C, Dugé de Bernonville T, Glévarec G, Pichon O, Papon N, Courdavault V, St-Pierre B, Giglioli-Guivarc'H N, Lanoue A, Besseau S (2015) Characterization of a spermidine hydroxycinnamoyltransferase in *Malus domestica* highlights the evolutionary conservation of trihydroxycinnamoyl spermidines in pollen coat of core Eudicotyledons. *J Exp Bot* 66:7271–7285. doi:[10.1093/jxb/erv423](https://doi.org/10.1093/jxb/erv423)
- Gas-Pascual E, Berna A, Bach TJ, Schaller H (2014) Plant oxidosqualene metabolism: cycloartenol synthase-dependent sterol biosynthesis in *Nicotiana benthamiana*. *PLoS One* 9:e109156. doi:[10.1371/journal.pone.0109156](https://doi.org/10.1371/journal.pone.0109156)
- González-Coloma A, López-Balboa C, Santana O, Reina M, Fraga BM (2011) Triterpene-based plant defenses. *Phytochem Rev* 10:245–260. doi:[10.1007/s11101-010-9187-8](https://doi.org/10.1007/s11101-010-9187-8)
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40:1178–1186. doi:[10.1093/nar/gkr944](https://doi.org/10.1093/nar/gkr944)
- Gu P, Ishii Y, Spencer TA, Shechter I (1998) Function-structure studies and identification of three enzyme domains involved in the catalytic activity in rat hepatic squalene synthase. *J Biol Chem* 273:12515–12525. doi:[10.1074/jbc.273.20.12515](https://doi.org/10.1074/jbc.273.20.12515)
- Guirimand G, Burlat V, Oudin A, Lanoue A, St-Pierre B, Courdavault V (2009) Optimization of the transient transformation of *Catharanthus roseus* cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. *Plant Cell Rep* 28:1215–1234. doi:[10.1007/s00299-009-0722-2](https://doi.org/10.1007/s00299-009-0722-2)
- Hartmann M-A (1998) Plant sterols and the membrane environment. *Trends Plant Sci* 3:170–175. doi:[10.1016/S1360-1385\(98\)01233-3](https://doi.org/10.1016/S1360-1385(98)01233-3)
- Hata S, Sanmiya K, Kouchi H, Matsuoka M, Yamamoto N, Izui K (1997) cDNA cloning of squalene synthase genes from mono- and dicotyledonous plants and expression of the gene in rice. *Plant Cell Physiol* 38:1409–1413. doi:[10.1093/oxfordjournals.pcp.a029137](https://doi.org/10.1093/oxfordjournals.pcp.a029137)
- Haudenschild C, Hartmann M-A (1995) Accumulation of furanocoumarins in parsley cell suspension cultures. *Phytochemistry* 40:1117–1124. doi:[10.1016/0031-9422\(95\)00434-9](https://doi.org/10.1016/0031-9422(95)00434-9)
- Hayashi H, Hirota A, Hiraoka N, Ikeshiro Y (1999) Molecular cloning and characterization of two cDNAs for *Glycyrrhiza glabra* squalene synthase. *Biol Pharm Bull* 22:947–950
- He X, Rui HL (2007) Triterpenoids isolated from apple peels have potent antiproliferative activity and may be partially responsible for apple's anticancer activity. *J Agric Food Chem* 55:4366–4370. doi:[10.1021/jf063563o](https://doi.org/10.1021/jf063563o)
- Herrera JBR, Bartel B, Wilson WK, Matsuda SP (1998) Cloning and characterization of the *Arabidopsis thaliana* lupeol synthase gene. *Phytochemistry* 49:1905–1911. doi:[10.1016/S0031-9422\(98\)00366-5](https://doi.org/10.1016/S0031-9422(98)00366-5)
- Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR (1989) Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* 77:51–59. doi:[10.1016/0378-1119\(89\)90358-2](https://doi.org/10.1016/0378-1119(89)90358-2)
- Hu Y, Bao F, Li J (2000) Promotive effect of brassinosteroids on cell division involves a distinct CycD3-induction pathway in *Arabidopsis*. *Plant J* 24:693–701. doi:[10.1046/j.1365-313X.2000.00915.x](https://doi.org/10.1046/j.1365-313X.2000.00915.x)
- Huang Z, Jiang K, Pi Y, Hou R, Liao Z, Cao Y, Han X, Wang Q, Sun X, Tang K (2007) Molecular cloning and characterization of the yew gene encoding squalene synthase from *Taxus cuspidata*. *J Biochem Mol Biol* 40:625–635
- Jäger S, Trojan H, Kopp T, Laszczyk MN, Scheffler A (2009) Pentacyclic triterpene distribution in various plants—rich sources for a new group of multi-potent plant extracts. *Molecules* 14:2016–2031. doi:[10.3390/molecules14062016](https://doi.org/10.3390/molecules14062016)
- Jennings SM, Tsay YH, Fisch TM, Robinson GW (1991) Molecular cloning and characterization of the yeast gene for squalene synthetase. *Proc Natl Acad Sci USA* 88:6038–6042
- Kim HB, Lee H, Oh CJ, Lee HY, Eum HL, Kim HS, Hong YP, Lee Y, Choe S, An C, Choi SB (2010) Postembryonic seedling lethality in the sterol-deficient *Arabidopsis cyp51A2* mutant is partially mediated by the composite action of ethylene and reactive oxygen species. *Plant Physiol* 152:192–205. doi:[10.1104/pp.109.149088](https://doi.org/10.1104/pp.109.149088)
- Kim T-D, Han J-Y, Huh GH, Choi Y-E (2011) Expression and functional characterization of three squalene synthase genes associated with saponin biosynthesis in *Panax ginseng*. *Plant Cell Physiol* 52:125–137. doi:[10.1093/pcp/pcq179](https://doi.org/10.1093/pcp/pcq179)
- Kolesnikova MD, Xiong Q, Lodeiro S, Hua L, Matsuda SPT (2006) Lanosterol biosynthesis in plants. *Arch Biochem Biophys* 447:87–95. doi:[10.1016/j.abb.2005.12.010](https://doi.org/10.1016/j.abb.2005.12.010)
- Kribii R, Arró M, Del Arco A, González V, Balcells L, Delourme D, Ferrer A, Karst F, Boronat A (1997) Cloning and characterization of the *Arabidopsis thaliana SQS1* gene encoding squalene synthase—involved in the C-terminal region of the enzyme in the channeling of squalene through the sterol pathway. *Eur J Biochem* 249:61–69. doi:[10.1111/j.1432-1033.1997.00061.x](https://doi.org/10.1111/j.1432-1033.1997.00061.x)
- Kushiro T, Shibuya M, Ebizuka Y (1998) Cloning of oxidosqualene cyclase that catalyzes the formation of the most popular triterpene among higher plants. *Eur J Biochem* 256:238–244
- Lee MH, Jeong JH, Seo JW, Shin CG, Kim YS, In JG, Yang DC, Yi JS, Choi YE (2004) Enhanced triterpene and phytosterol biosynthesis in *Panax ginseng* overexpressing squalene synthase gene. *Plant Cell Physiol* 45:976–984. doi:[10.1093/pcp/pch126](https://doi.org/10.1093/pcp/pch126)
- Lindsey K, Pullen ML, Topping JF (2003) Importance of plant sterols in pattern formation and hormone signalling. *Trends Plant Sci* 8:521–525. doi:[10.1016/j.tplants.2003.09.012](https://doi.org/10.1016/j.tplants.2003.09.012)
- Ma CM, Cai SQ, Cui JR, Wang RQ, Tu PF, Hattori M, Daneshtalab M (2005) The cytotoxic activity of ursolic acid derivatives. *Eur J Med Chem* 40:582–589. doi:[10.1016/j.ejmech.2005.01.001](https://doi.org/10.1016/j.ejmech.2005.01.001)
- Manavalan LP, Chen X, Clarke J, Salmeron J, Nguyen HT (2012) RNAi-mediated disruption of squalene synthase improves drought tolerance and yield in rice. *J Exp Bot* 63:163–175. doi:[10.1093/jxb/err258](https://doi.org/10.1093/jxb/err258)
- Mckenzie TL, Jiang G, Straubhaar JR, Conrad DG, Shechters I (1992) Molecular cloning, expression, and characterization of the cDNA for the rat hepatic squalene synthase. *J Biol Chem* 267:21368–21374
- Miyazawa Y, Nakajima N, Abe T, Sakai A, Fujioka S, Kawano S, Kuroiwa T, Yoshida S (2003) Activation of cell proliferation by brassinolide application in tobacco BY-2 cells: effects of brassinolide on cell multiplication, cell-cycle-related gene expression, and organellar DNA contents. *J Exp Bot* 54:2669–2678. doi:[10.1093/jxb/erg312](https://doi.org/10.1093/jxb/erg312)
- Müssig C (2005) Brassinosteroid-promoted growth. *Plant Biol* 7:110–117. doi:[10.1055/s-2005-837493](https://doi.org/10.1055/s-2005-837493)
- Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23:895–910. doi:[10.1105/tpc.111.083667](https://doi.org/10.1105/tpc.111.083667)
- Nelson BK, Cai X, Nebenführ A (2007) A multicolored set of in vivo organelle markers for co-localization studies in *Arabidopsis* and other plants. *Plant J* 51:1126–1136. doi:[10.1111/j.1365-313X.2007.03212.x](https://doi.org/10.1111/j.1365-313X.2007.03212.x)
- Nieuwenhuizen NJ, Green SA, Chen X, Bailleul EJD, Matich AJ, Wang MY, Atkinson RG (2013) Functional genomics reveals that a compact terpene synthase gene family can account for terpene volatile production in apple. *Plant Physiol* 161:787–804. doi:[10.1104/pp.112.208249](https://doi.org/10.1104/pp.112.208249)

- Novotny C, Karst F (1994) Sterol dependent growth and ethanol tolerance of a sterol-auxotrophic *erg9:HIS3* mutant of *Saccharomyces cerevisiae*. *Biotechnol Lett* 16:539–542. doi:[10.1007/BF01023340](https://doi.org/10.1007/BF01023340)
- Pandit J, Danley DE, Schulte GK, Mazzalupo S, Pauly T, Hayward CM, Hamanaka ES, Thompson JF, Harwood HJ (2000) Crystal structure of human squalene synthase. *Biochemistry* 275:30610–30617. doi:[10.1074/jbc.M004132200](https://doi.org/10.1074/jbc.M004132200)
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2016) Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. Preprint BioarXiv. doi:[10.1101/021592](https://doi.org/10.1101/021592)
- Pechous SW, Whitaker BD (2004) Cloning and functional expression of an (E, E)-farnesene synthase cDNA from peel tissue of apple fruit. *Planta* 219:84–94. doi:[10.1007/s00425-003-1191-4](https://doi.org/10.1007/s00425-003-1191-4)
- Phillips DR, Rasberry JM, Bartel B, Matsuda SP (2006) Biosynthetic diversity in plant triterpene cyclization. *Curr Opin Plant Biol* 9:305–314. doi:[10.1016/j.pbi.2006.03.004](https://doi.org/10.1016/j.pbi.2006.03.004)
- Racovita RC, Jetter R (2016) Composition of the epicuticular waxes coating the adaxial side of *Phyllostachys aurea* leaves: identification of very-long-chain primary amides. *Phytochemistry* 130:1–10. doi:[10.1016/j.phytochem.2016.06.005](https://doi.org/10.1016/j.phytochem.2016.06.005)
- Robinson GW, Tsay YIMH, Kienzle BK, Smith-monroy CA, Bishop RW (1993) Conservation between human and fungal squalene synthetases: similarities in structure, function, and regulation. *Mol Cell Biol* 13:2706–2717. doi:[10.1128/MCB.13.5.2706](https://doi.org/10.1128/MCB.13.5.2706)
- Sasaki S, Yamagishi N, Yoshikawa N (2011) Efficient virus-induced gene silencing in apple, pear and Japanese pear using *Apple latent spherical virus* vectors. *Plant Methods* 7:15–25. doi:[10.1186/1746-4811-7-15](https://doi.org/10.1186/1746-4811-7-15)
- Schaller H (2003) The role of sterols in plant growth and development. *Prog Lipid Res* 42:163–175. doi:[10.1016/S0163-7827\(02\)00047-4](https://doi.org/10.1016/S0163-7827(02)00047-4)
- Schaller H (2004) New aspects of sterol biosynthesis in growth and development of higher plants. *Plant Physiol Biochem* 42:465–476. doi:[10.1016/j.plaphy.2004.05.012](https://doi.org/10.1016/j.plaphy.2004.05.012)
- Seo JW, Jeong JH, Shin CG, Lo SC, Han SS, Yu KW, Harada E, Han JY, Choi YE (2005) Overexpression of squalene synthase in *Eleutherococcus senticosus* increases phytosterol and triterpene accumulation. *Phytochemistry* 66:869–877. doi:[10.1016/j.phytochem.2005.02.016](https://doi.org/10.1016/j.phytochem.2005.02.016)
- Shi Z, Buntel CJ, Griffin JH (1994) Isolation and characterization of the gene encoding 2,3-oxidosqualene-lanosterol cyclase from *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 91:7370–7374
- Singh AK, Dwivedi V, Rai A, Pal S, Reddy SGE, Rao DV, Shasany AK, Nagegowda D (2015) Virus-induced gene silencing of *Withania somnifera* squalene synthase negatively regulates sterol and defence-related genes resulting in reduced withanolides and biotic stress tolerance. *Plant Biotechnol J* 13:1287–1299. doi:[10.1111/pbi.12347](https://doi.org/10.1111/pbi.12347)
- Sonawane PD, Pollier J, Panda S, Szymanski J, Massalha H, Yona M et al (2016) Plant cholesterol biosynthetic pathway overlaps with phytosterol metabolism. *Nat Plants* 3:16205–16218. doi:[10.1038/nplants.2016.205](https://doi.org/10.1038/nplants.2016.205)
- Summers C, Karst F, Charles AD (1993) Cloning, expression and characterisation of the cDNA encoding human hepatic squalene synthase, and its relationship to phytoene synthase. *Gene* 136:185–192. doi:[10.1016/0378-1119\(93\)90462-C](https://doi.org/10.1016/0378-1119(93)90462-C)
- Suzuki H, Achnine L, Xu R, Matsuda SPT, Dixon R (2002) A genomics approach to the early stages of triterpene saponin biosynthesis in *Medicago truncatula*. *Plant J* 32:1033–1048. doi:[10.1046/j.1365-313X.2002.01497.x](https://doi.org/10.1046/j.1365-313X.2002.01497.x)
- Suzuki M, Xiang T, Ohyama K, Seki H, Saito K, Muranaka T, Van Montagu M, Kushnir S, Schaller H (2006) Lanosterol synthase in dicotyledonous plants. *Plant Cell Physiol* 47:565–571. doi:[10.1093/pcp/pcj031](https://doi.org/10.1093/pcp/pcj031)
- Uchida H, Yamashita H, Kajikawa M, Ohyama K, Nakayachi O, Sugiyama R, Yamato KT, Muranaka T, Fukuzawa H, Takemura M, Ohyama K (2009) Cloning and characterization of a squalene synthase gene from a petroleum plant, *Euphorbia tirucalli* L. *Planta* 229:1243–1252. doi:[10.1007/s00425-009-0906-6](https://doi.org/10.1007/s00425-009-0906-6)
- Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A et al (2010) The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nat Genet* 42:833–839. doi:[10.1038/ng.654](https://doi.org/10.1038/ng.654)
- Vincken JP, Heng L, de Groot A, Gruppen H (2007) Saponins, classification and occurrence in the plant kingdom. *Phytochemistry* 68:275–297. doi:[10.1016/j.phytochem.2006.10.008](https://doi.org/10.1016/j.phytochem.2006.10.008)
- Vögeli U, Chappell J (1988) Induction of sesquiterpene cyclase and suppression of squalene synthetase activities in plant cell cultures treated with fungal elicitor. *Plant Physiol* 88:1291–1296. doi:[10.1104/pp.88.4.1291](https://doi.org/10.1104/pp.88.4.1291)
- Wang Z, Guhling O, Yao R, Li F, Yeats TH, Rose JKC, Jetter R (2011) Two oxidosqualene cyclases responsible for biosynthesis of tomato fruit cuticular triterpenoids. *Plant Physiol* 155:540–552. doi:[10.1104/pp.110.162883](https://doi.org/10.1104/pp.110.162883)
- Wang K, Senthil-Kumar M, Ryu C-M, Kang L, Mysore KS (2012) Phytosterols play a key role in plant innate immunity against bacterial pathogens by regulating nutrient efflux into the apoplast. *Plant Physiol* 158:1789–1802. doi:[10.1104/pp.111.189217](https://doi.org/10.1104/pp.111.189217)
- Wang W, Bai MY, Wang ZY (2014) The brassinosteroid signaling network—a paradigm of signal integration. *Curr Opin Plant Biol* 21:147–153. doi:[10.1016/j.pbi.2014.07.012](https://doi.org/10.1016/j.pbi.2014.07.012)
- Wegel E, Koumproglou R, Shaw P, Osbourn A (2009) Cell type-specific chromatin decondensation of a metabolic gene cluster in oats. *Plant Cell* 21:3926–3936. doi:[10.1105/tpc.109.072124](https://doi.org/10.1105/tpc.109.072124)
- Zhang H, Dugé de Bernonville T, Body M, Glévarec G, Reichelt M, Unsicker S, Bruneau M, Renou JP, Huguet E, Dubreuil G, Giron D (2016) Leaf-mining by *Phyllonorycter blancae* reprograms the host-leaf transcriptome to modulate phytohormones associated with nutrient mobilization and plant defense. *J Insect Physiol* 84:114–127. doi:[10.1016/j.jinsphys.2015.06.003](https://doi.org/10.1016/j.jinsphys.2015.06.003)



# SCIENTIFIC REPORTS



OPEN

## Folivory elicits a strong defense reaction in *Catharanthus roseus*: metabolomic and transcriptomic analyses reveal distinct local and systemic responses

Received: 20 July 2016

Accepted: 06 December 2016

Published: 17 January 2017

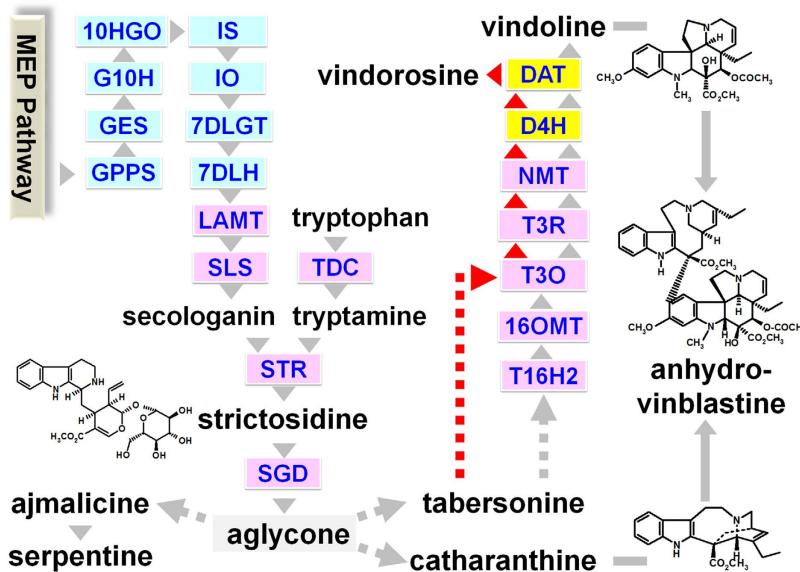
Thomas Dugé de Bernonville<sup>1,\*</sup>, Inês Carqueijeiro<sup>1,\*</sup>, Arnaud Lanoue<sup>1</sup>, Florent Lafontaine<sup>1</sup>, Paloma Sánchez Bel<sup>2</sup>, Franziska Liesecke<sup>1</sup>, Karine Musset<sup>3</sup>, Audrey Oudin<sup>1</sup>, Gaëlle Glévarec<sup>1</sup>, Olivier Pichon<sup>1</sup>, Sébastien Besseau<sup>1</sup>, Marc Clastre<sup>1</sup>, Benoit St-Pierre<sup>1</sup>, Victor Flors<sup>2</sup>, Stéphane Maury<sup>4</sup>, Elisabeth Huguet<sup>3</sup>, Sarah E. O'Connor<sup>5</sup> & Vincent Courdavault<sup>1</sup>

Plants deploy distinct secondary metabolisms to cope with environment pressure and to face bio-aggressors notably through the production of biologically active alkaloids. This metabolism-type is particularly elaborated in *Catharanthus roseus* that synthesizes more than a hundred different monoterpene indole alkaloids (MIAs). While the characterization of their biosynthetic pathway now reaches completion, still little is known about the role of MIAs during biotic attacks. As a consequence, we developed a new plant/herbivore interaction system by challenging *C. roseus* leaves with *Manduca sexta* larvae. Transcriptomic and metabolic analyses demonstrated that *C. roseus* respond to folivory by both local and systemic processes relying on the activation of specific gene sets and biosynthesis of distinct MIAs following jasmonate production. While a huge local accumulation of strictosidine was monitored in attacked leaves that could repel caterpillars through its protein reticulation properties, newly developed leaves displayed an increased biosynthesis of the toxic strictosidine-derived MIAs, vindoline and catharanthine, produced by up-regulation of MIA biosynthetic genes. In this context, leaf consumption resulted in a rapid death of caterpillars that could be linked to the MIA dimerization observed in intestinal tracts. Furthermore, this study also highlights the overall transcriptomic control of the plant defense processes occurring during herbivory.

The Madagascar periwinkle (*Catharanthus roseus* (L.) G. Don; Apocynaceae) is one of the most studied plants displaying an active secondary metabolism (also called specialized metabolism), reaching up the status of “model non-model system” during the last decade<sup>1</sup>. *C. roseus* synthesizes a myriad of monoterpene indole alkaloids (MIAs) that have been proposed to mediate plant adaptation to the environment, especially during biotic interactions<sup>2</sup>. A good example of such a role relies on the proposed phytoanticipin function of strictosidine which upon leaf attack and membrane leakage can be enzymatically deglucosylated to form a highly reactive aglycone. This conversion induces a massive protein reticulation that was suggested to limit aggressor attacks, the so-called “nuclear time-bomb” process<sup>3</sup>. Other studies have also established the toxicity of several MIAs against pests or herbivores but essentially by feeding experiments using high concentrations of selected MIAs or total leaf extracts<sup>4–6</sup>.

<sup>1</sup>Université François-Rabelais de Tours, EA2106 “Biomolécules et Biotechnologies Végétales”, Tours, France.

<sup>2</sup>Metabolic Integration and Cell Signaling Group, Plant Physiology Section, Department of CAMN, Universitat Jaume I, Spain. <sup>3</sup>Institut de Recherche sur la Biologie de l’Insecte, UMR 7261, CNRS/Université François-Rabelais de Tours, Tours, France. <sup>4</sup>Université d’Orléans, COST, Laboratoire de Biologie des Ligneux et des Grandes Cultures (LBLGC), EA 1207, USC1328 INRA, Orléans, France. <sup>5</sup>The John Innes Centre, Department of Biological Chemistry, Norwich NR4 7UH, United Kingdom. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to V.C. (email: vincent.courdavault@univ-tours.fr)



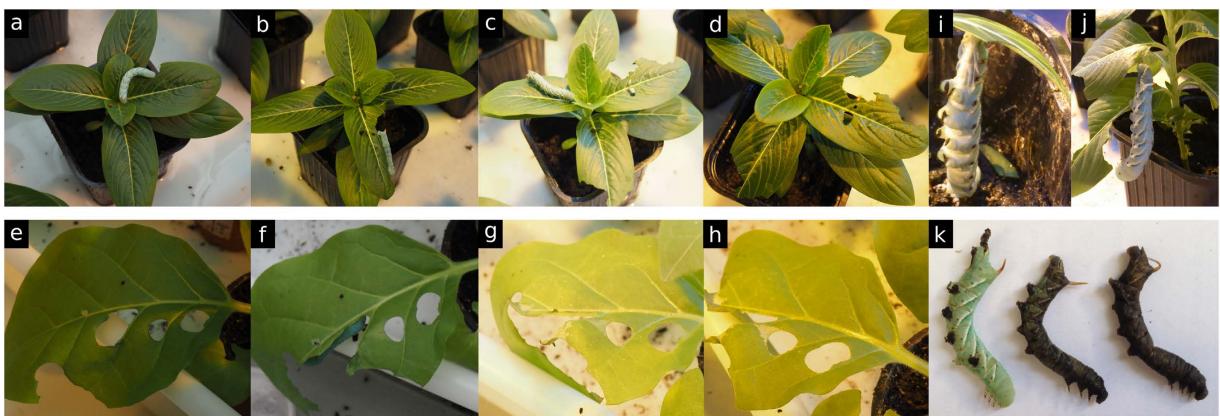
**Figure 1.** Biosynthetic pathway of MIAs in *C. roseus*. Simplified representation of the MIA biosynthesis in *C. roseus* leaves highlighting the cellular organization of the pathway in internal phloem associated parenchyma (blue rectangles), epidermis (pink rectangles), laticifers/idioblasts (yellow rectangles). Known single enzymatic steps are indicated by grey/red arrowheads and abbreviation of enzyme names. Broken grey/red arrows indicate unknown enzymatic steps. Conversion of tabersonine may occur in two ways to generate vindorosine (red arrows) or vindoline (grey). MEP, methyl-D-erythritol phosphate; GPPS, geranyl diphosphate synthase; GES, geraniol synthase; G10H, geraniol 10-hydroxylase; 10HGO, 10-hydroxygeraniol oxidoreductase; IO, iridoid oxidase; IS, iridoid synthase; 7DLGT, 7-deoxyloganic acid glucosyltransferase; 7DLH, 7-deoxyloganic acid 7-hydroxylase; LAMT, loganic acid O-methyltransferase; SLS, secologanin synthase; TDC, tryptophan decarboxylase; STR, strictosidine synthase; SGD, strictosidine  $\beta$ -D-glucosidase; T16H2, tabersonine 16-hydroxylase 2; 16OMT, 16-hydroxytabersonine O-methyltransferase; T3O, tabersonine 3-oxidase; T3R, tabersonine 3-reductase; NMT, 16-methoxy-2,3-dihydrotabersonine N-methyltransferase; D4H, desacetoxyvindoline 4-hydroxylase; DAT, deacetylvindeoline 4-O-acetyltransferase.

For several decades, these cytotoxic MIAs have been valorized as pharmaceutical compounds used to treat human diseases such as the antineoplastic vinblastine and vincristine inhibiting tubulin polymerization, or the antihypertensive ajmalicine<sup>7</sup>. However, their very low amounts *in planta* restrict their supply and have prompted the search for alternative production sources. In addition, MIA complex stereochemistry renders complete chemical synthesis uneconomical. To date, the dimeric MIAs vinblastine and vincristine used in anticancer treatments are produced by the chemical condensation of their monomeric precursors vindoline and catharanthine extracted from leaves of *C. roseus*<sup>8</sup> (Fig. 1). The recent development of bioengineering approaches based on the elaboration of MIA producing yeast strains following multiple gene transfer as well as on metabolic pathway expression in heterologous plants offer new alternatives but requires the elucidation of the whole MIA biosynthetic pathway<sup>9–11</sup>.

The synthesis of MIAs *in planta* relies on a complex route involving at least thirty enzymatic steps, characterized over 25 years and especially during the last 5 years<sup>12</sup> (Fig. 1). Basically, MIAs originate from the condensation of a monoterpene precursor, secologanin with an indole precursor, tryptamine, catalyzed by strictosidine synthase (STR). The resulting strictosidine is subsequently deglucosylated by strictosidine  $\beta$ -D-glucosidase (SGD) leading to the formation of the plethora of MIAs synthesized in *C. roseus* such as vindoline and catharanthine as well as additional scaffolds in other MIA producing plants<sup>2</sup>. *In planta*, the condensation of vindoline and catharanthine initiates the synthesis of dimeric MIAs through the formation of anhydrovinblastine that also displays cytotoxic properties<sup>13</sup>.

While MIAs accumulate in nearly all *C. roseus* organs, leaves constitute one of the main sites of accumulation and display a high diversity of MIAs with unique biosynthetic traits like the synthesis of vindoline and its demethoxylated derivative vindorosine<sup>6,14,15</sup>. The complexity in the number of steps is overlaid by the spatial organization of the pathway which is distributed in at least three different cell-types and five distinct subcellular compartments<sup>16</sup> (Fig. 1). This multi-site organization implies potential inter- and intra-cellular transport of metabolites but also involves the evolution of distinct enzyme isoforms harboring specific functions in MIAs synthesis, as exemplified with two recently identified secologanin synthase (SLS) isoforms displaying specific and complementary gene expression profiles within plant organs<sup>17</sup>. In the last years, the combination of RNA-seq based transcriptome analyses and gene validation procedures based on virus-induced gene silencing allowed the identification of these isoforms as well as other missing genes from this pathway<sup>18–24</sup>. However, as the pathway deciphering progresses, new complementary transcriptome resources are still required to facilitate and complete the identification of still uncovered enzymes or regulators of the pathway.

The use of contrasting physiological states is known to facilitate primary or specialized metabolism understanding as well as pathway discovery in phytochemical genomic, in particular through gene and metabolite

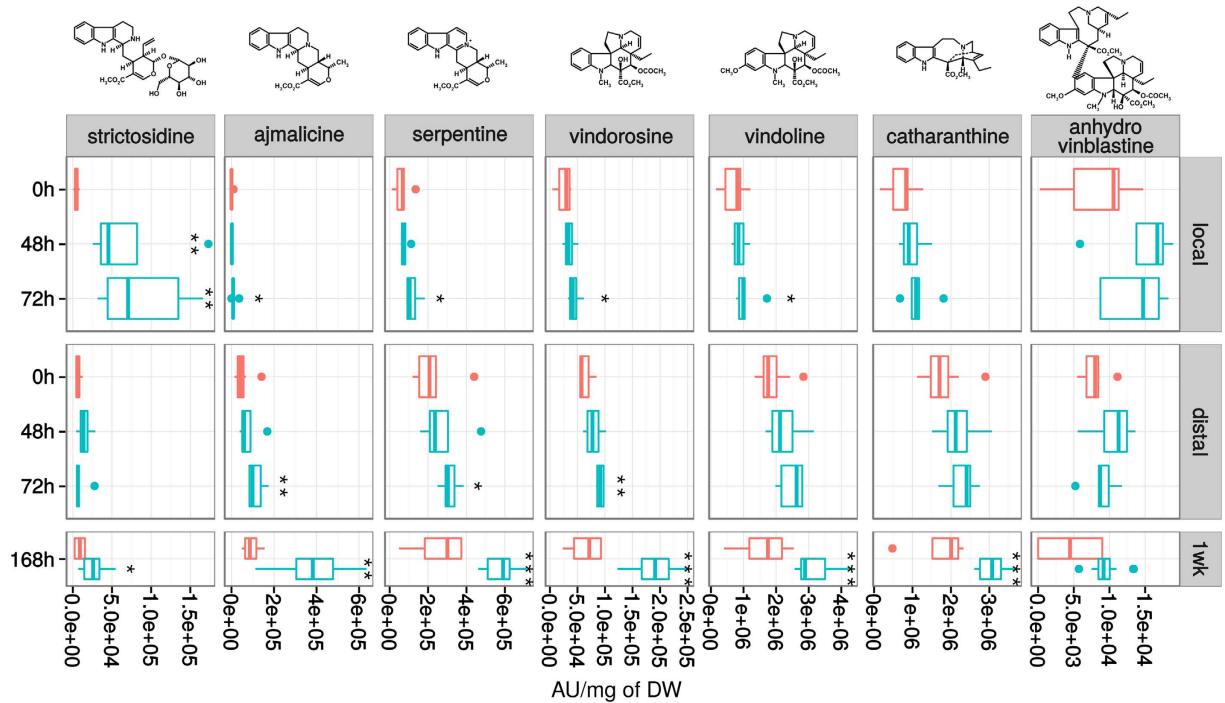


**Figure 2. *Manduca sexta* larvae feeding of *C. roseus* (a–d) and *N. tabacum* (e–h) leaves.** Pictures were taken 2 h (a,e), 24 h (b,f), 48 h (c,g) and 72 h (d,h) after placing caterpillars on leaves. (i,j) Typical morphologic alteration of caterpillars feeding on *C. roseus*, hanging by their softened end. k, browning phenotypes usually obtained after 72 h of feeding.

clustering analyses<sup>19</sup>. From this point of view, the biosynthesis of MIAs in *C. roseus* appears to be a tightly regulated process influenced notably by environmental factors and phytohormones. For instance, light and UV stimulate MIA production by promoting MIA biosynthetic gene expression and the corresponding enzymatic activities<sup>25–29</sup>. Furthermore, mechanical wounding engenders a similar positive effect while drought stress causes the decrease of the MIA content<sup>30,31</sup>. Plant hormones also exhibit pronounced and well documented effects such as the stimulatory properties of cytokinins and ethylene as well as the antagonist role of auxins on MIA biosynthesis<sup>32–34</sup>. On the other hand, jasmonic acid (JA) derivatives have been classically described to mediate plant responses to wounding, necrotrophic fungi and chewing insects, often in a synergistic way with ethylene<sup>35–37</sup>. In *C. roseus*, JA and its derivative methyl jasmonate (MeJA) also trigger the biosynthesis of MIAs mainly through the activation of octadecanoid-responsive Catharanthus AP2/ERF-domain transcription factors (ORCA)<sup>38</sup>. Those transcription factors are able to bind JA-responsive elements (JERE) in the promoters of MIA genes such as STR<sup>39,40</sup>. Interestingly, the JA signal also seems to mediate or potentiate the effects of some abiotic and biotic stimuli in *C. roseus* including light exposure or fungal extract treatments<sup>41,42</sup>. Such treatments with fungal extracts drastically increase MIA biosynthesis through the rapid stimulation of MIA biosynthetic gene expression including tryptophan decarboxylase (TDC) and STR in cell cultures and hairy roots of *C. roseus*<sup>43–45</sup>. In the light of those data, the use of biotic agents combining several of these signals could lead to a strong and robust modulation of the MIA pathway, thus constituting a valuable tool to decipher MIA metabolism<sup>46</sup>. This postulate is supported by the prominent results retrieved from the herbivory of the tobacco hornworm *Manduca sexta* on *Nicotiana* sp., enabling notably the elucidation of nicotine alkaloid metabolism<sup>47</sup>. While *C. roseus* can be challenged by different types of aggressors, only mollicute infections and especially phytoplasma have been shown to trigger MIA metabolism on whole plants through induction of gene expression<sup>48–51</sup>. To our knowledge, no other interaction between *C. roseus* plants and other bioaggressors has been characterized at the molecular level, may be due to the lack of easily propagated pests of this plant. Using the model *M. sexta/Nicotiana* sp. as a guideline, we developed a new plant-herbivore system based on the non-host interaction of *C. roseus* and *M. sexta* larvae, to investigate how *C. roseus* deploys MIA metabolism in response to herbivory. By combining targeted metabolic analyses and RNA sequencing, we demonstrated that folivory of *C. roseus* caused both local and systemic induction of MIA biosynthesis resulting from the induction of specific MIA biosynthetic gene expression.

## Results and Discussion

***M. sexta* larvae consumed *C. roseus* leaves and died.** To determine whether an efficient interaction can be established between *C. roseus* and *M. sexta*, caterpillars were placed on leaves of *C. roseus* and its host plant *N. tabacum*. In both cases, *M. sexta* larvae were able to consume substantial amounts of leaves during the first 2 h (Fig. 2a,b,e,f). However, independent choice experiments showed an expected and clear preference for *N. tabacum* (Supplemental Figure S1a). Feeding on tobacco leaves remained constant for the next 20 h but progressively decreased on periwinkle up to a total arrest usually observed from 72 h onwards (Fig. 2c,d,g,h). In this context, no weight gain was measured for larvae feeding on *C. roseus* in contrast to larvae feeding on the host plant *N. tabacum* (Supplemental Figure S1b). This was also accompanied by pronounced morphologic alterations (such as body softening; Fig. 2i,j) and intense browning followed by the death of caterpillars (Fig. 2k). Dissection of larvae revealed internal alterations/damages compared to control caterpillars fed on tobacco (Supplemental Figure S1c). Metabolic analyses of the larval gut also revealed the presence of multiple MIAs including the monomers vindoline and catharanthine and especially of their condensation product anhydrovinblastine. This qualitative MIA composition was quite similar to that found in leaves, thus confirming leaf ingestion. However, the dimer/monomer (anhydrovinblastine/catharanthine and vindoline) ratio was dramatically increased compared to those observed in the whole leaf (Supplemental Figure S1d) suggesting that the observed dimerization process contributes to caterpillar intoxication and may rely on the gathering of vindoline and catharanthine that were suggested to accumulate in different leaf compartments<sup>6</sup>. This differs from the MIA content of the intestinal tract of *Bombyx*



**Figure 3. Monoterpene Indole Alkaloid (MIA) accumulation in *C. roseus* leaves fed by *M. sexta*.**

MIA accumulation was monitored 48 h and 72 h after an initial 2 h feeding period, as well as in newly emerged leaves one week after (168 h). Asterisks denote significance levels at \* $p < 0.05$ , \*\* $p < 0.01$  and \*\*\* $p < 0.001$  (non parametric Wilcoxon rank sum test). Red boxes, control (independent, intact plants; see Supplemental Figure S2a); blue boxes, after feeding with *M. sexta*. AU, arbitrary unit normalized by sample mass.

*mori* fed with *C. roseus* leaf extracts, where guts of the insect only contained monomers<sup>6</sup>. Taken altogether, our results indicate that *M. sexta* is able to feed long enough on *C. roseus* despite of its toxic compounds to establish a model for analyzing metabolic and transcriptomic changes in *C. roseus*.

**MIA metabolism is induced by herbivory.** Leaf consumption by *M. sexta* has been reported to induce the biosynthesis of several defense compounds in *Nicotiana* sp around three days post-attack<sup>52</sup>. We therefore analyzed the MIA composition of damaged *C. roseus* leaves in a similar time range. Leaves were challenged for 2 h with larvae and MIA accumulation was monitored at 48 h and 72 h post-feeding in both locally damaged and distal non damaged leaves, as well as in leaves newly emerged 1 week (168 h) after attack and finally compared to the MIA content of control leaves that had not encountered any herbivory and that were sampled accordingly (Supplemental Figure S2a). Relative quantification was performed on the main MIAs accumulating in leaves (catharanthine and vindoline) but also on minor compounds including ajmalicine, serpentine, anhydrovinblastine, vindorosine and strictosidine. In damaged and distal leaves, slight but significant changes ( $p < 0.05$ ) were observed 48 h and 72 h post attack except for ajmalicine, serpentine, vindorosine and vindoline (Fig. 3). The largest increase was observed for strictosidine at 48 h and 72 h in local leaves ( $p = 0.003$  and  $p = 0.001$  respectively). This latter was present at trace levels in leaves of control plants ( $4,549 \pm 2,863$  AU) but strongly accumulated 48 h and 72 h post-attack (up to  $\times 15$  and  $\times 19$  respectively in locally damaged leaves). By contrast, only minor increases of strictosidine were monitored at such times when leaves were mechanically wounded suggesting that strictosidine accumulation was specifically induced by the biotic interaction (Supplemental Figure S2b). In addition, in newly emerged leaves 1-week post attack, all MIAs except anhydrovinblastine were significantly enhanced suggesting that a systemic signal induced by herbivory may have triggered the accumulation of MIAs as a probable defense response. The absence of anhydrovinblastine increase in young leaves was expected since it has already been shown to exclusively accumulate in older leaves<sup>6</sup>. Interestingly, only a moderate increase of strictosidine (1.8 times more accumulated;  $p = 0.02$ , Wilcoxon rank sum test) was observed in newly emerged leaves that might reflect the consumption of this compound to allow the dramatic increase in the biosynthesis of the downstream MIAs including catharanthine, vindoline or anhydrovinblastine. Based on these results, one could speculate that *C. roseus* set up two distinct responses to herbivory, a local and quickly induced one relying on strictosidine accumulation which may cross link proteins following its deglucosylation catalyzed by SGD after membrane leakage<sup>3</sup>, and a systemic and long-term mechanism involving a higher accumulation of toxic MIAs in newly developing organs.

**Herbivory of *C. roseus* leaves led to a marked transcriptional reprogramming.** As described for several specialized metabolisms, increase in MIA biosynthesis in *C. roseus* is usually preceded by the activation of the corresponding biosynthetic genes<sup>33</sup>. Moreover, we already reported that MIA biosynthetic genes also respond

Sample name	Sample description	Sample accession	Read counts	Percent of reads mapped on CDF97
Ms6h	Manduca damaged leaves 6 h	ERR1512369	23,772,392	98.81%
Ctrl6h	Leaves from control plants 6 h	ERR1512370	27,591,053	98.82%
Ms8h	Manduca damaged leaves 8 h	ERR1512371	27,774,796	98.93%
Ctrl8h	Leaves from control plants 8 h	ERR1512372	34,082,314	98.87%
MsDamaged24h	Manduca damaged leaves 24 h (damaged part of the leaf)	ERR1512372	31,252,515	98.88%
MsIntact24h	Manduca damaged leaves 24 h (intact part of the leaf)	ERR1512372	18,495,538	98.84%
Ctrl24h	Leaves from control plants 24 h	ERR1512372	30,061,186	98.97%
Ms1wk	New leaf from Manduca damaged plant after 1 week	ERR1512376	27,394,833	98.63%
Ctrl1wk	New leaf from control plant after 1 week	ERR1512376	27,226,966	98.68%

**Table 1.** Description of samples deposited in EBI ENA under accession number PRJEB14626.

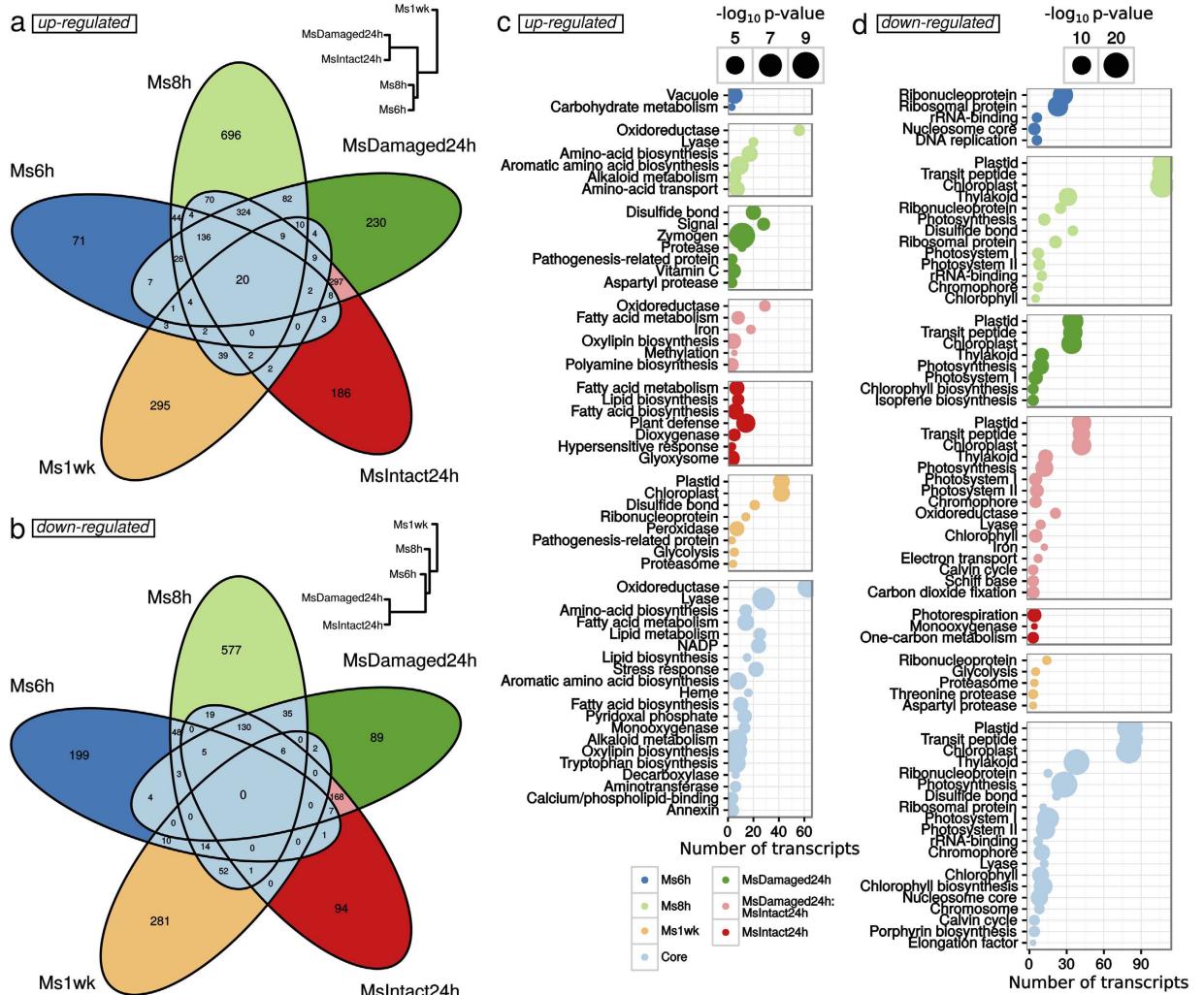
Compared conditions	Time	Up-regulated (Log2 fold change > 0)	Down-regulated (Log2 fold change < 0)
Ms6h/Ctrl6h	6 h	333	291
Ms8h/Ctrl8h	8 h	1,470	890
MsDamaged24h/Ctrl24h	24 h	1,171	449
MsIntact24h/Ctrl24h	24 h	1,072	431
Ms1wk/Ctrl1wk	1 week	402	366

**Table 2.** Number of differentially accumulated transcripts.

to fungal elicitors and hormones by an enhanced expression from 8 h to 24 h post-treatment<sup>53,54</sup>. We therefore analyzed the transcriptional reprogramming of *C. roseus* leaves subjected to herbivory in four independent experiments following this timeline (Supplemental Figure S2c, Table 1). In all experiments, caterpillars were allowed to feed for 2 h and subsequently removed, indicating the beginning of the kinetics. Total RNA were extracted from Manduca-damaged (Ms) and control (Ctrl, intact plants of the same age and physiological status) leaves at 6 h (Ms6h and Ctrl6h), 8 h (Ms8h and Ctrl8h) and 24 h after the feeding period. In this latter time, leaves were split along the midrib to analyze the damaged (MsDamaged24h) and the intact (MsIntact24h) halves separately. In addition, we analyzed the systemic responses in intact and newly formed leaves obtained 1 week after initial consumption of plant leaves (Ms1wk) and in non-attacked control plants (Ctrl1wk). These 9 samples were then sequenced with the Illumina technology in a paired-end design at an average sequencing depth of 27 million of reads per sample. The resulting reads were pseudo-aligned at a very good rate (>98%) to the CDF97 reference transcriptome<sup>17</sup> for *C. roseus* to estimate the abundance of each transcript with Salmon<sup>55</sup>.

At each time, log<sub>2</sub> of Transcript Per Million (TPM) fold changes between the attacked leaf and the control leaves were calculated (Ms6h/Ctrl6h, Ms8h/Ctrl8h, MsDamaged24h/Ctrl24h, MsIntact24h/Ctrl24h and Ms1wk/Ctrl1wk). Differentially expressed genes were identified in the 5 comparisons at a *p*-value < 0.001 without setting a cut-off threshold for log<sub>2</sub> fold changes because of slight differences in attacked areas between samples. The number of transcripts accumulating at higher or lower amounts in attacked leaves compared to control leaves (considered as up-regulated and down-regulated genes respectively) considerably changed over time (Table 2). Overall, changes were more important at 8 h (a total of 2,360 transcripts) and 24 h (1,620 and 1,503 for MsDamaged24h and MsIntact24h respectively). This demonstrated that the intensity of transcriptional responses to herbivory reached a maximum at 8 h and 24 h as previously observed with plant hormone treatments. By contrast, only 333 and 402 up-regulated transcripts and 291 and 366 down-regulated ones were retained for shorter (Ms6h) and prolonged responses (Ms1wk), respectively. These results showed that challenging leaves of *C. roseus* with a chewing insect led to a transcriptional response at least 6 h after initial feeding and to a systemic transcriptome reprogramming in newly developed leaves.

To gain more insights into leaf responses to *M. sexta* herbivory, the lists of differentially expressed transcripts (up-regulated and down-regulated) obtained from the 5 comparisons (attacked vs control) were cross-compared with a Venn diagram analysis to evaluate similarities and specificities of each (Fig. 4a, Supplemental Figure S3). Concerning up-regulated transcripts, a total of 2,588 were more highly expressed in attacked leaves in at least one comparison. It is noteworthy that this number includes a certain degree of redundancy between transcripts, owing to the reference transcriptome which was specifically built to capture isoform complexity in *C. roseus*<sup>17</sup>. Out of these 2,588 transcripts, we defined 5 specific transcript sets induced at each sampling time corresponding to 21% of the up-regulated transcripts for Ms6h (Fig. 4a, dark blue, 71 transcripts over 333), 47% for Ms8h (light green, 696 transcripts over 1,470), 19% for MsDamaged24h (dark green, 230 transcripts over 1,171), 17% for MsIntact24h (red, 186 transcripts over 1,072) and 73% for Ms1wk (orange, 295 transcripts over 402). In addition, we also defined a core set of 813 transcripts (light blue, 31% of the 2,588 transcripts) found in at least 2 different sampling times, including 20 transcripts (0.7%) commonly found in the 5 comparisons (Fig. 4a). In the light of this distribution, largest specificities were observed at 8 h and 1 week after attack in newly emerged leaves reflecting once again the previously measured modification of the MIA content.



**Figure 4. Functional classification of differentially expressed transcripts in leaves of *C. roseus* fed by *M. sexta* compared to undamaged control leaves.** For each of the 5 comparisons ((Ms6h/Ctrl6h, Ms8h/Ctrl8h, MsDamaged24h/Ctrl24h, MsIntact24h/Ctrl24h and Ms1wk/Ctrl1wk), lists of significantly ( $p < 0.001$ ) up-regulated (a,c) or down-regulated transcripts (b,d) were analyzed with Venn diagrams (a,b). In a and b, hierarchical clustering trees depict dissimilarities between samples calculated on the 2,588 and 1,745 transcripts that were respectively significantly up-regulated or down-regulated in at least one comparison. Intersections between these lists were used to identify core sets of herbivory-modulated transcripts (modulation at least at two different sampling times) clear blue sectors) and sample-specific transcripts (other sectors). The pink sectors correspond to transcripts commonly modulated both in MsDamaged24h and MsIntact24h and were therefore not included in the core sets. (c,d) UniProt keyword classification of lists of differentially accumulated transcripts. The number of transcripts is indicated for each term by a dot in which the diameter is proportional to the  $p$ -value of enrichment testing (Hypergeometric distribution). Only keywords attributed to more than 2 transcripts and that were significantly enriched (Hypergeometric distribution,  $p$ -value  $< 0.001$ ) are represented. Lists of up- and down-regulated transcripts are available in Supplementary Tables S1 and S2 respectively. Details for UniProt keyword analysis are presented in Supplementary Table 3.

Besides up-regulated genes, a lower number of transcripts (1,745) were found to be down-regulated in response to attack in at least one comparison. In this case, overlaps between differentially expressed transcripts were less pronounced than observed for up-regulated ones (Fig. 4b). Indeed, only 337 transcripts (20%, light blue) were common to at least 2 sampling times and were defined as the core set of down-regulated transcripts during herbivory. The Ms8h and Ms1w specific transcript sets (light green and orange, respectively) were once again the most divergent with 577 and 281 down-regulated transcripts, reinforcing thus the prominence of these two conditions in response to folivory.

Finally, out of the 4,333 (2,588 + 1,745) transcripts whose expression was altered upon leaf herbivory, we were able to find homologies (Blastx,  $e$ -value  $< 1e-10$ ) for 3,189 transcripts (73%) with UniProt proteins. This mapping was used to retrieve PFAM and UniProt keyword information in order to classify the functions characterizing the *C. roseus* responses to *M. sexta*. The next paragraphs focus on the main functions identified in the core sets

of up- and down-regulated transcripts but also in specific transcript sets according to the Venn diagram analysis (Fig. 4). All information on transcripts and functions are available in Supplemental Tables S1, S2, S3 and S4.

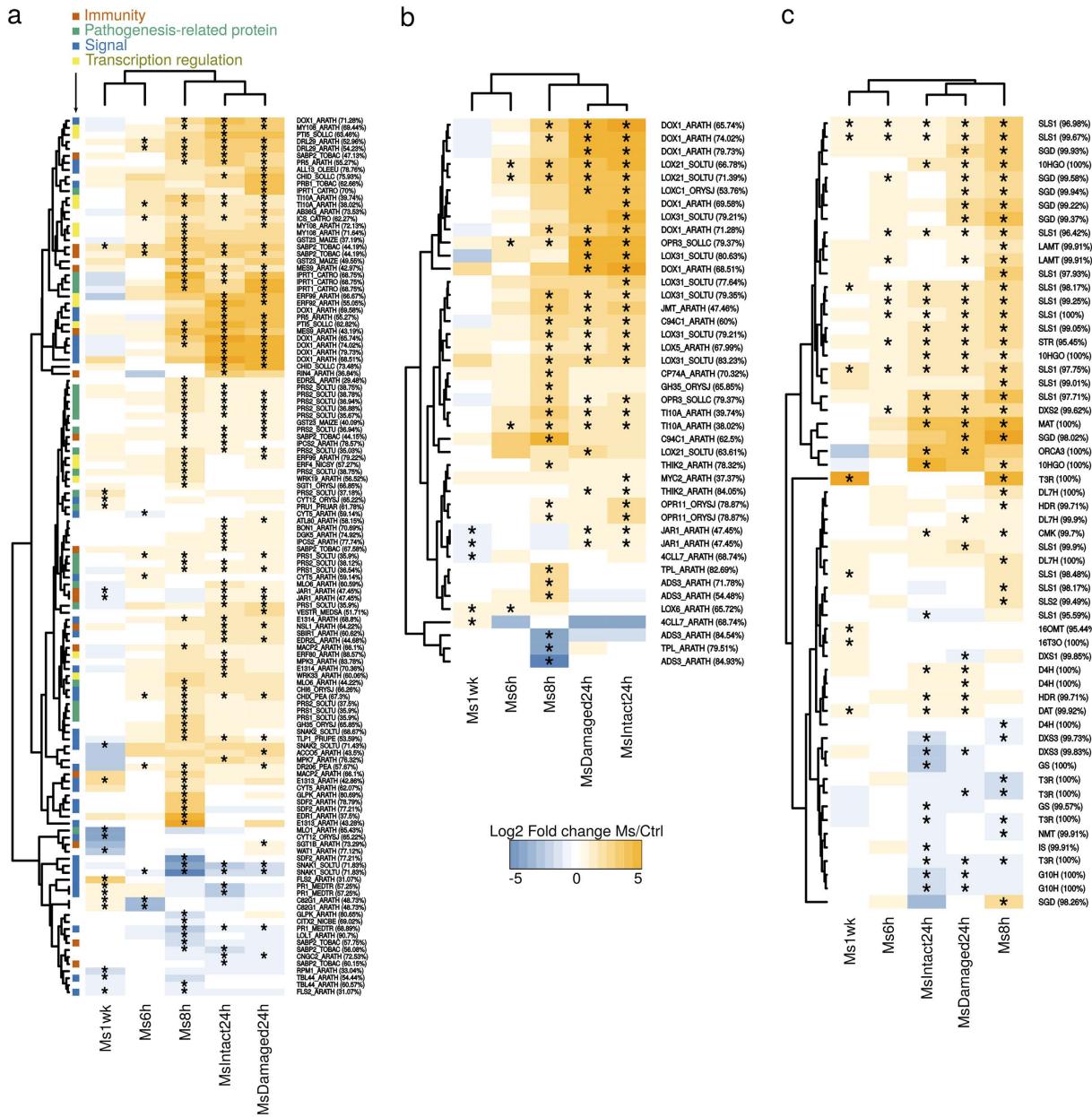
**C. roseus leaf consumption induces a potential photosynthesis breakdown.** Analysis of the core set of down-regulated transcripts revealed a strong representation of photosynthesis-related elements. This was the most striking feature of this set of genes. Indeed, many Uniprot keywords related to this process (e.g. Photosynthesis, Chlorophyll biosynthesis) were significantly enriched ( $p$ -value  $< 1e-10$ ) within transcripts from this core set (Fig. 4d, Supplementary Table S3 and Supplemental Figure S3b). Accordingly, the PFAM domain PF00504.18 Chlorophyll A/B binding protein was also well represented (24 transcripts) although not significantly enriched ( $p$ -value = 0.09; Supplementary Table S4). Interestingly, a similar trend was also detected in transcripts specifically associated with Ms8h, MsDamaged24h and MsIntact24h (Fig. 4d). This apparent alteration in photosynthesis-related processes might reflect a reallocation of cell activities towards other processes including defense. Such an effect has been already described in several plant herbivore interactions and particularly for *M. sexta*-mediated herbivory<sup>56,57</sup>. In addition, this down-regulation of photosynthesis has been shown to be associated with the production of JA in fed parts<sup>58</sup>. Moreover, many pentatricopeptide-repeat (PPR) containing proteins were also down-regulated at the different sampling times (Supplemental Figure S3b). These proteins were previously shown to be involved in organelle biosynthesis<sup>59</sup>. In our case, they might be linked to potential modifications in chloroplast biogenesis and associated to the proposed decrease in photosynthesis.

**Herbivory caused induction of PR-proteins, terpene synthases and JA metabolism.** Within our defined core set of 813 up-regulated genes, many transcripts related to known defense processes were identified (Supplemental Tables S3 and S3). For instance, it contained 17 transcripts related to PFAM domain PF00407.16 (Pathogenesis-related protein Bet v I family, from the PR-10 family). In particular, transcripts homologous to *Solanum tuberosum* pathogenesis-related protein STH-1 and 2 (PRS1\_SOLTU and PRS2\_SOLTU) and a previously identified probable intracellular pathogenesis-related protein T1 from *C. roseus* (IPRT1\_CATRO) were strongly induced in fed leaves (Figs 4c and 5a). A transcript encoding Pathogenesis-Related protein 5 (PR-5) was also identified. Interestingly, these genes types were also over-represented in specific gene sets and notably in the 24 h samples. The Uniprot keyword 'Pathogenesis-related protein' was significantly represented in the MsDamaged24h specific gene set ( $p = 0.0003$ ) and the PFAM domain PF00407.16 in the MsDamaged24h and MsIntact24h intersection set ( $p < 1e-05$ ) (Supplemental Tables S3 and S4). Interestingly, cysteine proteinase inhibitors homologous to CYT5\_ARATH and CYSP\_PEA known to be induced by herbivory and to inhibit insect proteolytic enzymes<sup>60</sup> were mostly found in these lists, in particular in the MsDamaged24h specific gene set but also in the core set of common transcripts such as the cystatin CYTI\_VIGUN (Supplemental Tables S3 and S4). This strong expression of defense proteins during herbivory indicated that *C. roseus* deployed defense responses complementary to MIA biosynthesis.

Aside from these genes related to direct defenses, several up-regulated transcripts suggested a recruitment of indirect defenses through modulations of terpene metabolism. This was illustrated by terpene synthase homologs including *Vitis vinifera* (-)-germacrene D synthase (TPSGD\_VITVI), *Quercus ilex* Myrcene synthase (MYRS\_QUEIL), *Malus domestica* (E,E)-alpha-farnesene synthase (AFS1\_MALDO), *Ricinus communis* Alpha-farnese synthase (TPS7\_RICCO) and *Fragaria ananassa* nerolidol Synthase (NES1\_FRAAN) catalyzing the synthesis of monoterpenes and sesquiterpenes. Such an activation could result in a *de novo* production of volatile terpene compounds upon herbivory that have been described to limit insect attacks by mediating attraction of parasitoids<sup>61</sup>. This apparent stimulation of sesquiterpene metabolism may also suggest modulations of leaf triterpene content, specifically ursolic acid, which has anti-insect feeding activity and accumulates at high level in *C. roseus* leaves<sup>62,63</sup>. However, only one transcript related to triterpene metabolism and encoding a putative squalene synthase from *N. benthamiana* (FDFT\_NICBE), was significantly up-regulated in the Ms1wk specific set.

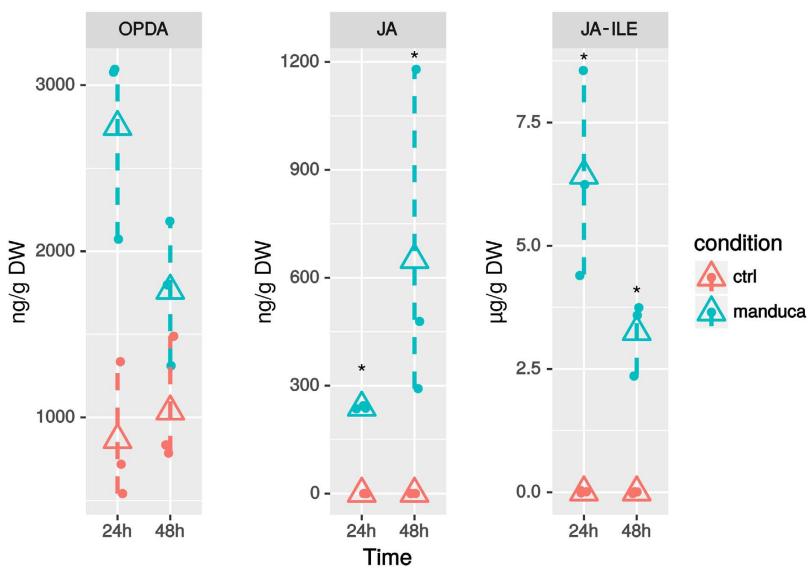
The second feature of the core set of transcripts up-regulated upon herbivory corresponded to the marked association with JA biosynthesis and signaling as revealed by the significant enrichments of 'Lipid metabolism' (25 transcripts,  $p = 1e-04$ ) and 'Oxylipin biosynthesis' terms (7 transcripts,  $p = 1e-07$ ) (Fig. 4c). For example, we observed a high expression of transcripts homologous to *S. tuberosum* Linoleate 13S-lipoxygenase 3-1 (LOX31\_SOLTU), *Arabidopsis thaliana* Linoleate 9S-lipoxygenase 5 (LOX5\_ARATH), *A. thaliana* Allene oxide cyclase 4 (AOC4\_ARATH; although this latter was not annotated in Uniprot as related to JA biosynthesis), *Solanum lycopersicum* 12-oxophytodienoate reductase 3 (OPR3\_SOLLC), *A. thaliana* Jasmonate O-methyltransferase (JMT\_ARATH) and *A. thaliana* MYB108 transcription factor (MY108\_ARATH) (Fig. 5b). In addition, 7 transcripts with a predicted Tify domain were also up-regulated (PF06200.11; Fig. 5b, Supplemental Table S4). Tify-domain containing proteins are known JA-related transcription factors like TI10A\_ARATH, reinforcing the probable activation of the JA signaling pathway upon feeding<sup>64</sup>. This was further illustrated by the presence of transcripts displaying putative AP2 domains and homology to the ORCA3 transcription factor, a key component of the JA-induced MIA biosynthesis<sup>39</sup>. Interestingly, orthologs of the jasmonate biosynthesis genes AOC4\_ARATH and LOX5\_ARATH were also found in the 20 genes common to the five gene lists, thus highlighting the prominence of JA biosynthesis in the periwinkle response to herbivory. This phenomenon was confirmed by quantification of oxophytodienoic acid (OPDA), JA and its active form jasmonoyl-isoleucine (JA-Ile) in *C. roseus* leaves, 24 h and 48 h after their intial consumption by *M. sexta* (performed during 2 hours as for other analyses; Fig. 6). In both cases, higher amounts of these compounds were detected in damaged leaves showing that synthesis of JA was strongly triggered during herbivory. Such JA accumulation in attacked leaves has been already observed during *Nicotiana attenuata/M. sexta* interaction<sup>65</sup> and suggests that the *C. roseus/M. sexta* interaction induces a representative response of plants to chewing insects.

In addition to JA, our analysis of *M. sexta*-induced gene expression in *C. roseus* also indicated the activation of several elements related to ethylene biosynthesis and signaling. The core set of transcripts contained a



**Figure 5. Thematic representation of *C. roseus* transcripts related to defense, jasmonate and MIA metabolisms.** Transcripts with significant log<sub>2</sub> fold changes (fed vs control) in at least one comparison were clustered according to their expression profiles. (a,b) Transcripts associated with “Plant defense” and “Oxylipin”/“Jasmonic acid” Uniprot keywords respectively. Transcripts from CDF97 assembly were annotated by searching homologies (Blastx, *e*-value < 1e-10) against the Uniprot database (the % of identity is indicated between brackets). Transcripts associated with the keyword “Plant defense” were further grouped according to the other keywords they were annotated with, i.e. Pathogenesis-Related proteins, Signal, Immunity and Transcription regulation. (c) Transcripts homologous to known MIA genes (blast score > 1100, %id > 95%). Asterisks indicate that the expression of the corresponding transcript significantly differed between the attacked and the control samples (Linear model, *p*-value < 0.001).

transcript homologous to 1-aminocyclopropane-1-carboxylate oxidase (ACCO\_ACTDE), also known as the ethylene-forming enzyme<sup>66</sup>. Furthermore, we observed the up-regulation of a MAPKK (M2K9\_ARATH) related to the MPK3/MPK6 signaling pathway that leads to the phosphorylation of ETHYLENE INSENSITIVE3 (EIN3) to trigger its activation resulting in the transactivation of its target genes<sup>67</sup>. Consistently, transcription factors related to ethylene signaling were also found in the same lists such as ERF08\_ARATH and RAP24\_ARATH. In this respect, we found that the regulatory motif GCCGC(C/G) was significantly enriched in the promoters of the intersections between lists of up-regulated transcripts at 8 and 24 h (Supplemental Figure S4a). This motif was reported to be an important binding site of ERF transcription factors<sup>68</sup>. Such activation could illustrate the



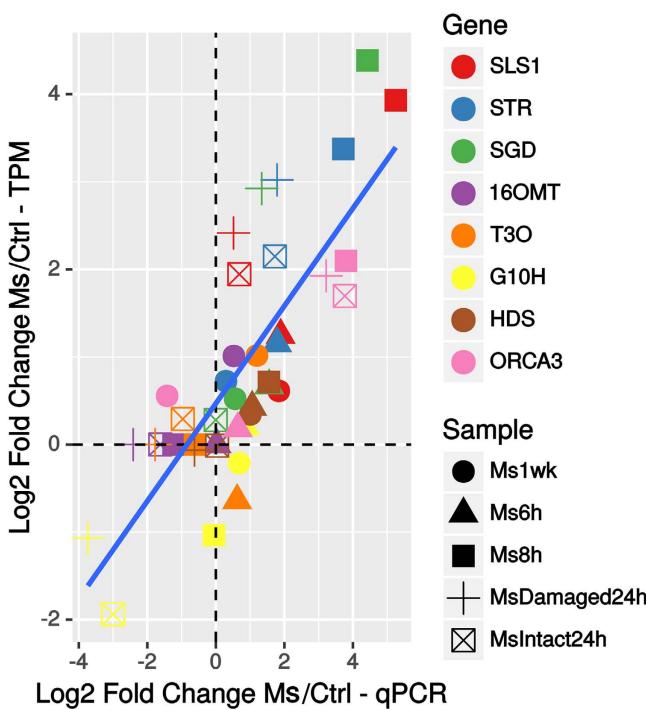
**Figure 6.** Oxylipin accumulation in *C. roseus* leaves fed by *M. sexta* after 24 and 48 h. JA, jasmonic acid; OPDA, oxo-phytodienoic acid; JA-Ile, Jasmonoyl-isoleucine. Triangles indicate arithmetic mean and bars correspond to confidence limits (non-parametric bootstrap). Asterisks denote significant differences between fed and control leaves (Wilcoxon rank sum test,  $p$ -value  $< 0.05$ ,  $n = 3$ ).

ethylene burst occurring after folivory by *M. sexta* that could be involved in the cross talk with jasmonate signaling known to mediate plant responses to herbivore interactions<sup>35,36</sup>.

**Herbivory sequentially activated two sets of MIA biosynthetic genes.** Consistent with the huge modification of the MIA content of *C. roseus* plants challenged with *M. sexta* larvae (Fig. 3), the expression of transcripts associated with MIA metabolism was strongly altered. The Uniprot keyword “Alkaloid metabolism” was significantly (7 transcripts,  $p = 1e-08$ ) enriched in the core set of up-regulated transcripts (Fig. 4c, Supplemental Tables S1 and S3). This set also contained transcripts representing 3 steps (TRPE\_ARATH, TRPD\_ARATH and TRPA2\_ARATH) out of 5 for tryptophan biosynthesis (pathway significantly enriched,  $p = 1e-07$ ) which leads to tryptamine, the indole MIA precursor. Many of the previously known MIA biosynthetic genes were similarly identified in the global gene lists (Fig. 5c).

Interestingly, these transcripts displayed a 2-step induction profile directly correlated with the variation of the leaf MIA content. The pronounced accumulation of strictosidine observed 48 h and 72 h after leaf attack was indeed preceded by a strong induction of the epidermis gene set ensuring LAMT, SLS1, STR and SGD gene expression. Indeed the 4 corresponding transcripts were commonly found upregulated in Ms6h, Ms8h and MsDamaged24 h (Fig. 5c). Such induction of gene expression was confirmed by qPCR analyses as illustrated for SLS1, STR and SGD (Fig. 7). Interestingly, expression of genes related to preceding reactions of biosynthesis localized in IPAP cells (from the MEP pathway genes to 7DLH) did not display such high induction suggesting a restriction of the response to epidermis expressed genes (Figs 1 and 5c). Similarly, no marked induction of genes catalyzing later steps of the MIA pathway were observed in agreement with the absence of vindoline, catharanthine or ajmalicine induction 72 h after leaf attack (Figs 2 and 5C). On the other hand, the induction of STR expression can be explained by the increase in JA biosynthesis probably leading to the induction of ORCA3 expression that has been previously described to transactivate the STR promoter by binding to conserved motif in its target gene promoters<sup>40,69</sup>. We found that this motif, CACGTG was significantly ( $e$ -value = 2e-008) represented in the promoters of the core set of induced genes thus confirming the prominent role of ORCA3 in plant response to herbivory (Supplemental Figure S4b). Interestingly, we also observed the local induction of the MATE transporter (CRO\_T006097 – encoded by SRR342023\_TR31426\_c4\_g1\_i1\_len = 2080) located close to STR in the *C. roseus* genome, which may reinforce its proposed involvement in the transport of MIA precursors<sup>70</sup>. Altogether, the concomitant dramatic accumulation of strictosidine and strong induction of SGD (Figs 3 and 5c) are in good agreement with the “nuclear time-bomb” defense system hypothesis. The potential massive formation of the reactive strictosidine aglycon upon deglycosylation ensuing might cause protein reticulation limiting leaf digestibility and/or damages in the early phase of the infestation as observed on *M. sexta* dissected guts (Fig. 2, Supplemental Figure S1c).

The second step of the MIA metabolism variation was observed in newly developed leaves, one week after attack (Ms1wk) and was characterized by the induction of a distinct set of genes (Fig. 4a, Fig. 5c, Supplemental Table S1). While the induction of expression of SLS, STR and SGD was not sustained, up-regulation of downstream genes of the pathway was observed and notably for those involved in vindoline and vindorosine biosynthesis including 16OMT (for vindoline only), T3O and T3R. This induction, validated by qPCR analysis, might have caused the increase of the vindoline amount in newly developed leaves reflecting the systemic response deployed by *C. roseus* against *M. sexta* (Fig. 7). While no molecular explanation of the increase of the



**Figure 7. Comparison of expression levels (log<sub>2</sub>) obtained by qPCR and RNA-seq (Transcripts per Million) for candidate genes.** A regression line (blue) and its confidence intervals (95%, shaded) are depicted.

other MIAs can be provided now, it is tempting to hypothesize that some of the missing biosynthetic genes could be retrieved from this analysis.

## Conclusion

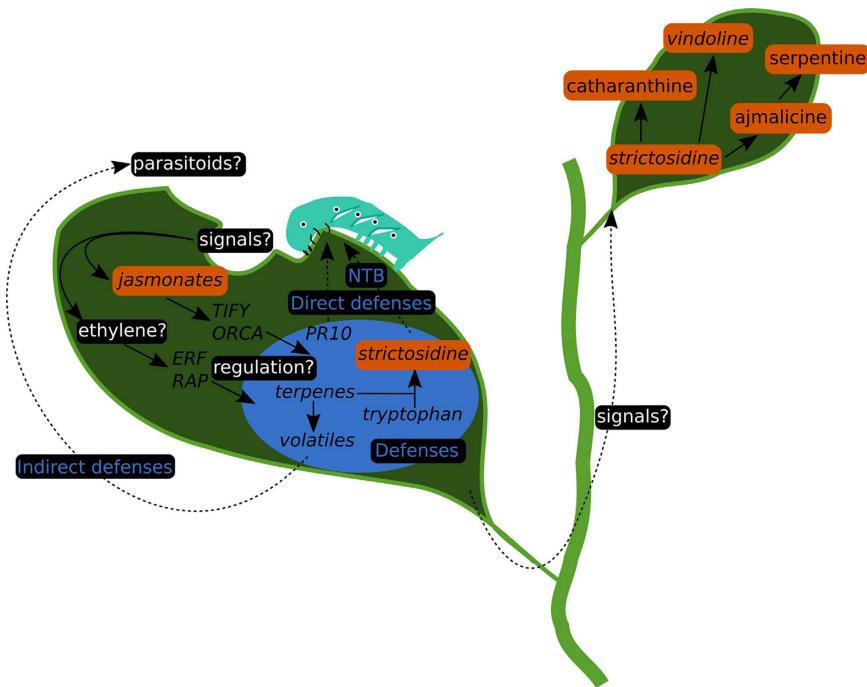
Despite the fact that MIA biosynthesis and toxicity have been studied for more than 50 years, still little is known about the role of these compounds in Madagascar periwinkle against biotic attacks. By studying the interaction between *M. sexta* and *C. roseus* using targeted metabolic and transcriptome analyses, we provided compelling evidence of the activation of the JA and ethylene signaling pathways and of distinct plant defense processes (Fig. 8). One of the main features of the folivory response was the local and systematic activation of MIA metabolism that relied on the induction of specific MIA biosynthetic gene subsets and differential MIA production (Figs 3 and 5c). The huge local accumulation of strictosidine combined with SGD expression may illustrate the importance of the ‘nuclear time bomb’ mechanism and the reactive strictosidine aglycone as a first barrier against aggressors whilst the synthesis of downstream MIAs in newly developing leaves may ensure an enhanced protection to overcome later attacks. In this context, leaf consumption resulted in quick and marked effects on caterpillar health status leading to death (Fig. 2i–k). Interestingly, analyses of the intestinal tracts of dying larvae compared to leaves revealed the main presence of anhydrovinblastine suggesting that catharanthine and vindoline dimerization only occurred after leaf consumption (Supplemental Figure S1d). Finally, the strong induction of MIA biosynthetic gene expression upon herbivory suggests that our RNA-seq data will constitute a new valuable resource to pursue the identification of missing genes involved in these complex metabolic pathways.

## Methods

**Feeding of *M. sexta* larvae on *C. roseus* leaves.** Seeds of *C. roseus* (Apricot Sunstorm cultivar, B and T world seeds, Aigues Vives, France) were germinated in a greenhouse and the resulting plants were grown in individual pot at 28 °C under a 16 h light/8 h dark cycle for eight weeks. *M. sexta* larvae were reared on artificial diet at 27 °C under 16 h light/8 h dark photoperiod and 70 ± 5% relative humidity until reaching the 3<sup>rd</sup> instar. For transcriptomics and metabolic analysis young caterpillars were laid on leaves in the morning and allowed to feed for 2 hours before being removed as depicted in Supplemental Figure S2a–c.

**Weight gain measurements, choice experiment and larvae analysis.** For the weight gain experiment, larvae were allowed to feed continuously on *N. tabacum* and *C. roseus* plants at 23 °C. Individual caterpillars were identified, initially weighted and laid on individual plants. Data were recorded 24, 48, and 72 h after the beginning of the experiment. For choice experiment, larvae were fed with 2.2 cm-diameter leaf disks of *C. roseus* and *N. tabacum* placed on wet filter paper in 15 cm-diameter Petri dishes. Leaf disk consumption was monitored during 120 min and the experiment was repeated twice.

**RNA extraction and sequencing.** RNAs were extracted from leaves with Trizol (Life Technologies) following the manufacturer’s recommendations with slight modifications. After precipitating with isopropanol and



**Figure 8. Molecular events associated with folivory in *C. roseus*.** Orange boxes, supported by compound measurements, italics, supported by gene expression measurements. Potential signals in the oral secretions of *M. sexta* together with damage-associated molecular patterns are likely to activate jasmonate and signaling pathways in *C. roseus*. Those two pathways might respectively use TIFY/ORCA and ERF/RAP transcription factors to control the activation of sets of potential defenses (within blue circle). Strictosidine may interfere with caterpillars by acting as a Nuclear Time Bomb (NTB). NBT involves a massive production of strictosidine aglycone, an efficient protein-cross-linker, by the nucleus-localized SGD, following vacuole disruption and the resultant release of strictosidine. Still unknown signals may control the increased biosynthesis of MIA in distal, newly formed leaves.

washing with 70% ethanol, RNA pellets were re-suspended in 100 µL of RNase free water and remaining sugars were precipitated by adding 10% ethanol (final concentration) and incubating 5 minutes at 4 °C. The supernatant obtained after centrifugation (5 min at 15,000g) was further precipitated by addition of 0.1 volume of 3 M sodium acetate pH 5.2 and 2.5 volume of 100% ethanol for 2 h at -20 °C. The tubes were centrifuged 15 minutes at 12,000 g and 4 °C and the resulting pellet was washed with 70% ethanol and re-suspended in RNase-free water. RNA concentration was estimated with a Nanodrop spectrophotometer (Thermo). A total of 9 transcriptomes (*Ms6h/Ctrl6h*, *Ms8h/Ctrl8h*, *MsDamaged24h/Ctrl24h*, *MsIntact24h/Ctrl24h* and *Ms1wk/Ctrl1wk*) were sequenced as single replicates by Eurofins Genomics using the Illumina HiSeq2000/2500 technology. Samples were sequenced in the paired-end mode (2 × 100 pb). The resulting fastq files were cleaned with Trimmomatic with default parameters (using TruSeq3 primer sequence). For quantification of transcript accumulation, reads were pseudo-aligned on CDF97 reference transcript sequences<sup>17</sup> and counted with Salmon<sup>55</sup> in the variational bayesian optimized (-vbo) quasi-mapping mode with bias correction (-biasCorrect). The resulting quant.sf files were combined and processed with R. Differentially accumulated transcripts were identified by considering each experiment as unique (without replicate) and by fitting a linear model to each gene with the 'exactTest' function of the edgeR Bioconductor package<sup>71</sup>. Biological variability was estimated by setting the square-root dispersion at 0.4. To balance the absence of biological replicates, differentially expressed genes were set if the *p*-value in the test was below 0.001. In addition, functions altered in response to folivory were focused on genes that were significantly modulated in at least two comparisons.

**Annotation and term enrichment analysis.** The CDF97 reference transcriptome was annotated with the Trinotate v3.0 pipeline against Uniprot (Blastx and Blastp on Transdecoder predicted ORFs) and PFAM (hmmscan) databases. Uniprot keywords were retrieved by using Uniprot predicted homologs in CDF97. Enrichment tests of functional terms were performed by comparing effectives to a hypergeometric distribution (phyper function in R). All graphics were made with ggplot2 package. For promoter analysis, scaffolds and predicted CDS of *C. roseus* genome sequencing project<sup>70</sup> were retrieved from Dryad (<http://datadryad.org/resource/doi:10.5061/dryad.hs593>). CDS were mapped on scaffolds using megablast (BLAST + suite 2.2.29<sup>72</sup>). When possible the 500 pb upstream the start codon were obtained for each CDS. The MEME suite (v4.11.2) was used to detect new ungapped motifs with DREME program using default parameter<sup>73</sup>. GOMO was next used to analyze the representation of candidate motifs in *Arabidopsis* genome.

**qPCR analysis.** Targeted gene expression measurement was performed by qPCR using primers described in Supplemental Table S5, after cDNA synthesis as described previously<sup>17</sup>.

**MIA quantification.** MIAs were extracted from lyophilized samples (*C. roseus* leaves and *M. sexta* intestinal tracts) by grinding tissues with a mixer mill (Restch, MM 400) during 3 min at the maximal frequency. The resulting powders were incubated in 1 ml methanol (containing 0.1% formic acid) and under vigorous shaking during 1 hour at 24 °C. After centrifugation (15,000 g; 15 minutes), supernatants were collected and used for quantification.

The MIA content of *C. roseus* leaves and of *M. sexta* intestinal tracts were determined using an UPLC-MS chromatography system coupled to a SQD mass spectrometer equipped with an electrospray ionization (ESI) source controlled by Masslynx 4.1 software (Waters, Milford, MA). Analyte separation was performed on a Waters Acquity HSS T3 C18 column (150 mm × 2.1 mm, i.d. 1.8 μm) with a flow rate of 0.4 mL/min at 55 °C and the volume of injection was 5 μL. The following linear elution gradient was used: acetonitrile-water-formic acid from 10:90:0.1 to 50:50:0.1 over 5 min. The capillary and sample cone voltages were 3,000 V and 30 V, respectively. The cone and desolvation gas flow rates were 60 and 800 Lh – 1. MS experiments were carried out in positive mode in the selected ion-monitoring mode using m/z 337 for catharanthine ([M + H]<sup>+</sup>, RT = 12.33 min), m/z 457 for vindoline ([M + H]<sup>+</sup>, RT = 14.69 min), m/z 793 for anhydrovinblastine ([M + H]<sup>+</sup>, RT = 16.5 min), m/z 427 for vindorosine ([M + H]<sup>+</sup>, RT = 15.03 min), m/z 353 for ajmalicine ([M + H]<sup>+</sup>, RT = 11.7 min), m/z 531 for strictosidine ([M + H]<sup>+</sup>, RT = 10.39 min), m/z 349 for serpentine ([M + H]<sup>+</sup>, RT = 13.01 min). The acquired data was processed by the QuanLynx™ software (Waters, UK). Relative quantification was performed by correcting peak areas by sample masses.

**Quantification of oxophytodienoic acid, jasmonic acid, and jasmonoyl-isoleucine.** Samples stored at –80 °C were freeze dried and powdered for subsequent analysis. Thirty milligrams of freeze dried powder were extracted at 4 °C with 1 ml of H<sub>2</sub>O:MeOH (90:10) containing 100 ng/ml of internal standards. After 20 min of incubation, samples were centrifuged at full speed for 15 min at 4 °C. The supernatant was recovered and adjusted to pH 2.8 with 6% acetic acid, and subsequently partitioned twice against diethylether. The fractions were pooled and dried in a speed vacuum and resuspended in H<sub>2</sub>O:MeOH (90:10). A 20 μL aliquot was injected into an Acquity ultra-performance liquid chromatography system (UPLC) (Waters, Mildford, MA, USA) interfaced to a triple quadrupole mass spectrometer (TQD, Waters, Manchester, UK). The LC separation was performed by HPLC Kinetex C18 analytical column 5 μm particle size, 2.1 × 100 mm (Phenomenex). The chromatographic conditions and mass spectrometry were performed as described previously<sup>74</sup>.

**Statistical procedures.** Statistical differences between means were non-parametrically tested in R<sup>75</sup>. The Wilcoxon rank sum test was used to avoid any conflict with the distribution of data. This was particularly the case for peak area in the metabolic analysis which displayed heteroscedasticity. P-values of tests were corrected with False Discovery Rate. Specific statistical procedures for RNA-seq data are presented above. Graphics were made with the ‘ggplot2’ package<sup>76</sup>.

## References

1. Facchini, P. J. & De Luca, V. Opium poppy and Madagascar periwinkle: model non-model systems to investigate alkaloid biosynthesis in plants. *Plant J.* **54**, 763–784 (2008).
2. St-Pierre, B. et al. Deciphering the evolution, cell biology and regulation of monoterpene indole alkaloids. *Adv. Bot. Res.* **68**, 73–109 (2013).
3. Guirimand, G. et al. Strictosidine activation in Apocynaceae: towards a ‘nuclear time bomb?’ *BMC Plant Biol.* **10**, 182 (2010).
4. Chockalingam, S., Sundari, M. S. N. & Thenmozhi, S. Impact of the extract of *Catharanthus roseus* on feeding and enzymatic digestive activities of *Spodoptera litura*. *J. Environ. Biol.* **10**, 303–307 (1989).
5. Luijendijk, T. J. C., van der Meijden, E. & Verpoorte, R. Involvement of strictosidine as a defensive chemical in *Catharanthus roseus*. *J. Chem. Ecol.* **22**, 1355–1366 (1996).
6. Roepke, J. et al. Vinca drug components accumulate exclusively in leaf exudates of Madagascar periwinkle. *Proc. Natl. Acad. Sci. USA* **107**, 15287–15292 (2010).
7. van der Heijden, R., Jacobs, D. I., Snoeijer, W., Hallard, D. & Verpoorte, R. The Catharanthus alkaloids: pharmacognosy and biotechnology. *Curr. Med. Chem.* **11**, 607–628 (2004).
8. Ishikawa, H., Colby, D. A. & Boger, D. L. Direct coupling of catharanthine and vindoline to provide vinblastine: total synthesis of (+)- and ent-(–)-vinblastine. *J. Am. Chem. Soc.* **130**, 420–421 (2008).
9. Brown, S., Clastre, M., Courdavault, V. & O'Connor, S. E. De novo production of the plant-derived alkaloid strictosidine in yeast. *Proc. Natl. Acad. Sci. USA* **112**, 3205–3210 (2015).
10. Qu, Y. et al. Completion of the seven-step pathway from tabersonine to the anticancer drug precursor vindoline and its assembly in yeast. *Proc. Natl. Acad. Sci. USA* **112**, 6224–6229 (2015).
11. Miettinen, K. et al. The seco-iridoid pathway from *Catharanthus roseus*. *Nat. Commun.* **5**, 3606 (2014).
12. Thamm, A. M. K., Qu, Y. & De Luca, V. Discovery and metabolic engineering of iridoid/secoiridoid and monoterpenoid indole alkaloid biosynthesis. *Phytochem. Rev.* **15**, 339–361 (2016).
13. Costa, M. M. R. et al. Molecular cloning and characterization of a vacuolar class III peroxidase involved in the metabolism of anticancer alkaloids in *Catharanthus roseus*. *Plant Physiol.* **146**, 403–417 (2008).
14. Besseau, S. et al. A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in *Catharanthus roseus*. *Plant Physiol.* **163**, 1792–803 (2013).
15. Westekemper, P. et al. Radioimmunoassay for the determination of the indole alkaloid vindoline in *Catharanthus*. *Planta Med.* **39**, 24–37 (1980).
16. Courdavault, V. et al. A look inside an alkaloid multisite plant: the *Catharanthus* logistics. *Curr. Opin. Plant Biol.* **19C**, 43–50 (2014).
17. Duge de Bernonville, T. et al. Characterization of a second secologanin synthase isoform producing both secologanin and secoxyloganin allows enhanced de novo assembly of a *Catharanthus roseus* transcriptome. *BMC Genomics* **16**, 619 (2015).
18. Carqueijeiro, I. et al. Virus-induced gene silencing in *Catharanthus roseus* by biolistic inoculation of tobacco rattle virus vectors. *Plant Biol.* **17**, 1242–1246 (2015).

19. Dugé de Bernonville, T. *et al.* Phytochemical genomics of the Madagascar periwinkle: Unravelling the last twists of the alkaloid engine. *Phytochemistry* **113**, 9–23 (2015).
20. Góngora-Castillo, E. *et al.* Development of transcriptomic resources for interrogating the biosynthesis of monoterpenoid indole alkaloids in medicinal plant species. *PLoS One* **7**, e52506 (2012).
21. Liscombe, D. K. & O'Connor, S. E. A virus-induced gene silencing approach to understanding alkaloid metabolism in *Catharanthus roseus*. *Phytochemistry* **72**, 1969–1977 (2011).
22. Salim, V., Yu, F., Altarejos, J. & De Luca, V. Virus-induced gene silencing identifies *Catharanthus roseus* 7-deoxyloganic acid-7-hydroxylase, a step in iridoid and monoterpenoid indole alkaloid biosynthesis. *Plant J.* **76**, 754–65 (2013).
23. Van Moerkerke, A. *et al.* CathaCyc, a metabolic pathway database built from *Catharanthus roseus* RNA-Seq data. *Plant Cell Physiol.* **54**, 673–85 (2013).
24. Xiao, M. *et al.* Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. *J. Biotechnol.* **166**, 122–134 (2013).
25. Aerts, R. J. & De Luca, V. Phytochrome is involved in the light-regulation of vindoline biosynthesis in *Catharanthus*. *Plant Physiol.* **100**, 1029–1032 (1992).
26. Binder, B. Y. K., Peebles, C. A. M., Shanks, J. V. & San, K.-Y. The effects of UV-B stress on the production of terpenoid indole alkaloids in *Catharanthus roseus* hairy roots. *Biotechnol. Prog.* **25**, 861–865 (2009).
27. De Luca, V., Fernandez, J. A., Campbell, D. & Kurz, W. G. W. Developmental regulation of enzymes of indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Physiol.* **86**, 447–450 (1988).
28. Ouwerkerk, P. B. F., Hallard, D., Verpoorte, R. & Memelink, J. Identification of UV-B light-responsive regions in the promoter of the tryptophan decarboxylase gene from *Catharanthus roseus*. *Plant Mol. Biol.* **41**, 491–503 (1999).
29. Ramani, S. & Chelliah, J. UV-B-induced signaling events leading to enhanced production of catharanthine in *Catharanthus roseus* cell suspension cultures. *BMC Plant Biol.* **7**, 1 (2007).
30. Frischknecht, P. M., Bättig, M. & Baumann, T. W. Effect of drought and wounding stress on indole alkaloid formation in *Catharanthus roseus*. *Phytochemistry* **26**, 707–710 (1987).
31. Vázquez-Flota, F., Carrillo-Pech, M., Minero-García, Y. & de Lourdes Miranda-Ham, M. Alkaloid metabolism in wounded *Catharanthus roseus* seedlings. *Plant Physiol. Biochem.* **42**, 623–628 (2004).
32. Aerts, R. J., Gisi, D., Carolis, E., Luca, V. & Baumann, T. W. Methyl jasmonate vapor increases the developmentally controlled synthesis of alkaloids in *Catharanthus* and *Cinchona* seedlings. *Plant J.* **5**, 635–643 (1994).
33. Courdavault, V. *et al.* CaaX-prenyltransferases are essential for expression of genes involved in the early stages of monoterpenoid biosynthetic pathway in *Catharanthus roseus* cells. *Plant Mol. Biol.* **57**, 855–870 (2005).
34. Papon, N. *et al.* Cytokinin and ethylene control indole alkaloid production at the level of the MEP/terpenoid pathway in *Catharanthus roseus* suspension cells. *Planta Med.* **71**, 572–574 (2005).
35. Kahl, J. *et al.* Herbivore-induced ethylene suppresses a direct defense but not a putative indirect defense against an adapted herbivore. *Planta* **210**, 336–342 (2000).
36. Onkokesung, N., Baldwin, I. T. & Gális, I. The role of jasmonic acid and ethylene crosstalk in direct defense of *Nicotiana attenuata* plants against chewing herbivores. *Plant Signal. Behav.* **5**, 1305–1307 (2010).
37. Wasternack, C. & Hause, B. Jasmonates: biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in Annals of Botany. *Ann. Bot.* **111**, 1021–1058 (2013).
38. Gantet, P. & Memelink, J. Transcription factors: tools to engineer the production of pharmacologically active plant metabolites. *Trends Pharmacol. Sci.* **23**, 563–569 (2002).
39. van der Fits, L. & Memelink, J. ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science* **289**, 295–297 (2000).
40. Vom Endt, D., Soares e Silva, M., Kijne, J. W., Pasquali, G. & Memelink, J. Identification of a bipartite jasmonate-responsive promoter element in the *Catharanthus roseus* ORCA3 transcription factor gene that interacts specifically with AT-Hook DNA-binding proteins. *Plant Physiol.* **144**, 1680–1689 (2007).
41. Vazquez-Flota, F. A. & De Luca, V. Developmental and light regulation of desacetoxyvindoline 4-hydroxylase in *Catharanthus roseus* (L.) G. Don. Evidence of a multilevel regulatory mechanism. *Plant Physiol.* **117**, 1351–1361 (1998).
42. Menke, F. L. H., Parchmann, S., Mueller, M. J., Kijne, J. W. & Memelink, J. Involvement of the octadecanoid pathway and protein phosphorylation in fungal elicitor-induced expression of terpenoid indole alkaloid biosynthetic genes in *Catharanthus roseus*. *Plant Physiol.* **119**, 1289–1296 (1999).
43. Eilert, U., De Luca, V., Constabel, F. & Kurz, W. G. W. Elicitor-mediated induction of tryptophan decarboxylase and strictosidine synthase activities in cell suspension cultures of *Catharanthus roseus*. *Arch. Biochem. Biophys.* **254**, 491–497 (1987).
44. Namdeo, A., Patil, S. & Fulzele, D. P. Influence of fungal elicitors on production of ajmalicine by cell cultures of *Catharanthus roseus*. *Biotechnol. Prog.* **18**, 159–162 (2002).
45. Pasquali, G. *et al.* Coordinated regulation of two indole alkaloid biosynthetic genes from *Catharanthus roseus* by auxin and elicitors. *Plant Mol. Biol.* **18**, 1121–1131 (1992).
46. Acevedo, F. E., Rivera-Vega, L. J., Chung, S. H., Ray, S. & Felton, G. W. Cues from chewing insects—the intersection of DAMPs, HAMPs, MAMPs and effectors. *Curr. Opin. Plant Biol.* **26**, 80–86 (2015).
47. Wu, J. & Baldwin, I. T. Herbivory-induced signalling in plants: perception and action. *Plant. Cell Environ.* **32**, 1161–1174 (2009).
48. Favali, M. A., Musetti, R., Benvenuti, S., Bianchi, A. & Pressacco, L. *Catharanthus roseus* L. plants and explants infected with phytoplasmas: alkaloid production and structural observations. *Protoplasma* **223**, 45–51 (2004).
49. Jagoueix-Eveillard, S. *et al.* *Catharanthus roseus* genes regulated differentially by mollicute infections. *Mol. Plant-Microbe Interact.* **14**, 225–233 (2001).
50. Liu, L.-Y. D. *et al.* High-throughput transcriptome analysis of the leafy flower transition of *Catharanthus roseus* induced by peanut witches'-broom phytoplasma infection. *Plant Cell Physiol.* **55**, 942–57 (2014).
51. Srivastava, S., Pandey, R., Kumar, S. & Nautiyal, C. S. Correspondence between flowers and leaves in terpenoid indole alkaloid metabolism of the phytoplasma-infected *Catharanthus roseus* plants. *Protoplasma* **251**, 1307–1320 (2014).
52. Fragoso, V., Rothe, E., Baldwin, I. T. & Kim, S.-G. Root jasmonic acid synthesis and perception regulate folivore-induced shoot metabolites and increase *Nicotiana attenuata* resistance. *New Phytol.* **202**, 1335–1345 (2014).
53. Courdavault, V., Burlat, V., St-Pierre, B. & Giglioli-Guivarc'h, N. Proteins prenylated by type I protein geranylgeranyltransferase act positively on the jasmonate signalling pathway triggering the biosynthesis of monoterpenoid indole alkaloids in *Catharanthus roseus*. *Plant Cell Rep.* **28**, 83–93 (2009).
54. Oudin, A., Courtois, M., Rideau, M. & Clastre, M. The iridoid pathway in *Catharanthus roseus* alkaloid biosynthesis. *Phytochem. Rev.* **6**, 259–276 (2007).
55. Patro, R., Duggal, G. & Kingsford, C. Accurate, fast, and model-aware transcript expression quantification with Salmon. *bioRxiv* **21592** (2015).
56. Barron-Gafford, G. A. *et al.* Herbivory of wild *Manduca sexta* causes fast down-regulation of photosynthetic efficiency in *Datura wrightii*: an early signaling cascade visualized by chlorophyll fluorescence. *Photosynth. Res.* **113**, 249–260 (2012).
57. Nabity, P. D., Zavala, J. A. & DeLucia, E. H. Herbivore induction of jasmonic acid and chemical defences reduce photosynthesis in *Nicotiana attenuata*. *J. Exp. Bot.* **64**, 685–694 (2013).

58. Attaran, E. *et al.* Temporal dynamics of growth and photosynthesis suppression in response to jasmonate signaling. *Plant Physiol.* **165**, 1302–1314 (2014).
59. Lurin, C. *et al.* Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16**, 2089–2103 (2004).
60. Koiwa, H., Bressan, R. A. & Hasegawa, P. M. Regulation of protease inhibitors and plant defense. *Trends Plant Sci.* **2**, 379–384 (1997).
61. Van Poecke, R. M. P., Posthumus, M. A. & Dicke, M. Herbivore-induced volatile production by *Arabidopsis thaliana* leads to attraction of the parasitoid *Cotesia rubecula*: chemical, behavioral, and gene-expression analysis. *J. Chem. Ecol.* **27**, 1911–1928 (2001).
62. Varanda, E. M., Zúñiga, G. E., Salatino, A., Roque, N. F. & Corcuer, L. J. Effect of ursolic acid from epicuticular waxes of *Jacaranda decurrens* on *Schizaphis graminum*. *J. Nat. Prod.* **55**, 800–803 (1992).
63. Usia, T., Watabe, T., Kadota, S. & Tezuka, Y. Cytochrome P450 2D6 (CYP2D6) inhibitory constituents of *Catharanthus roseus*. *Biol. Pharm. Bull.* **28**, 1021–1024 (2005).
64. Vanholme, B., Grunewald, W., Bateman, A., Kohchi, T. & Gheysen, G. The tify family previously known as ZIM. *Trends Plant Sci.* **12**, 239–244 (2007).
65. Onkokesung, N. *et al.* Jasmonic acid and ethylene modulate local responses to wounding and simulated herbivory in *Nicotiana attenuata* leaves. *Plant Physiol.* **153**, 785–798 (2010).
66. Wang, K. L.-C., Li, H. & Ecker, J. R. Ethylene biosynthesis and signaling networks. *Plant Cell* **14**, S131–S151 (2002).
67. Yoo, S.-D., Cho, Y.-H., Tena, G., Xiong, Y. & Sheen, J. Dual control of nuclear EIN3 by bifurcate MAPK cascades in C2H4 signalling. *Nature* **451**, 789–795 (2008).
68. Fujimoto, S. Y., Ohta, M., Usui, A., Shinshi, H. & Ohme-Takagi, M. Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. *Plant Cell* **12**, 393–404 (2000).
69. Van Der Fits, L. & Memelink, J. The jasmonate-inducible AP2/ERF-domain transcription factor ORCA3 activates gene expression via interaction with a jasmonate-responsive promoter element. *Plant J.* **25**, 43–53 (2001).
70. Kellner, F. *et al.* Genome-guided investigation of plant natural product biosynthesis. *Plant J.* **82**, 680–692 (2015).
71. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
72. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
73. Bailey, T. L. *et al.* Meme Suite: tools for motif discovery and searching. *Nucleic Acids Res.* W202–8 (2009).
74. Gamir, J., Pastor, V., Cerezo, M. & Flors, V. Identification of indole-3-carboxylic acid as mediator of priming against *Cucumberina*. *Plant Physiol. Biochem.* **61**, 169–179 (2012).
75. Team, R. R. Development Core Team. R: A Language and Environment for Statistical Computing. R foundation for statistical computing. <https://www.R-project.org/> (2013).
76. Wickham, H. ggplot2: elegant graphics for data analysis (ed. Springer Verlag, Springer Science & Business Media, 2009).

## Acknowledgements

We gratefully acknowledge the financial support from the “Région Centre” (France, ABISAL grant) and from the University of Tours. We also thank Jean-Louis Rouet and Laurent Catherine for access and support to the CCSC computing resources (Cascimodot Federation, CNRS, Orléans).

## Author Contributions

T.D.D.B., M.C., B.St.-P., S.M., S.E.O.C. and V.C. designed the project; T.D.D.B., I.C., A.L., F.La., K.M., A.O., G.G., O.P., S.B. and E.H. performed molecular and metabolic analyses; T.D.D.B. and F.Li. performed transcript annotation and term enrichment analysis; P.S. and V.F. performed J.A. quantification; T.D.D.B. and V.C. supervised the work and wrote the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Dugé de Bernonville, T. *et al.* Folivory elicits a strong defense reaction in *Catharanthus roseus*: metabolomic and transcriptomic analyses reveal distinct local and systemic responses. *Sci. Rep.* **7**, 40453; doi: 10.1038/srep40453 (2017).

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



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017





# CHASE-Containing Histidine Kinase Receptors in Apple Tree: From a Common Receptor Structure to Divergent Cytokinin Binding Properties and Specific Functions

Dimitri Daudu<sup>1</sup>, Elsa Allion<sup>1</sup>, Franziska Liesecke<sup>1</sup>, Nicolas Papon<sup>2</sup>, Vincent Courdavault<sup>1</sup>, Thomas Dugé de Bernonville<sup>1</sup>, Céline Mélin<sup>1</sup>, Audrey Oudin<sup>1</sup>, Marc Clastre<sup>1</sup>, Arnaud Lanoue<sup>1</sup>, Martine Courtois<sup>1</sup>, Olivier Pichon<sup>1</sup>, David Giron<sup>3</sup>, Sabine Carpin<sup>4</sup>, Nathalie Giglioli-Guivarc'h<sup>1</sup>, Joël Crèche<sup>1</sup>, Sébastien Besseau<sup>1</sup> and Gaëlle Glévarec<sup>1\*</sup>

## OPEN ACCESS

### Edited by:

Ján A. Miernyk,  
Agricultural Research Service (USDA),  
United States

### Reviewed by:

Jan Hejatkó,  
Masaryk University, Czechia  
Dong Xu,  
University of Missouri, United States

### \*Correspondence:

Gaëlle Glévarec  
gaelle.glevarec@univ-tours.fr

### Specialty section:

This article was submitted to  
Plant Physiology,  
a section of the journal  
*Frontiers in Plant Science*

Received: 16 June 2017

Accepted: 04 September 2017

Published: 20 September 2017

### Citation:

Daudu D, Allion E, Liesecke F, Papon N, Courdavault V, Dugé de Bernonville T, Mélin C, Oudin A, Clastre M, Lanoue A, Courtois M, Pichon O, Giron D, Carpin S, Giglioli-Guivarc'h N, Crèche J, Besseau S and Glévarec G (2017) CHASE-Containing Histidine Kinase Receptors in Apple Tree: From a Common Receptor Structure to Divergent Cytokinin Binding Properties and Specific Functions.

Front. Plant Sci. 8:1614.

doi: 10.3389/fpls.2017.01614

<sup>1</sup> EA 2106 Biomolécules et Biotechnologies Végétales, Université François-Rabelais, Tours, France, <sup>2</sup> EA 3142 Groupe d'Etude des Interactions Hôte-Pathogène, Université Angers, Angers, France, <sup>3</sup> UMR 7261 Institut de Recherche sur la Biologie de l'Insecte, Centre National de la Recherche Scientifique (CNRS), Université François-Rabelais, Tours, France, <sup>4</sup> EA 1207 Laboratoire de Biologie des Ligneux et des Grandes Cultures, Université d'Orléans, Orléans, France

Cytokinin signaling is a key regulatory pathway of many aspects in plant development and environmental stresses. Herein, we initiated the identification and functional characterization of the five CHASE-containing histidine kinases (CHK) in the economically important *Malus domestica* species. These cytokinin receptors named MdCHK2, MdCHK3a/MdCHK3b, and MdCHK4a/MdCHK4b by homology with *Arabidopsis* AHK clearly displayed three distinct profiles. The three groups exhibited architectural variations, especially in the N-terminal part including the cytokinin sensing domain. Using a yeast complementation assay, we showed that MdCHK2 perceives a broad spectrum of cytokinins with a substantial sensitivity whereas both MdCHK4 homologs exhibit a narrow spectrum. Both MdCHK3 homologs perceived some cytokinins but surprisingly they exhibited a basal constitutive activity. Interaction studies revealed that MdCHK2, MdCHK4a, and MdCHK4b homodimerized whereas MdCHK3a and MdCHK3b did not. Finally, qPCR analysis and bioinformatics approach pointed out contrasted expression patterns among the three MdCHK groups as well as distinct sets of co-expressed genes. Our study characterized for the first time the five cytokinin receptors in apple tree and provided a framework for their further functional studies.

**Keywords:** CHASE-containing histidine kinase, cytokinin, yeast complementation assay, RNAseq data, protein-protein interaction, *Malus domestica*

## INTRODUCTION

Cytokinins are essential adenine-derived plant hormones, gathering more than 40 structures substituted at the N<sup>6</sup>-position by an isoprenoid or aromatic chain (Spíchal, 2012; Osugi and Sakakibara, 2015). They are involved in numerous physiological processes such as cell division, delayed senescence, vascular tissue development, root architecture and light responses (Sakakibara, 2006; Kieber and Schaller, 2014; Zürcher and Müller, 2016). Cytokinins also play roles in the

interaction with both biotic and abiotic factors (Frugier et al., 2008; Giron and Glévarec, 2014; Naseem et al., 2014; Zwack and Rashotte, 2015).

Plant cytokinin perception is mediated by CHASE domain-containing histidine kinase receptors (CHK) as first actors of cytokinin signaling (Inoue et al., 2001). These receptors display a complex multidomain structure with a N-terminal part including at least two hydrophobic membrane-spanning domains (TM) that border an extracytosolic sensing domain referred to as CHASE (Cyclase/Histidine kinase Associated Sensory Extracellular) (Anantharaman and Aravind, 2001; Mougel and Zhulin, 2001) as well as a cytoplasmic C-terminal part containing a catalytic histidine kinase domain (HK) and both receiver and pseudo-receiver domains (REC and REC-like, respectively) (Ueguchi et al., 2001). The HK domain is composed of an HK dimerization and phosphoacceptor domain (HisKA) and an HK catalytic domain called the HK-like ATPase domain (HATPase). The cytokinin perception by the CHASE domain leads to the autophosphorylation of a conserved histidine within the HK domain. The phosphate residue is then transferred to the REC domain on a conserved aspartate residue (Inoue et al., 2001). Although the pseudo-receiver domain of CHK is structurally similar to the REC domain, its functionality has not been yet elucidated (Ueguchi et al., 2001; Lomin et al., 2012). Subsequently, the signal is transferred by phosphorelay to Response Regulators (RR) through histidine-containing phosphotransfer shuttle proteins (HPt). While type-B RRs (RRB) are transcription factors that play a positive role in mediating cytokinin-regulated gene expression, type-A RRs (RRA) act as negative regulators of cytokinin responses (To et al., 2004; Mason et al., 2005; Ginis et al., 2012). In addition, Cytokinin Response Factors (CRF) interact directly with HPts and were reported to influence a subset of cytokinin responses (Cutcliffe et al., 2011; Raines et al., 2016).

Cytokinin receptors were shown to localize mainly to the endoplasmic reticulum (ER) both in *Arabidopsis thaliana* and *Zea mays* (Caesar et al., 2011; Lomin et al., 2011; Wulfetange et al., 2011). They are supposed to interact with each other, forming potential homo- and hetero-dimers probably enabling the *trans*-phosphorylation of the HK domain following cytokinin perception (Dortay et al., 2006; Hothorn et al., 2011). However, the signal transmission process across the membrane remains unknown. The ligand-binding properties of the cytokinin receptors have been investigated mostly using heterologous assay systems through their expression in *Escherichia coli* or *Saccharomyces cerevisiae* cells (Inoue et al., 2001; Romanov et al., 2006; Stolz et al., 2011; von Schwartzenberg et al., 2016). More recently, a plant assay system has been developed to overcome the problem of alien membrane environment and the difficulty to express some membrane receptors in bacteria or yeast (Lomin et al., 2015). Overall, the cytokinin receptors differ in their preference toward cytokinin forms (Yonekura-Sakakibara et al., 2004; Lomin et al., 2011; Kuderová et al., 2015) but their functional and specific properties as well as the structural changes caused by cytokinin binding remain to be elucidated.

While it is established that the CHK receptors operate mostly in a redundant fashion, the extensive studies of *Arabidopsis*

mutants have attributed some specific roles to single receptors. Among others, AHK4 is the main regulator of primary root growth and vascular morphogenesis whereas AHK2 and AHK3 are commonly involved in chlorophyll retention during leaf senescence (Kim et al., 2006; Riefler et al., 2006). CHKs are also involved in response to environmental changes (Zwack and Rashotte, 2015). The three AHKs are also known to function as negative regulators in osmotic stress responses (Tran et al., 2007; Kumar and Verslues, 2015). AHK2 and AHK3 play an additional negative regulatory role in cold stress (Jeon et al., 2016) and ensure a protective function during light stress (Cortleven et al., 2014). Moreover, cytokinin receptors also take part in a large range of responses to biotic interactions. In legume plants, cytokinin receptors regulate nodule formation (Tirichine et al., 2007; Held et al., 2014; Boivin et al., 2016). In *Arabidopsis*, the success of the pathogens *Rhodococcus fascians* and *Hyaloperonospora arabidopsidis* depends on some AHKs (Petry et al., 2009; Argueso et al., 2012). Finally, NaCHK2 and NaCHK3 modulate herbivory-induced defense signaling and defenses in *Nicotiana attenuata* (Schäfer et al., 2015). If the knowledge on the cytokinin receptors is increasingly important, their study in various plant models is necessary for a complete understanding of their biological functions.

Previous works on *M. domestica* reported a large accumulation of cytokinins in the leaves infected by the insect *Phyllonorycter blancardella*. This increase is responsible for the preservation of nutrient green tissues when leaves are otherwise turning yellow (Giron et al., 2007; Kaiser et al., 2010; Zhang et al., 2016). Based on the involvement of cytokinins in this plant-biotic interaction, we initiated the study of cytokinin signaling in apple tree with a special focus on cytokinin receptors. Indeed, considering that apple tree is one of the most cultivated fruit-tree with a continual worldwide production increase, a greater knowledge of cytokinin signaling pathway in this species could provide new opportunities for agronomical and economical purposes. This study discloses an overall and complete characterization of the five *M. domestica* CHASE Histidine Kinases (MdCHKs).

## MATERIALS AND METHODS

### *In Silico* Sequence Analysis and Receptor Identification

To identify MdCHK receptors, BLAST searches were performed against the Genome Database for Rosaceae (GDR; Jung et al., 2014) using *A. thaliana* cytokinin receptor sequences as queries (AHK2, AHK3, and AHK4). Five sequences were identified based on genome analysis and corresponding cDNA were amplified from various plant organs using specific primers (Supplementary Table S1). Sequences were registered in Genbank as *MdCHK2* (KM114879), *MdCHK3a* (KM114880), *MdCHK3b* (KM114881), *MdCHK4a* (KM114883) and *MdCHK4b* (KM114882).

*MdCHKs* gene organization has been visualized using the FancyGene program (Rambaldi and Ciccarelli, 2009). Phylogeny analyses were performed on conserved domains and local similarities among proteins sequences. To this aim, multiple protein sequence alignments were done using the COBALT tool

(Papadopoulos and Agarwala, 2007) and sequences were curated with Gblocks prior the construction of a bootstrap neighbor joining tree. Protein domain predictions were acquired using the SMART (Letunic et al., 2015) and PROSITE (Sigrist et al., 2002) programs, and transmembrane regions were identified with TMHMM (Krogh et al., 2001) and TMpred tools (Hofmann and Stoffel, 1993). Visualization of the transmembrane helices has been performed with a helical wheel drawing program<sup>1</sup>.

## Yeast Complementation Assay

The *S. cerevisiae* strain YIL147C, deficient in SLN1 receptor (*MATa/α, ura3, leu2, his3, can1Δ::LEU2-MFA1pro-HIS3/CAN1, sln1Δ::KanMX/SLN1*) was used in complementation assays. Full-length coding sequences of MdCHKs were amplified and cloned into the yeast expression vector pYES2 under the control of the *GAL1* gene promoter using the *NotI* restriction site (for primers, see Supplementary Table S1). The *S. Cerevisiae* strain was transformed as follows. Cells were grown in 200 mL YPD liquid medium (150 rpm, 28°C) to 0.4–0.6 OD<sub>600</sub>, harvested by centrifugation (3000 g, 10 min) and resuspended in 0.1 M lithium acetate, 10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 10 mM DTT. After 1 h incubation at 30°C, cells were washed twice with ice-cold 1 M sorbitol and resuspended in 1–5 mL ice-cold 1 M sorbitol. Plasmid DNA (0.5–1 μg DNA) was added to a 180 μL cell suspension and transferred to a 0.2 cm gap width electroporation cuvette. Electroporation was performed using a Bio-Rad gene pulser with an electric pulse of 2.5 kV, 25 μF and 200 Ω. Cells were immediately washed out from the cuvettes after the electroporation with YPD, plated on selective CSM-URA medium and incubated at 28°C for 3–5 days. Fresh colonies were then grown 3 h in liquid YPD at 30°C and meiosis was induced by pouring the suspension cell on ACK medium (10 g/L potassium acetate, 2.5 g/L yeast extract, 20 g/L agar). These plates were incubated 1 week at 20°C and haploids were finally selected on MMAS medium plates [20 g/L galactose, 7 g/L YNB (WA), 0.6 g/L DOB-LEU-HIS-ARG-URA, 0.06 g/L L-canavanine, 0.2 g/L G418, 20 g/L agar] supplemented with 10 μM *trans*-zeatin at 28°C for 3–5 days. Suspensions of transformants were then spotted onto dropout media containing or not 10 μM of *trans*-zeatin with 2% galactose and grown for 48 h at 28°C. For specificity and sensitivity assays, complemented-yeast growth was carried out in liquid YGal (7 g/L YNB, 0.8 g/L CSM-URA, 20 g/L galactose) supplemented with various types and concentrations of cytokinins for 48 h at 28°C. Cell growth was measured at 630 nm (BioHit Reader BP800).

## Chemicals

Pure standards of isopentenyladenosine 5'-monophosphate, *trans*-zeatin riboside 5'-monophosphate, *cis*-zeatin riboside 5'-monophosphate, isopentenyladenine, *trans*-zeatin, *cis*-zeatin, dihydrozeatin, isopentenyladenosine, *trans*-zeatin riboside, *cis*-zeatin riboside, dihydrozeatin riboside, dihydrozeatin

riboside 5'-monophosphate, *trans*-zeatin O-glucoside, *trans*-zeatin O-glucoside riboside, *trans*-zeatin N7-glucoside, 2-methylthio-isopentyladenine, 2-methylthio-*trans*-zeatin, 2-methylthio-*cis*-zeatin, 2-methylthio-isopentyladenosine, 2-methylthio-*trans*-zeatin riboside, 2-methylthio-*cis*-zeatin riboside were purchased from Olchemim (Olomouc, Czechia).

## RNA Isolation and Gene Expression Analysis

Extraction of total RNA from *M. domestica* organs was performed using the NucleoSpin RNA extraction kit (Marcheray-Nagel), with improved lysis step (McKenzie et al., 1997). First-strand cDNA were synthesized from 1 μg of total RNA using the iScript cDNA Synthesis Kit (Bio-Rad). Quantitative real-time PCR measurements were carried out in triplicate using SsoAdvanced Universal SYBR Green (Bio-Rad) in a 15 μL final volume containing 6 μL diluted template cDNA and specific primers (0.5 μM) (Supplementary Table S1). Amplification was performed on a CFX96 Touch real-time PCR system (Bio-Rad) with the following conditions: 95°C for 7 min and 40 cycles at 95°C for 10 s and 60°C for 40 s. Amplification was followed by a melt curve analysis. Absolute quantification of transcript copy number was assessed with calibration curves. Transcript levels were then normalized with EF1α.

## Subcellular Localization Experiments

Subcellular localization of MdCHK receptors were studied in *Catharanthus roseus* C20D cells transiently transformed using plasmid-coated particles bombardment as described in Guirimand et al. (2009). The full length MdCHK sequences were amplified and cloned into the *SpeI* restriction site of pSCA-YFP plasmid (for primers, see Supplementary Table S1), in frame with the 5' extremity of the YFP coding sequence. The endoplasmic reticulum (ER) cyan fluorescent protein (CFP) marker (Guirimand et al., 2010) was used in co-transformation assays.

Dynamic localization of MdCHK receptors was also studied in yeast *S. cerevisiae* strain WT303 (*MATa/α, leu2, trp1, ura3, ade2, his3*) transformed by pESC-LEU plasmids (Foureau et al., 2016) containing *MdCHK* sequences, except for MdCHK2, cloned in pYES2 and transformed in *sln1Δ* yeast strain to bypass the sequence toxicity in microorganisms. The CYP450 T16H2 sequence was clone in the pESC-TRP plasmid in fusion with the 5'end of the CFP sequence and used as an ER marker (Besseau et al., 2013). *MdCHK* sequences were cloned under galactose inducible promoter and fused at the 5' end with the YFP sequence. Transformed colonies were cultivated on selective plates (CSM-LEU or CSM-URA, supplemented by 2% glucose, respectively, for pESC-LEU and pYES2) at 30°C for 48 h and then transferred in inducing liquid media (CSM-LEU or CSM-URA, supplemented by 2% galactose, respectively, for pESC-LEU and pYES2) with or without iP (5 μM) for additional overnight culture at 28°C.

An Olympus BX51 epifluorescence microscope equipped with the Olympus DP71 digital camera and Cell\*D imaging software (Soft Imaging System Olympus) was used for image capture and

<sup>1</sup>[http://www-nmr.cabm.rutgers.edu/bioinformatics/Proteomic\\_tools/Helical\\_wheel/](http://www-nmr.cabm.rutgers.edu/bioinformatics/Proteomic_tools/Helical_wheel/)

merging false-colored images of both C20D cells and *S. cerevisiae* colonies expressing YFP.

## BiFC Interaction Assays

Bimolecular Fluorescent Complementation (BiFC) experiments were conducted using SPYNE and SPYCE plasmids (Waadt and Kudla, 2008). MdCHK sequences were amplified and cloned into the *SpeI* restriction site (for primers, see Supplementary Table S1), in frame with the 5' extremity of a truncated YFP coding sequence. Transient transformation of *C. roseus* cells by particle bombardment and YFP imaging were performed according to Guirimand et al. (2009) with adaptation for BiFC assays (Guirimand et al., 2010). Interactions were tested in triplicates using three independent plasmid clones.

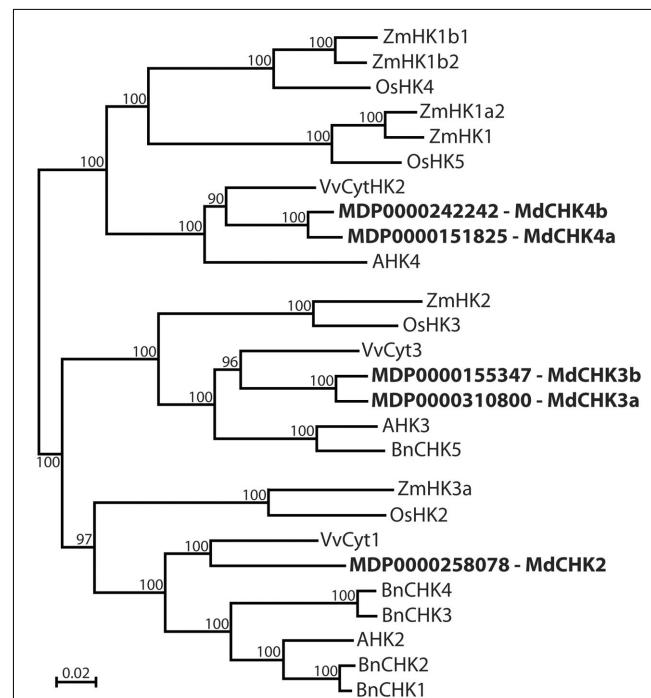
## RNA-seq Data Analysis

Available RNAseq data for *M. domestica* was downloaded from NCBI via SRA toolkit2.6.2. The recovered SRA files were transformed in fastq format with the “fastq-dump” command from SRA toolkit. The files were cleaned with Trimmomatic 0.36 with default parameters and using provided adapter sequences for TruSeq2 and TruSeq3. The transcription quantification was performed with Salmon 0.6.1 using the Variational Bayesian EM algorithm and biase correction. TPM (transcripts per million) values from the resulting quant.sf files were combined under R 3.3.0 in an expression matrix containing 95,232 predicted genes (*Malus domestica* v3.0)  $\times$  250 experimental conditions. Using the expression matrix, Pearson Correlation Coefficients (PCC) and further Highest Reciprocal Ranks (HRR) computation were performed using a homemade program written in C [HRR (gene A, gene B)] = max [rank (gene A, gene B), rank (gene B, gene A)] to establish the co-expressed genes lists for MdCHK and GO enrichment tests. For each MdCHK highly co-expressed genes, i.e., genes with a HRR  $\leq$  500 were selected. The procedure was repeated with publicly available *A. thaliana* RNAseq data. We similarly prepared an expression matrix containing 33,604 transcripts (*Arabidopsis* TAIR v10 genome annotation) and 1,676 samples. Co-expressed genes lists were obtained after calculating PCC and ranking them with HRR. For each AtCHK (AT5G35750.1, AT1G27320.1, AT2G01830.2), gene pairs having an HRR  $<$  500 were considered to be significantly co-expressed. Orthology between *Arabidopsis* and apple tree was obtained from Plaza 3.0 (Van Bel et al., 2012). The functions represented by coexpressed genes of each MdCHK were analyzed with the Gene Ontology classification. A BlastX was performed on the *M. domestica* genome v3.0 to recover correspondent protein sequences and Pfam domains were identified using Hmmer. The functional annotations of the *M. domestica* genome v3.0 were generated using Trinotate on the previous data. In order to determine potential functional enrichment for every target gene, enrichment of GO terms was tested by comparing effectives to a hypergeometric distribution ( $p$ -value cut-off = 0.001) using the R “phyper” function. To compare redundancies in the five co-expressed genes lists, a Venn diagram was drawn using the venneuler package 1.1 for R.

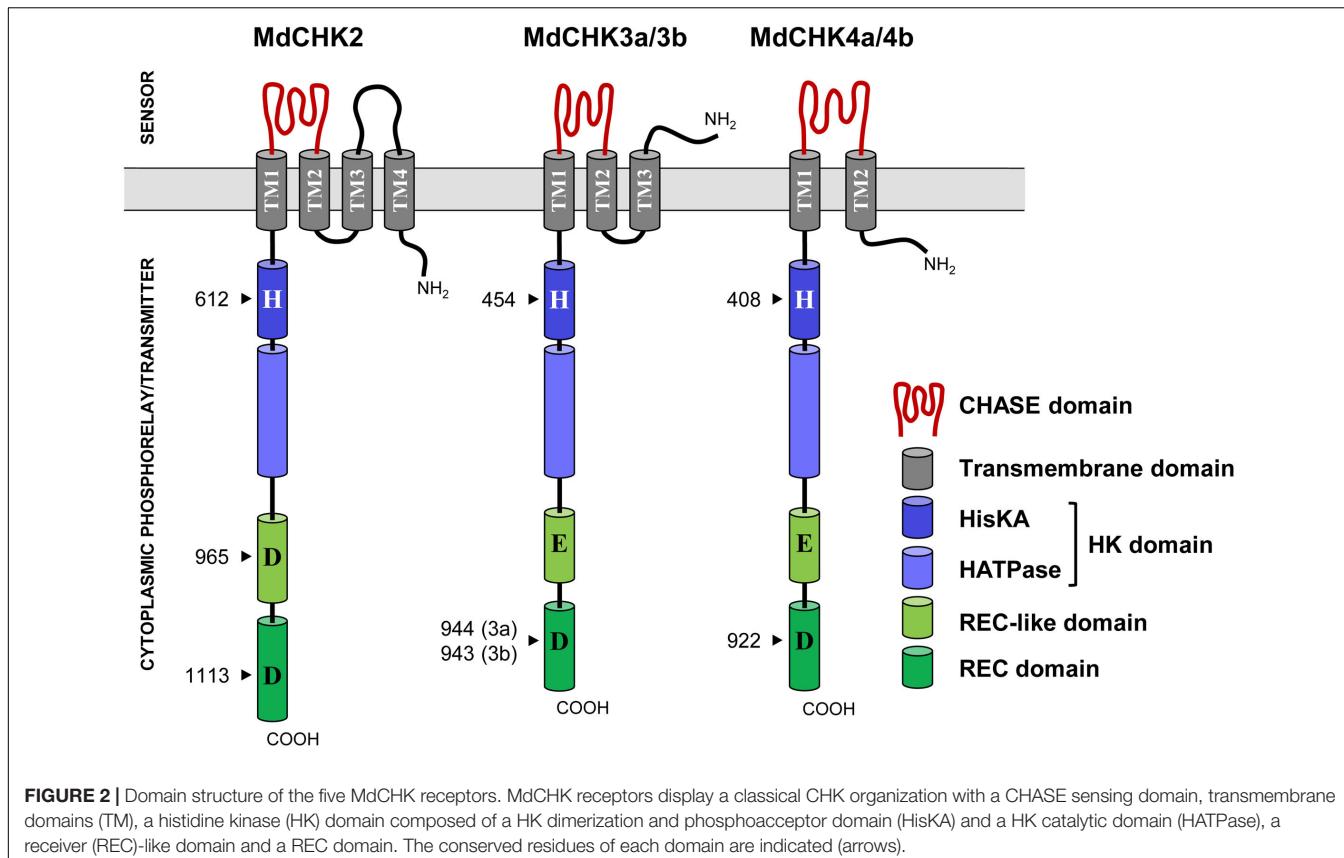
## RESULTS

### Identification of Five *Malus domestica* CHASE-Containing Histidine Kinases (MdCHKs)

Based on the CHASE domain of the three *A. thaliana* CHKs, apple tree genome was browsed to identify putative *CHK* sequences in *M. domestica*. Five candidates were identified and named *MdCHK2* (locus tag MDP0000258078), *MdCHK3a* (locus tag MDP0000310800), *MdCHK3b* (locus tag MDP0000155347), *MdCHK4a* (locus tag MDP0000151825) and *MdCHK4b* (locus tag MDP0000242242) according to their distribution within the three classical groups homologous to *AHK2*, *AHK3*, and *AHK4*, as shown by phylogenetic analysis (Figure 1). Genomic sequences revealed that *MdCHK2* possesses 13 exons and is located on chromosome 9 (Supplementary Figure S1). Both *MdCHK3a* and *MdCHK3b*, respectively, located on chromosomes 16 and 13, display 10 exons with a similar organization regarding



**FIGURE 1 |** Phylogenetic analysis of *Malus domestica* cytokinin receptors. Apple tree CHASE Histidine Kinase receptors *MdCHK2* (KM114879), *MdCHK3a* (KM114880), *MdCHK3b* (KM114881), *MdCHK4a* (KM114883), and *MdCHK4b* (KM114882) were compared to cytokinin receptors characterized in angiosperms. The tree was constructed by neighbor-joining distance analysis on conserved protein sequence domains. *Brassica napus*: BnCHK1 (KF621029), BnCHK2 (KF621030), BnCHK3 (KF621031), BnCHK4 (KF621032), BnCHK5 (KF621033); *Arabidopsis thaliana*: AHK2 (At5g35750), AHK3 (At1g27320), AHK4 (At2g01830); *Oryza sativa*: OsHK5 (Os02g50480), OsHK2 (Os10g21810), OsHK3 (Os01g69920), OsHK4 (Os03g50860); *Vitis vinifera*: VvCyt3 (CAO42401), VvCyt1 (GSVIVT01030058001), VvCyt2 (CAO66151); *Zea mays*: ZmHK1 (NP\_001104859), ZmHK2 (NP\_001104866), ZmHK3a (AB102957), ZmHK1a2 (NP\_001105857), ZmHK1b1 (NP\_001105858), ZmHK1b2 (NP\_001105913).



**FIGURE 2 |** Domain structure of the five MdCHK receptors. MdCHK receptors display a classical CHK organization with a CHASE sensing domain, transmembrane domains (TM), a histidine kinase (HK) domain composed of a HK dimerization and phosphoacceptor domain (HisKA) and a HK catalytic domain (HATPase), a receiver (REC)-like domain and a REC domain. The conserved residues of each domain are indicated (arrows).

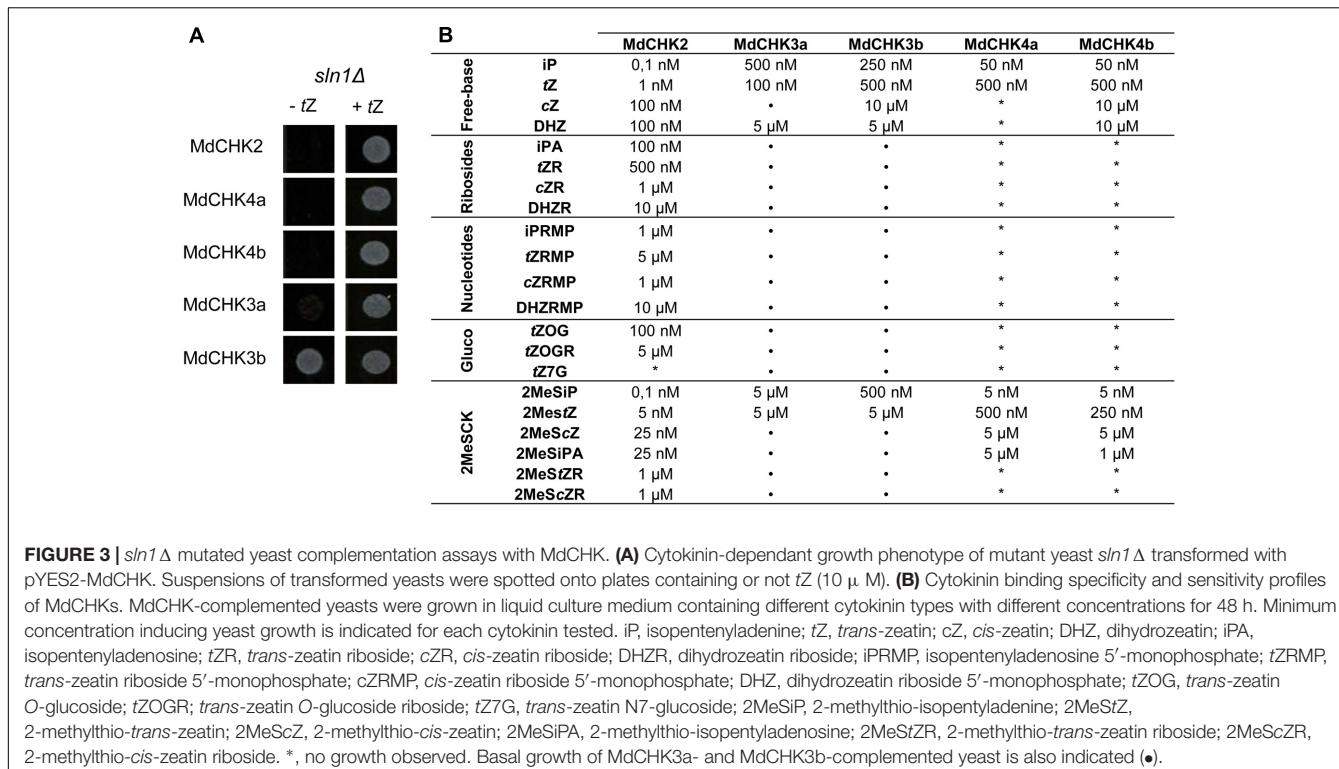
intron positions and intron/exon sizes. The same observation was made for *MdCHK4a* and *MdCHK4b* respectively located on chromosomes 13 and 10 and containing 11 exons. Such similarity may reflect the gene duplication events leading to the couples of CHKs homologous *MdCHK3a/MdCHK3b* and *MdCHK4a/MdCHK4b* which share respectively 94.67 and 95.67% nucleotide identity. The full-length cDNAs of the five MdCHKS were cloned and deposited at NCBI under the GenBank accession numbers KM114879 to KM114883. They contained large open-reading frames ranging from 3027 to 3612 bp encoding proteins of 1008 to 1203 amino acids.

The computational analysis of the protein sequences of MdCHKS revealed the presence of the four basic conserved domains called Cyclases/Histidine kinases Associated SEnsory (CHASE), Histidine Kinase (HK), Receiver (REC), and Receiver-like (REC-like) (Figure 2 and Supplementary Figure S2). While these modular MdCHKS share a similar multidomain architecture in their cytosolic C-terminal part (HK, REC-like and REC domains), they differ in the TM domain topology of the N-terminus part. Indeed, the *MdCHK4a* and *MdCHK4b* have two predicted  $\alpha$ -helices transmembrane domains TM1 and TM2 (Supplementary Figure S3) bordering the CHASE domain (Figure 2). *MdCHK3a* and *MdCHK3b* possess equivalent TM1/CHASE/TM2 organization with an additional predicted transmembrane helice (TM3) in N-terminus. Finally, the N-terminus of *MdCHK2* includes a fourth predicted transmembrane domain

(TM4). Consequently, in addition of the extracytosolic loop containing the sensing CHASE domain, the TM3 and TM4 domains border a supplemental extracytosolic loop (140 aa) absent in *MdCHK3a/b* and *MdCHK4a/b* structures (Figure 2). Regarding the cytosolic part, each MdCHK possesses an HK domain containing the conserved phosphorylatable histidine residue as well as the C-terminal REC domain including the conserved phospho-accepting aspartate residue (Figure 2 and Supplementary Figure S2). In addition, the five receptors contain a second receiver domain located between the HK domain and the REC domain called pseudo-receiver domain or REC-like (Figure 2). Interestingly, the putative phospho-accepting aspartate residue in the REC-like domain of *MdCHK2* is conserved, suggesting that this domain may be functional in terms of phosphorelay reaction, whereas the corresponding residues in *MdCHK3a/MdCHK3b* and *MdCHK4a/MdCHK4b* are substituted with glutamate (Supplementary Figure S2).

## MdCHKS Function as Cytokinin Receptors in a Yeast Complementation Assay

To assess the function of the five MdCHKS as cytokinin receptors, we exploited the *S. cerevisiae sln1 $\Delta$*  deletion mutant strain which carries a lethal mutation in the *SLN1* gene encoding its unique osmosensing histidine kinase (Inoue et al., 2001).



The *sln1Δ* yeast mutants carrying pYES2-MdCHK2, pYES2-MdCHK4a or pYES2-MdCHK4b were lethal (Figure 3A). However, the addition of *trans*-zeatin (tZ) in culture medium allowed recovering a normal growth (Figure 3A). By depending on the presence and perception of cytokinin to complement *sln1* mutation, the three recombinant yeast strains clearly demonstrated that MdCHK2, MdCHK4a, and MdCHK4b act as cytokinin receptors in this heterologous system. Concerning the recombinant strains carrying pYES2-MdCHK3a and pYES2-MdCHK3b, they displayed an original phenotype since they exhibited a basal growth in absence of cytokinin, especially for MdCHK3b (Figure 3A and Supplementary Figure S4). However, the addition of tZ clearly induced the yeast growth pointing out that both MdCHK3a and MdCHK3b sense cytokinins (Figure 3A and Supplementary Figure S4). The basal constitutive activity of these two cytokinin receptors raised the question of their putative additional sensing function.

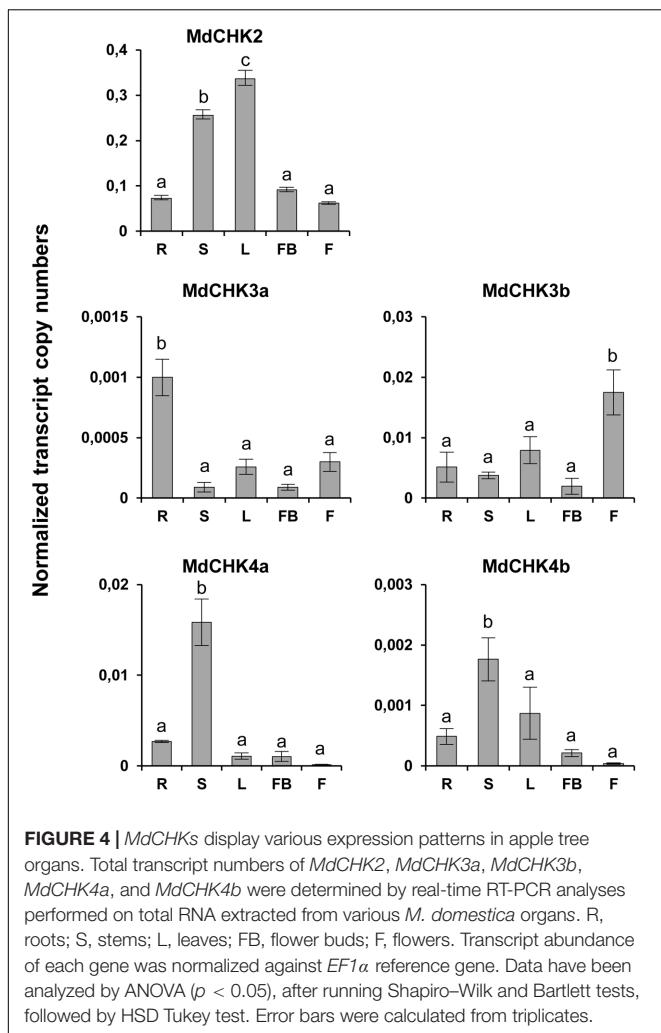
## MdCHK Receptors Show Different Binding Specificities toward Cytokinin

We further evaluated the substrate specificity and sensitivity of *M. domestica* cytokinin receptors. In this way, we measured the growth of the yeast cells in presence of various cytokinin-types at different concentrations including free-bases, ribosides, glucosides, and methylthio cytokinins. The specificity as well as the sensitivity (the minimal cytokinin concentration that induced yeast growth) were reported on Figure 3B and Supplementary Figure S4. Three distinct profiles corresponding to MdCHK2, MdCHK3, and MdCHK4 groups were observed.

First, MdCHK2 clearly perceived a wide range of cytokinin forms since each cytokinin type activated the receptor except tZ7G. Furthermore, this receptor presented a remarkable higher sensitivity than other MdCHks (until 0,1 nM for iP and 2MeSiP forms) and was the only one to be activated by some ribosides and glucosides cytokinin-types. Secondly, the strictly cytokinin-dependent MdCHK4a and MdCHK4b were activated by the free-bases iP and tZ and the methylthio-forms 2MeSiP, 2MeStZ, 2MeScZ, and 2MeSiPA (Supplementary Figure S4). High concentrations (10 μM) of cZ and DHZ were also effective on MdCHK4b. Nevertheless, the cytokinin-sensitivity of both MdCHK4a and MdCHK4b was obviously lower than MdCHK2. Finally, even if MdCHK3a and MdCHK3b showed a basal growth in absence of cytokinin, we were able to detect a significant difference of growth in presence of iP, tZ, DHZ, 2MeSiP, and 2MeStZ (Supplementary Figure S4).

## MdCHks Exhibit Distinct Expression Patterns

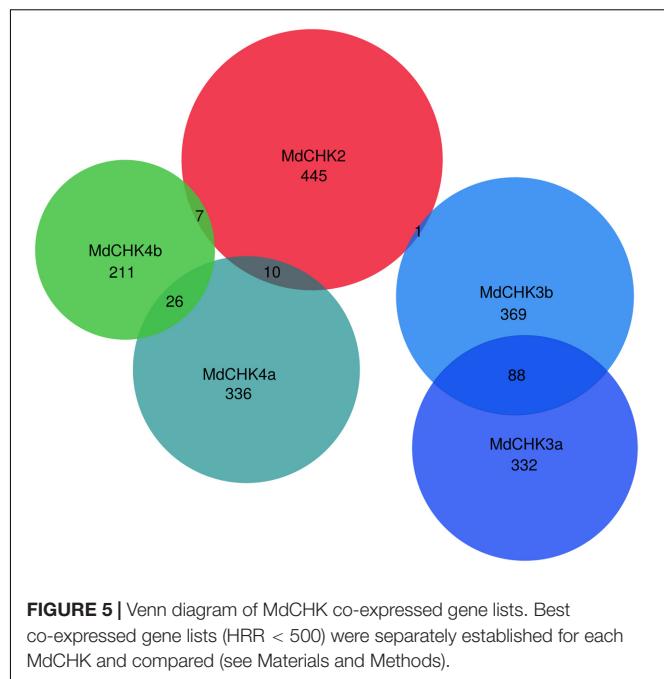
To examine the gene expression of *MdCHks*, RT-qPCR was carried out using distinct plant organs including roots, stems, leaves, flower buds and flowers. Transcripts of the five *MdCHks* were detected in all the tested organs, but with distinct expression pattern. Thus, *MdCHK2* reached higher expression level in leaves and stems (Figure 4). *MdCHK4a* and *MdCHK4b* displayed similar expression profiles with high expression level in stems whereas gene expression was hardly detected in flowers (Figure 4). *MdCHK3a* and *MdCHK3b* disclosed differential pattern of expression. While *MdCHK3a* was mainly



**FIGURE 4 |** *MdCHKs* display various expression patterns in apple tree organs. Total transcript numbers of *MdCHK2*, *MdCHK3a*, *MdCHK3b*, *MdCHK4a*, and *MdCHK4b* were determined by real-time RT-PCR analyses performed on total RNA extracted from various *M. domestica* organs. R, roots; S, stems; L, leaves; FB, flower buds; F, flowers. Transcript abundance of each gene was normalized against *EF1 $\alpha$*  reference gene. Data have been analyzed by ANOVA ( $p < 0.05$ ), after running Shapiro-Wilk and Bartlett tests, followed by HSD Tukey test. Error bars were calculated from triplicates.

expressed in roots, *MdCHK3b* showed its highest expression in flowers (Figure 4). Finally, substantial differences in the overall expression level of each *MdCHKs* were also observed. Interestingly, *MdCHK2* displayed the higher expression level whilst *MdCHK3b/MdCHK4a* and *MdCHK3a/MdCHK4b* retained a 10- and a 100-fold lower expression, respectively.

To complete qPCR analysis, we used the available RNAseq data to generate an expression matrix of apple tree genes (Supplementary Table S2). The best co-expressed genes with each *MdCHK* through our expression matrix were investigated and compared (Supplementary Table S3). A very weak degree of overlap among lists of genes co-expressed with each *MdCHK* was found. *MdCHK2* shared up to 18 genes with other cytokinin receptors. *MdCHK3* and *MdCHK4* groups did not have common co-expressed genes. *MdCHK3a/MdCHK3b* homologs as well as *MdCHK4a/MdCHK4b* homologs shared respectively 88 and 26 genes (Figure 5). This low overlapping of co-associated genes might support the limited functional redundancy of *MdCHKs*. Each *MdCHK* co-expressed genes list was compared to the list established for their *Arabidopsis* ortholog in order to highlight potential shared genes. We

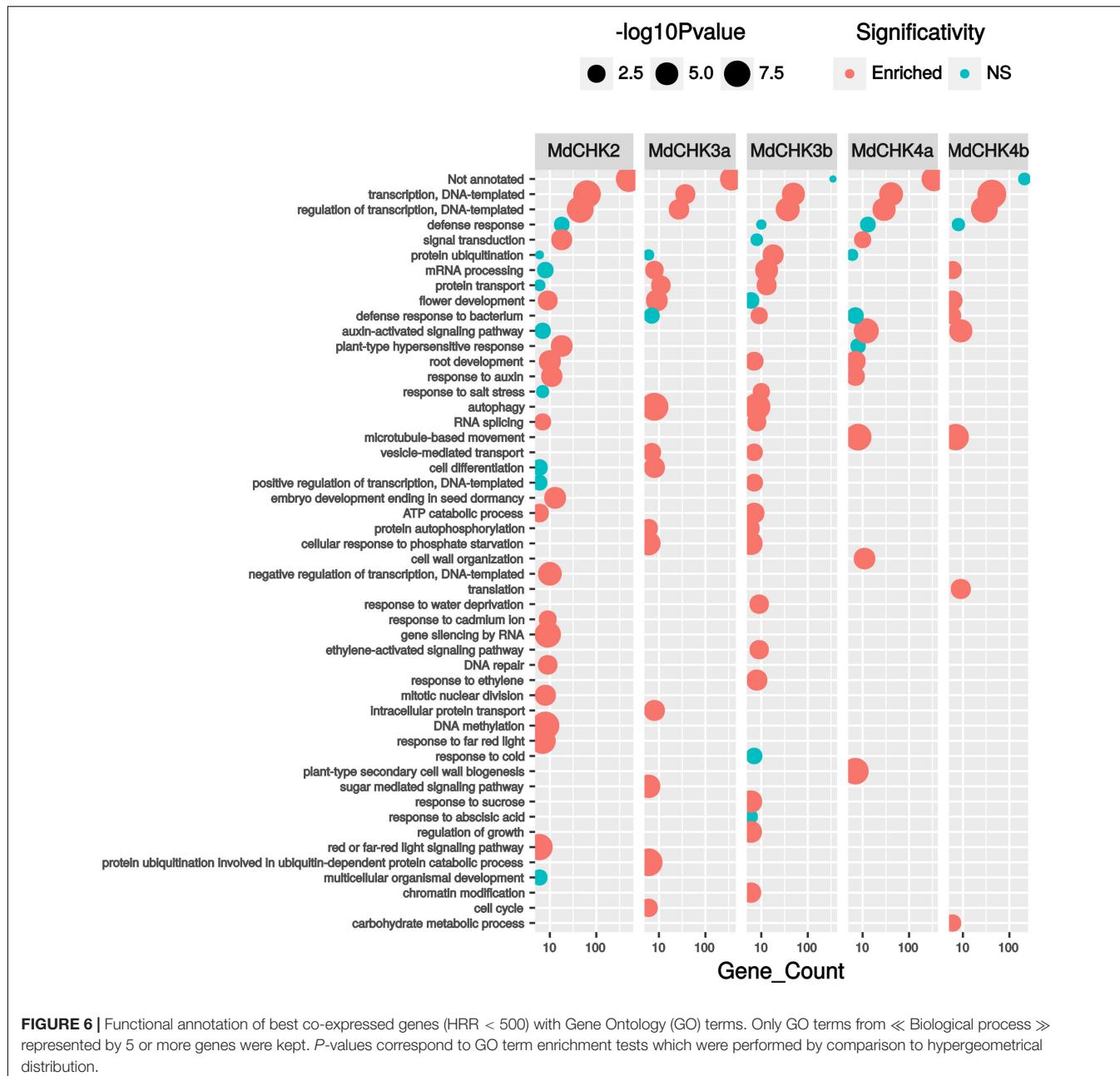


**FIGURE 5 |** Venn diagram of *MdCHK* co-expressed gene lists. Best co-expressed gene lists (HRR < 500) were separately established for each *MdCHK* and compared (see Materials and Methods).

found a relatively weak overlap between functions associated to either CHK (Supplementary Figure S5A and Table S4). For example, only 19 genes were similarly co-expressed between *MdCHK2* (1.8%) and *AtCHK2* (8.7%). While such a weak overlap could be due to the initial datasets used to calculate correlations which differ in size and experiments, conserved co-expressed genes may be good candidates for a further investigation of the cytokinin pathway. In addition, we found very small overlaps between co-expressed gene lists of *AtCHKs*, as observed for *MdCHKs*, reinforcing a potential specificity in CHK functions (Supplementary Figure S5B). Enrichment tests of Gene Ontology (GO) terms performed on each list of co-expressed genes also gave an overview of possible specific physiological processes associated with each receptor. For example, “Embryo development ending in seed dormancy” and “Response to cadmium ion” were exclusively enriched for *MdCHK2* whereas “Plant-type secondary cell wall biogenesis” and “Regulation of growth” were specifically enriched for *MdCHK4a* and *MdCHK3b*, respectively (Figure 6 and Supplementary Figure S6).

## MdCHKs Mainly Localize to the Endoplasmic Reticulum in Plant and Show a Dynamic Behavior in Response to Cytokinin in Yeast

We investigated the subcellular distribution of the *MdCHKs* using C-terminal YFP tagging to ensure the correct anchoring of the transmembrane domains. *MdCHK-YFP* constructs were transiently expressed in *C. roseus* cells that constitute a reliable model for studying protein subcellular localization (Foureau et al., 2016). In transiently transformed cells, the fusion proteins displayed a fluorescence signal located in

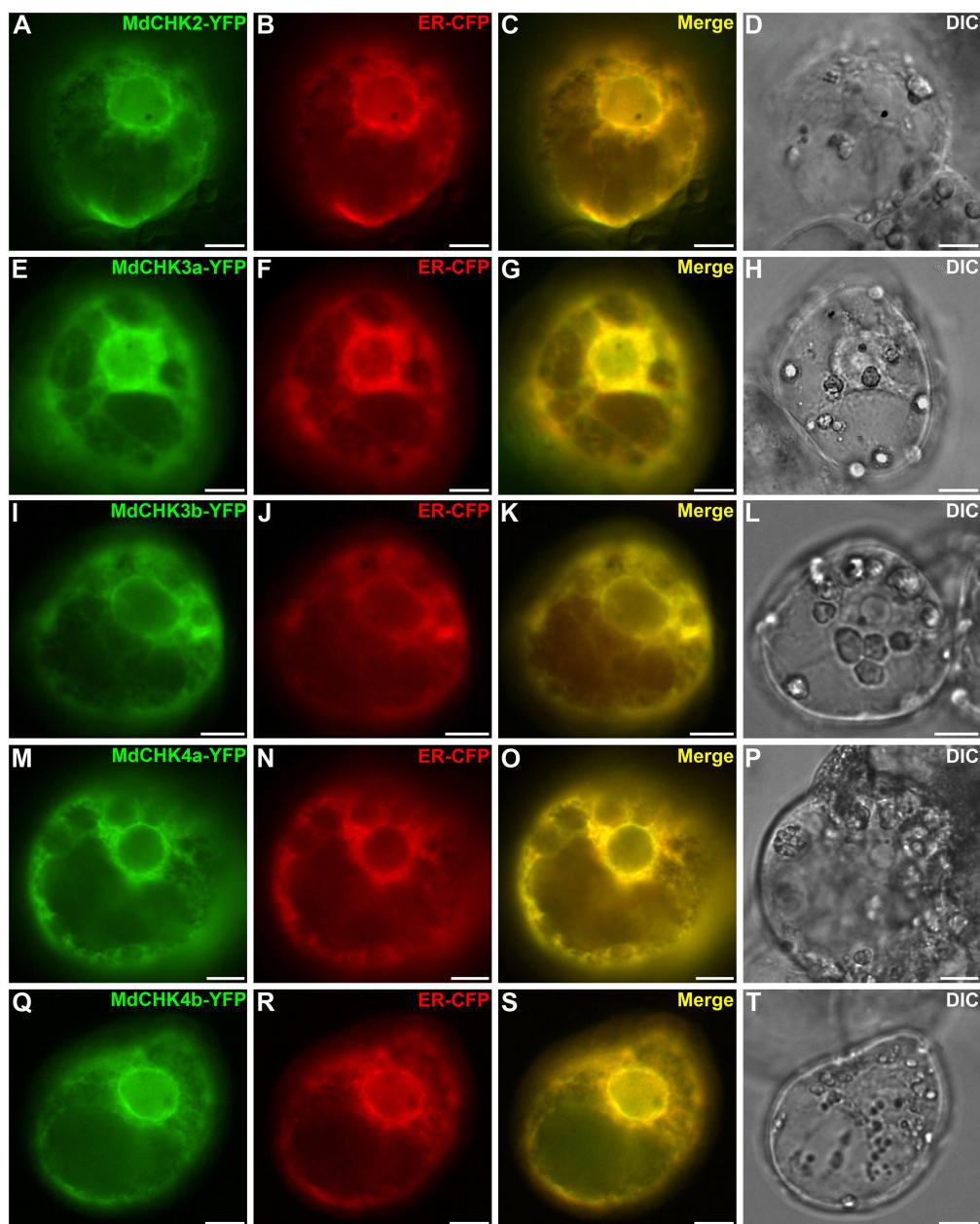


**FIGURE 6 |** Functional annotation of best co-expressed genes (HRR < 500) with Gene Ontology (GO) terms. Only GO terms from <> Biological process <> represented by 5 or more genes were kept. *P*-values correspond to GO term enrichment tests which were performed by comparison to hypergeometrical distribution.

the endoplasmic reticulum (ER) network throughout the cell as well as in the perinuclear space (**Figures 7A,E,I,M,Q**). The signal perfectly co-localized (**Figures 7C,G,K,O,S**) with the specific ER-CFP marker (**Figures 7B,F,J,N,R**), confirming that the five MdCHKs are located in the ER of plant cells. Noteworthy, cytokinin addition in plant cell medium did not alter the ER localization of the MdCHK-YFP fusions (data not shown).

Since plant cells may produce their own pool of cytokinins preventing the study influence of exogenous cytokinins on the localization of the MdCHKs, we therefore investigated the subcellular distribution of MdCHKs in the yeast *S. cerevisiae*,

by using YFP fusion proteins. In absence of cytokinins, a punctate fluorescence pattern was observed for the five MdCHKs (**Figures 8G1,I1,K1,M1,O1**), which accumulated in the ER forming structures comparable as organized smooth ER (Snapp et al., 2003) that is described to result from protein interactions. Upon cytokinin treatment, the ER localization of the five MdCHKs did not change. But interestingly, a reorganization of the fluorescent pattern was observed reflecting the decrease or disappearance of aggregate structures (**Figures 8H1,J1,L1,N1,P1**). Indeed, MdCHK2 and MdCHK4b displayed a strong perinuclear localization (**Figures 8H1,P1**). Concerning MdCHK3a, MdCHK3b, and MdCHK4a, the

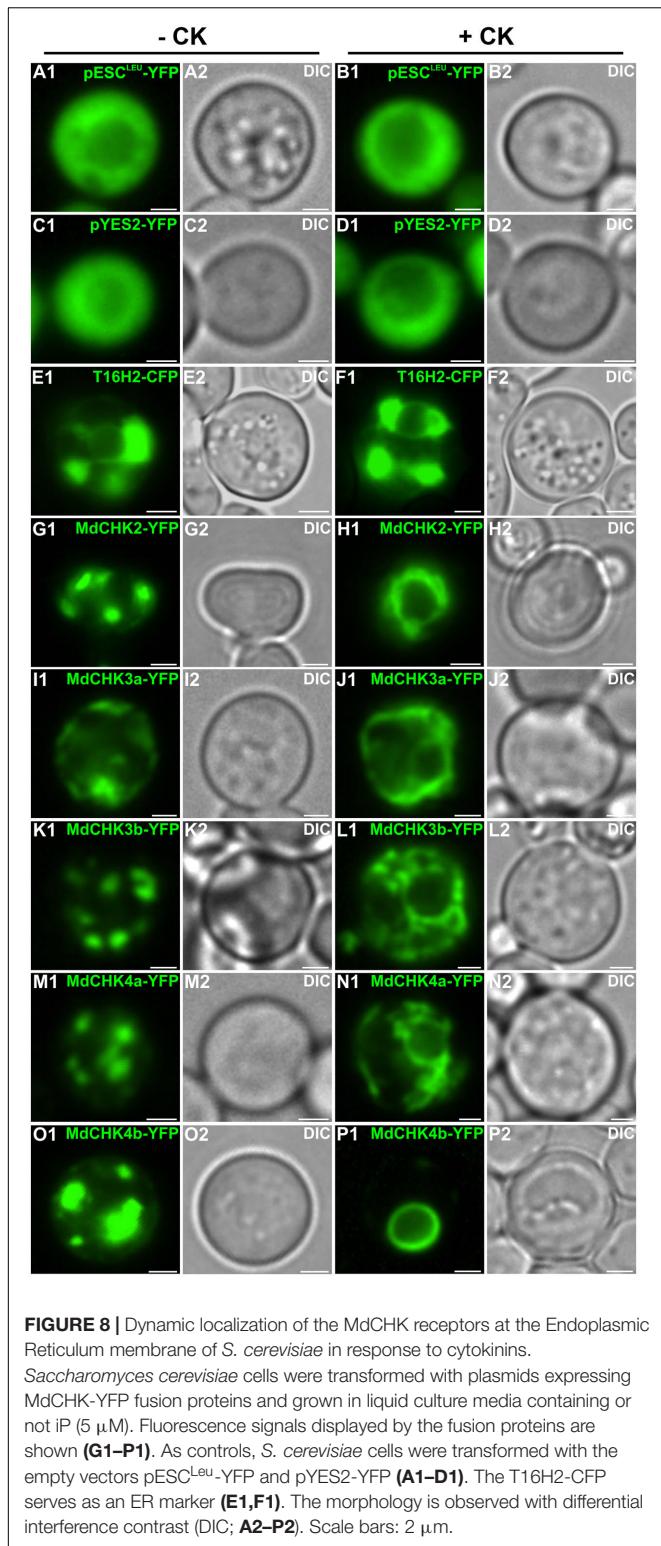


**FIGURE 7 |** MdCHK-YFP fusion proteins are localized at the Endoplasmic Reticulum (ER) membrane. *Catharanthus roseus* cells were transiently co-transformed with plasmids expressing MdCHK-YFP (**A,E,I,M,Q**) and endoplasmic reticulum-CFP marker (**B,F,J,N,R**). Co-localization of the two fluorescence signals appeared on the merged image (**C,G,K,O,S**). The morphology is observed with differential interference contrast (DIC; **D,H,L,P,T**). Scale bars: 10  $\mu$ m.

fluorescence signal appeared in a discontinuous pattern as well as in the perinuclear space (Figures 8J1,L1,N1). In order to ensure that cytokinins themselves had no impact on the architecture of the ER, we used T16H2-CFP construct as a specific ER marker (Besseau et al., 2013). Upon cytokinin treatment, no redistribution of fluorescence signal was observed, reinforcing the plausibility of a specific reorganization of MdCHKs in response to cytokinin signal (Figures 8E1,F1).

### Homo- and Heterodimerization Are Common Features of MdCHK2 and MdCHK4 in Contrast to the MdCHK3 Pair That Only Displays Specific Heterodimerization Characteristics

Cytokinin receptors were previously proposed to interact each other to enable the *trans*-phosphorylation of the HK domain after cytokinin perception (Dortay et al., 2006;



**FIGURE 8 |** Dynamic localization of the MdCHK receptors at the Endoplasmic Reticulum membrane of *S. cerevisiae* in response to cytokinins.

*Saccharomyces cerevisiae* cells were transformed with plasmids expressing MdCHK-YFP fusion proteins and grown in liquid culture media containing or not iP (5  $\mu$ M). Fluorescence signals displayed by the fusion proteins are shown (G1–P1). As controls, *S. cerevisiae* cells were transformed with the empty vectors pESC<sup>Leu</sup>-YFP and pYES2-YFP (A1–D1). The T16H2-CFP serves as an ER marker (E1,F1). The morphology is observed with differential interference contrast (DIC; A2–P2). Scale bars: 2  $\mu$ m.

Caesar et al., 2011; Hothorn et al., 2011; Wulfetange et al., 2011). Considering the multiple possibilities of interactions between the five MdCHks, homo- and hetero-dimerization were investigated by BiFC assays *in planta*. The full coding

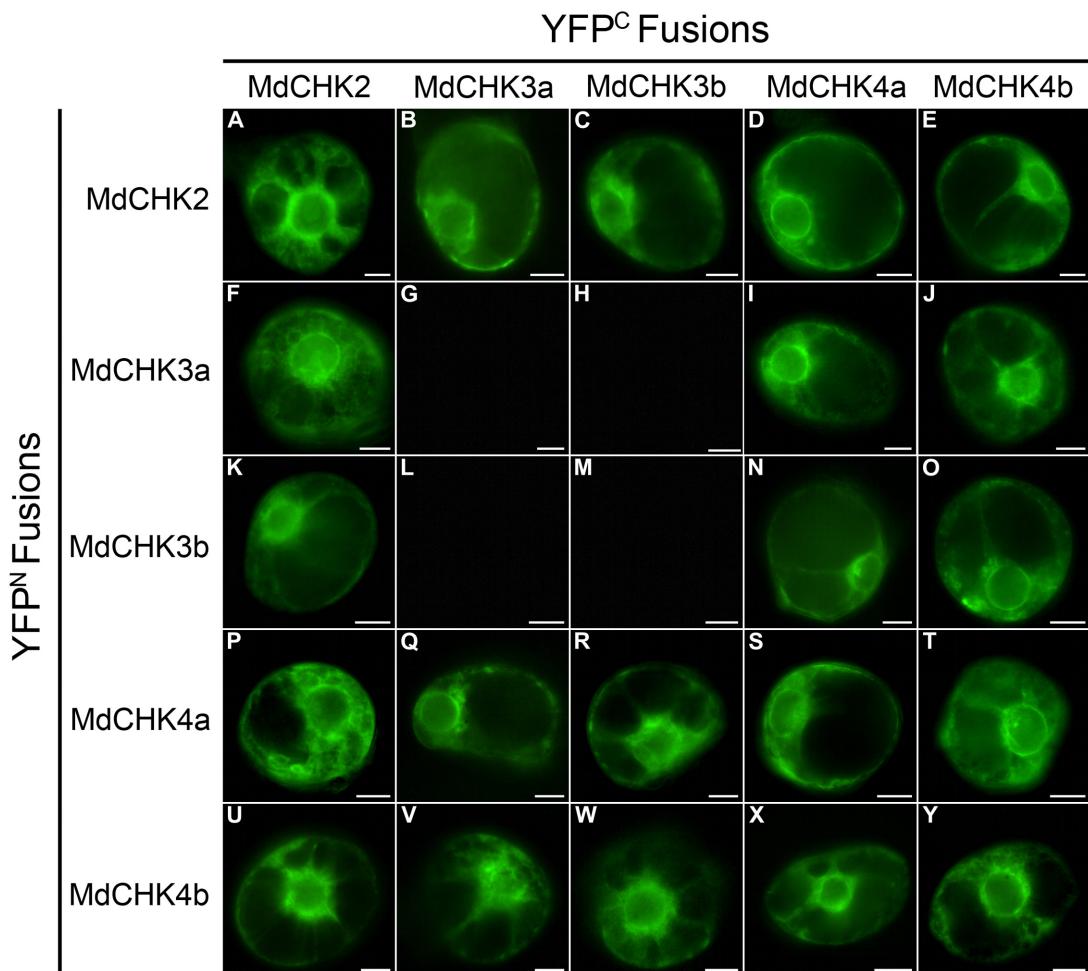
sequences of MdCHks were cloned upstream of the coding sequence of the two split-YFP fragments (YFP<sup>N</sup> and YFP<sup>C</sup>) to generate the MdCHK-YFP<sup>N</sup> and MdCHK-YFP<sup>C</sup> fusion proteins. BiFC analysis revealed that MdCHK2, MdCHK4a, and MdCHK4b were able to form homodimers within the ER network (Figures 9A,S,Y) whereas no BiFC complex reconstitution was observed when testing the MdCHK3a and MdCHK3b homodimers (Figures 9G,M). Moreover, no signal was detected with the MdCHK3a and MdCHK3b heterodimer combination (Figures 9H,L). Additionally, a cytokinin application did not result in the formation of a fluorescent signal within the three configurations (data not shown). Nevertheless, MdCHK3a and MdCHK3b were able to heterodimerize with MdCHK2, MdCHK4a and MdCHK4b within the ER (Figures 9B,C,E–K,N,O,Q,R,V,W). In addition, MdCHK2 and both MdCHK4 homologs shape heterodimers with each other (Figures 9D,E,P,U,T,X).

## DISCUSSION

Five CHASE domain-containing Histidine Kinases belonging to the three classical groups were identified in *M. domestica*: MdCHK2, the homolog of *Arabidopsis* AHK2 and the two pairs MdCHK3a/MdCHK3b and MdCHK4a/MdCHK4b, homologous to AHK3 and AHK4, respectively. These two pairs might be the direct consequence of apple genome-wide duplication (Velasco et al., 2010) as it was found in another hardwood tree *Populus trichocarpa* (Nieminan et al., 2008).

To confirm the functionality of MdCHks as cytokinin receptors, we conducted a cytokinin-responsive assay based on the use of the *sln1* $\Delta$  *S. cerevisiae* strain mutant (Maeda et al., 1994). We thus showed that MdCHK2, MdCHK4a, and MdCHK4b restore the viability of the *sln1* $\Delta$  mutant in a strictly cytokinin-dependent fashion providing convincing evidences of their cytokinin receptor function. By contrast, both MdCHK3a and MdCHK3b, they presented an unexpected profile since they conferred a basal growth to the *sln1* $\Delta$  mutant in absence of cytokinin. However, the enhancement of the yeast growth in presence of cytokinin confirmed their cytokinin receptor function. It is important to emphasize that the basal constitutive activity confers an originality for MdCHK3a and MdCHK3b compared to MdCHK2, MdCHK4a, and MdCHK4b. To our knowledge, a constitutive activity for a CHK protein has never been reported before.

The extensive exploitation of our five MdCHK-complemented yeast strains revealed that the receptors differed greatly in their cytokinin specificity and sensitivity. Three distinct specificity and sensitivity profiles clearly emerged. The most remarkable result comes from MdCHK2 which perceives an unprecedented range of cytokinins including nucleotide-type precursors as well as free-base-, riboside-, O-glucoside- and methylthio-forms, with a substantial sensitivity for iP, tZ, 2MeSiP, and 2MeStZ (Figure 3B and Supplementary Figure S4). Until now, due to its presumed toxicity in bacteria, few studies examined the ligand-binding properties of the full-length *Arabidopsis* AHK2 and its homologs in other plant species. Recently, an *E. coli* cytokinin-binding assay



**FIGURE 9 |** MdCHKs interact to form homodimers and heterodimers in BiFC assays. Cells of *C. roseus* were transiently co-transformed using the plasmids expressing the different MdCHK proteins fused with the YFP<sup>N</sup> and YFP<sup>C</sup> split in C-terminal. Homodimerizations (**A,G,M,S,Y**) and heterodimerizations (**B-F,H-L,N-R,T-X**) have been tested, with both YFP<sup>N</sup> and YFP<sup>C</sup> combinations. Three independent plasmid clones were used to test interactions. Scale bars: 10  $\mu$ m.

revealed that the full-length BnCHK1 and BnCHK3, two AHK2 homologs in *Brassica napus*, showed high affinity for tZ, iP, and tZR (Kuderová et al., 2015). A tobacco membrane assay also revealed that tZ and iP strongly interacted with AHK2, whereas their conjugated forms did not, suggesting that free bases were the sole biologically active cytokinin compounds (Lomin et al., 2015). Even if our experiments used a heterologous system in which the yeast membrane environment can potentially differ from those of plant, we highlighted not only the receptor ability to bind hormones but also their activation through cytokinin perception activating the phosphorelay in yeast. Thus, we can assume that our assay reflects the biological activity of cytokinin riboside and nucleotide forms on the receptor. Moreover, artefactual cytokinin activation or conversion occurrence in yeast can be omitted, since MdCHK4a and MdCHK4b are not activated in presence of the nucleotide-, riboside- or O-glucoside types. Therefore, the broad cytokinin spectrum of MdCHK2 raises the question of

its possible central role in *M. domestica*. Such hypothesis is also reinforced by the high expression level of MdCHK2 compared to other MdCHKs (Figure 4). Concerning MdCHK4a and MdCHK4b, they perceived a restricted spectrum of cytokinins such as free-base forms and some of the methylthiolated-forms with a lower sensitivity than MdCHK2 supporting previous works obtained with *Arabidopsis* AHK4 and AHK2 (Stolz et al., 2011). Concerning MdCHK3a and MdCHK3b, they perceived some cytokinin free-bases and methylthiolated forms. However, regarding their constitutive activity, we must consider that we could have under estimated their real cytokinin binding effectiveness and that it might not really reflect their complete capacity to perceive the diverse structures of cytokinins. The original activities of MdCHK3a and MdCHK3b were definitely interesting and need to be further investigated. In particular, did the monomeric or heterodimeric forms of MdCHK3a/b influence their cytokinin perception? In any case, our results

clearly supported previous works reporting that cytokinin-binding properties of AHK3 differed from those of AHK2 and AHK4 (Spíchal et al., 2004; Romanov et al., 2006; Stolz et al., 2011; Heyl et al., 2012).

These different properties between MdCHK receptors raised the question of their specialized functions in apple tree. MdCHks clearly showed an organ-specific gene expression pattern (**Figure 4**) and the analysis of gene co-expression profiles with each MdCHK unequivocally shed light on the weak degree of overlap among lists of co-expressed genes that strongly underlines the distinct roles of the five receptors in physiological processes (**Figure 5** and Supplementary Figure S6). Besides these distinct patterns, we can point out that both pairs of homologs MdCHK3 and MdCHK4 might acquire distinct functions after duplication.

From a structural point of view, the five MdCHks possess the conserved CHASE, HK, REC-like and REC domains also found in *Arabidopsis*, maize and rapeseed (Ueguchi et al., 2001; Yonekura-Sakakibara et al., 2004; Kuderová et al., 2015). While the architecture of cytoplasmic C-terminal part is similar within MdCHK, the topology of their N-terminus differs in the number of predicted transmembrane domains which surround the CHASE sensing domain (**Figure 2**). MdCHK4a/MdCHK4b and MdCHK3a/MdCHK3b possess two and three transmembrane domains, respectively, whereas MdCHK2 exhibits a fourth transmembrane helix that forms a unique additional extracytoplasmic loop. This variability might reflect the specific sensing activities of MdCHK receptors. It is well established that structural variations in the CHASE domain result in different ligand specificities of *Arabidopsis* receptors (Romanov et al., 2006; Heyl et al., 2007; Stolz et al., 2011). Nevertheless, the organization of the surrounding environment of the CHASE domain might also be important for receptor functioning (Steklov et al., 2013). Indeed, the CHASE flanking regions including transmembrane helices are assumed to play a substantial role in localization and intramolecular signaling (Steklov et al., 2013). Directed mutagenesis on a transmembrane helix highlighted its importance for the AHK4 receptor activation (Miwa et al., 2007). Moreover, experiments on the CHASE domain of AHK4 compared with the full-length receptor revealed significant differences in the binding affinity, highlighting the importance of the CHASE environment in cytokinin perception (Stolz et al., 2011). Thus, it cannot be excluded that architectural variations in the N-terminal part of the MdCHks somehow influence their distinct properties.

Regarding the cytoplasmic C-terminal part, only MdCHK2 harbors a phospho-accepting aspartate in the receiver-like domain (Supplementary Figure S2). Its presence is a common feature shared with AHK2 and the four AHK2 homologs in *B. napus* BnCHK1, BnCHK2, BnCHK3, and BnCHK4 (Ueguchi et al., 2001; Kuderová et al., 2015). However, the function of this receiver-like domain has not been yet elucidated. As reported above, MdCHK2 perceives a broad cytokinin spectrum with a substantial sensitivity compared to other MdCHks, thus it would be interesting to further investigate if the second phosphorylatable aspartate contribute

to the properties of this receptor. Indeed, if we consider that the MdCHK2 receiver-like domain is functional, it might optimize the phosphorelay reaction in addition to the receiver domain or might guide specific interactions with downstream HPts.

As previously described in other species, the five MdCHks are located at the ER membrane and the perinuclear space. Moreover, we revealed for the first time, a dynamic redistribution of cytokinin receptors in response to cytokinin application. The development of an approach in *S. cerevisiae* allowed us to overcome the use of plant cells, which probably produce their own cytokinins and prevent studying cytokinin influence. More precisely, we showed that MdCHks relocalized through the ER network, especially by getting closer to the nucleus (**Figure 8**). This result clearly supports the current concept of phosphotransfer enhancement through a perinuclear localization which overcomes intracellular distance and optimizes the signal transduction (Caesar et al., 2011; Wulfetange et al., 2011).

We also report herein a complete analysis of the full-length cytokinin receptors interactions *in planta*. Until now, only homodimerization of full-length AHK2 was demonstrated *in planta* (Wulfetange et al., 2011) and a partial study in yeast two-hybrid system based on full-length receptors showed the AHK3/AHK4 interaction as well as the formation of AHK3 homo-oligomers (Caesar et al., 2011). Here, we examined the homo- and the hetero-dimerization of the MdCHks *in planta* since histidine kinases are supposed to act as dimers. Not only MdCHK2, MdCHK4a, and MdCHK4b homodimerize, but they form heterodimers with each other. Surprisingly, MdCHK3a and MdCHK3b do not form homodimers. This feature might explain their singularity in cytokinin perception in our yeast system compared to MdCHK2, MdCHK4a, and MdCHK4b. Furthermore, MdCHK3a and MdCHK3b did not heterodimerize with each other, but exclusively heterodimerize with MdCHK2, MdCHK4a, and MdCHK4b. This particularity needs to be deeply addressed for the complete understanding of MdCHK3 functioning. To date, our overview clearly emphasized the complexity of cytokinin perception in *M. domestica* since MdCHks are able to form not only homodimers, but also heterodimers as well as monomers. Regarding the heterodimers in plant cells, they provided a new layer of intricacy since most of the histidine kinases form homodimers in order to autophosphorylate (Capra and Laub, 2012). Nevertheless, their physiological relevance *in planta* needs to be further determined. It cannot be excluded that CHK heterodimers might operate in cytokinin perception contributing to specify the cytokinin signaling pathways in order to regulate distinct physiological processes. For instance, the low gene expression of the MdCHK3 and MdCHK4 pairs compared to MdCHK2 suggested a MdCHK2 dimerization ratio in favor of homodimerization. Such homodimers could ensure signaling for the major physiological processes associated to cytokinins while MdCHK2 heterodimers would be associated to more discrete functions. In this way, the putative functions of MdCHK3 predicted through gene correlation analysis (**Figure 6**) would be assumed by

heterodimers with MdCHK2 or MdCHK4 since the MdCHK3 pair does not homo- or heterodimerize and is potentially not able to active the phosphorelay.

## CONCLUSION

This work provided a framework for further functional studies of cytokinin receptors in apple tree. In particular, it will be greatly interesting to focus on their involvement in response to the pathogens of apple tree. Furthermore, a structural approach would also contribute to gain insights into the key aspects of the mechanisms by which MdCHks are differentially activated by cytokinin signal.

## AUTHOR CONTRIBUTIONS

DD, EA, CM, and GG conducted experiments. FL and TDdB achieved bioinformatics analyses. NP, VC, AO, AL, MCl, OP, and SB participated in the design of the study and interpretation. DG, SC, NG-G, MCo, JC, and SB assisted in the supervision of this work. GG conceived, supervised and coordinated the work. DD and GG wrote the first draft of the manuscript, to which all authors contributed.

## REFERENCES

- Anantharaman, V., and Aravind, L. (2001). The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors. *Trends Biochem. Sci.* 26, 579–582. doi: 10.1016/S0968-0004(01)01968-5
- Argueso, C. T., Ferreira, F. J., Epple, P., To, J. P. C., Hutchison, C. E., Schaller, G. E., et al. (2012). Two-component elements mediate interactions between cytokinin and salicylic acid in plant immunity. *PLOS Genet.* 8:e1002448. doi: 10.1371/journal.pgen.1002448
- Besseau, S., Kellner, F., Lanoue, A., Thamm, A. M. K., Salim, V., Schneider, B., et al. (2013). A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in *Catharanthus roseus*. *Plant Physiol.* 163, 1792–1803. doi: 10.1104/pp.113.222828
- Boivin, S., Kazmierczak, T., Brault, M., Wen, J., Gamas, P., Mysore, K. S., et al. (2016). Different cytokinin histidine kinase receptors regulate nodule initiation as well as later nodule developmental stages in *Medicago truncatula*. *Plant Cell Environ.* 39, 2198–2209. doi: 10.1111/pce.12779
- Caesar, K., Thamm, A. M. K., Witthoft, J., Elgass, K., Huppenberger, P., Grefen, C., et al. (2011). Evidence for the localization of the *Arabidopsis* cytokinin receptors AHK3 and AHK4 in the endoplasmic reticulum. *J. Exp. Bot.* 62, 5571–5580. doi: 10.1093/jxb/err238
- Capra, E. J., and Laub, M. T. (2012). Evolution of two-component signal transduction systems. *Annu. Rev. Microbiol.* 66, 325–347. doi: 10.1146/annurev-micro-092611-150039
- Cortleven, A., Nitschke, S., Klaumunzer, M., AbdElgawad, H., Asard, H., Grimm, B., et al. (2014). A novel protective function for cytokinin in the light stress response is mediated by the *Arabidopsis* histidine kinase2 and *Arabidopsis* histidine kinase3 receptors. *Plant Physiol.* 164, 1470–1483. doi: 10.1104/pp.113.224667
- Cutcliffe, J. W., Hellmann, E., Heyl, A., and Rashotte, A. M. (2011). CRFs form protein-protein interactions with each other and with members of the cytokinin signalling pathway in *Arabidopsis* via the CRF domain. *J. Exp. Bot.* 62, 4995–5002. doi: 10.1093/jxb/err199
- Dortay, H., Mehnert, N., Bürkle, L., Schmülling, T., and Heyl, A. (2006). Analysis of protein interactions within the cytokinin-signaling pathway of *Arabidopsis thaliana*. *FEBS J.* 273, 4631–4644. doi: 10.1111/j.1742-4658.2006.05467.x
- Foureau, E., Carqueijeiro, I., de Bernonville, T. D., Melin, C., Lafontaine, F., Besseau, S., et al. (2016). Prequels to synthetic biology: from candidate gene identification and validation to enzyme subcellular localization in plant and yeast cells. *Methods Enzymol.* 516, 167–206. doi: 10.1016/bs.mie.2016.02.013
- Frugier, F., Kosuta, S., Murray, J. D., Crespi, M., and Szczyglowski, K. (2008). Cytokinin: secret agent of symbiosis. *Trends Plant Sci.* 13, 115–120. doi: 10.1016/j.tplants.2008.01.003
- Ginis, O., Oudin, A., Guirimand, G., Chebbi, M., Courdavault, V., Glévarec, G., et al. (2012). A type-B response regulator drives the expression of the hydroxymethylbutenyl diphosphate synthase gene in periwinkle. *J. Plant Physiol.* 169, 1571–1574. doi: 10.1016/j.jplph.2012.07.008
- Giron, D., and Glévarec, G. (2014). Cytokinin-induced phenotypes in plant-insect interactions: learning from the bacterial world. *J. Chem. Ecol.* 40, 826–835. doi: 10.1007/s10886-014-0466-5
- Giron, D., Kaiser, W., Imbault, N., and Casas, J. (2007). Cytokinin-mediated leaf manipulation by a leafminer caterpillar. *Biol. Lett.* 3, 340–343. doi: 10.1098/rsbl.2007.0051
- Guirimand, G., Burlat, V., Oudin, A., Lanoue, A., St-Pierre, B., and Courdavault, V. (2009). Optimization of the transient transformation of *Catharanthus roseus* cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. *Plant Cell Rep.* 28, 1215–1234. doi: 10.1007/s00299-009-0722-2
- Guirimand, G., Courdavault, V., Lanoue, A., Mahroug, S., Guihur, A., Blanc, N., et al. (2010). Strictosidine activation in Apocynaceae: towards a ‘nuclear time bomb’? *BMC Plant Biol.* 10:182. doi: 10.1186/1471-2229-10-182
- Held, M., Hou, H., Miri, M., Huynh, C., Ross, L., Hossain, M. S., et al. (2014). *Lotus japonicus* cytokinin receptors work partially redundantly to mediate nodule formation. *Plant Cell* 26, 678–694. doi: 10.1105/tpc.113.119362
- Heyl, A., Riefler, M., Romanov, G. A., and Schmülling, T. (2012). Properties, functions and evolution of cytokinin receptors. *Eur. J. Cell Biol.* 91, 246–256. doi: 10.1016/j.ejcb.2011.02.009

## FUNDING

This study was supported by the Région Centre-Val de Loire, France (SiSCyLi grant). Doctoral Fellow attributed to DD was jointly funded by the Région Centre-Val de Loire, France and the Ministère de l'Enseignement Supérieur et de la Recherche, France.

## ACKNOWLEDGMENTS

We thank Marie-Antoinette Marquet, Evelyne Danos, and Emeline Marais (EA2106 Biomolécules et Biotechnologies Végétales) for their help in maintaining cell cultures and Emilien Foureau for his technical help in subcellular localization. We also thank François Héricourt (LBLGC) for discussions concerning Histidine-Kinase receptors. We would also like to acknowledge the Fédération CaSciModOT (CCSC, Orléans, France) for accessing the Région Centre-Val de Loire computing grid.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fpls.2017.01614/full#supplementary-material>

- Heyl, A., Wulfetange, K., Pils, B., Nielsen, N., Romanov, G. A., and Schmülling, T. (2007). Evolutionary proteomics identifies amino acids essential for ligand-binding of the cytokinin receptor CHASE domain. *BMC Evol. Biol.* 7:62. doi: 10.1186/1471-2148-7-62
- Hofmann, K., and Stoffel, W. (1993). TM base-A database of membrane spanning proteins segments. *Biol. Chem. Hoppe Seyler* 374, 166.
- Hothorn, M., Dabi, T., and Chory, J. (2011). Structural basis for cytokinin recognition by *Arabidopsis thaliana* histidine kinase 4. *Nat. Chem. Biol.* 7, 766–768. doi: 10.1038/nchembio.667
- Inoue, T., Higuchi, M., Hashimoto, Y., Seki, M., Kobayashi, M., Kato, T., et al. (2001). Identification of CRE1 as a cytokinin receptor from *Arabidopsis*. *Nature* 409, 1060–1063. doi: 10.1038/35059117
- Jeon, J., Cho, C., Lee, M. R., Van Binh, N., and Kim, J. (2016). CYTOKININ RESPONSE FACTOR 2 (CRF2) and CRF3 regulate lateral root development in response to cold stress in *Arabidopsis*. *Plant Cell* 28, 1828–1843. doi: 10.1105/tpc.15.00909
- Jung, S., Ficklin, S. P., Lee, T., Cheng, C. H., Blenda, A., Zheng, P., et al. (2014). The genome database for rosaceae (GDR): year 10 update. *Nucleic Acids Res.* 42, 1237–1244. doi: 10.1093/nar/gkt1012
- Kaiser, W., Huguet, E., Casas, J., Commin, C., and Giron, D. (2010). Plant green-island phenotype induced by leaf-miners is mediated by bacterial symbionts. *Proc. Biol. Sci.* 277, 2311–2319. doi: 10.1098/rspb.2010.0214
- Kieber, J. J., and Schaller, G. E. (2014). Cytokinins. *Arabidopsis Book* 12:e0168. doi: 10.1199/tab.0168
- Kim, H. J., Ryu, H., Hong, S. H., Woo, H. R., Lim, P. O., Lee, I. C., et al. (2006). Cytokinin-mediated control of leaf longevity by AHK3 through phosphorylation of ARR2 in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 814–819. doi: 10.1073/pnas.0505150103
- Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315
- Kumar, M. N., and Verslues, P. E. (2015). Stress physiology functions of the *Arabidopsis* histidine kinase cytokinin receptors. *Physiol. Plant.* 154, 369–380. doi: 10.1111/ppl.12290
- Kudrová, A., Gallová, L., Kuricová, K., Nejedlá, E., Čurdová, A., Micenková, L., et al. (2015). Identification of AHK2- and AHK3-like cytokinin receptors in *Brassica napus* reveals two subfamilies of AHK2 orthologues. *J. Exp. Bot.* 66, 339–353. doi: 10.1093/jxb/eru422
- Letunic, I., Doerks, T., and Bork, P. (2015). SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* 43, 257–260. doi: 10.1093/nar/gku949
- Lomin, S. N., Krivosheev, D. M., Steklov, M. Y., Arkhipov, D. V., Osolodkin, D. I., Schmülling, T., et al. (2015). Plant membrane assays with cytokinin receptors underpin the unique role of free cytokinin bases as biologically active ligands. *J. Exp. Bot.* 66, 1851–1863. doi: 10.1093/jxb/eru522
- Lomin, S. N., Krivosheev, D. M., Steklov, M. Y., Osolodkin, D. I., and Romanov, G. A. (2012). Receptor properties and features of cytokinin signaling. *Acta Naturae* 43, 31–45.
- Lomin, S. N., Yonekura-Sakakibara, K., Romanov, G. A., and Sakakibara, H. (2011). Ligand-binding properties and subcellular localization of maize cytokinin receptors. *J. Exp. Bot.* 62, 5149–5159. doi: 10.1093/jxb/err220
- Maeda, T., Wurgler-Murphy, S. M., and Saito, H. (1994). A two-component system that regulates an osmosensing MAP kinase cascade in yeast. *Nature* 369, 242–245. doi: 10.1038/369242a0
- Mason, M. G., Mathews, D. E., Argyros, D. A., Maxwell, B. B., Kieber, J. J., Alonso, J. M., et al. (2005). Multiple type-B response regulators mediate cytokinin signal transduction in *Arabidopsis*. *Plant Cell* 17, 3007–3018. doi: 10.1105/tpc.105.035451
- McKenzie, D. J., McLean, M. A., Mukerji, S., and Green, M. (1997). Improved RNA extraction from woody plants for the detection of viral pathogens by reverse transcription-polymerase chain reaction. *Plant Dis.* 81, 222–226. doi: 10.1094/PDIS.1997.81.2.222
- Miwa, K., Ishikawa, K., Terada, K., Yamada, H., Suzuki, T., Yamashino, T., et al. (2007). Identification of amino acid substitutions that render the *Arabidopsis* cytokinin receptor histidine kinase AHK4 constitutively active. *Plant Cell Physiol.* 48, 1809–1814. doi: 10.1093/pcp/pcm145
- Mougel, C., and Zhulin, I. B. (2001). CHASE: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants. *Trends Biochem. Sci.* 26, 582–584. doi: 10.1016/S0968-0004(01)01969-7
- Naseem, M., Wölfling, M., and Dandekar, T. (2014). Cytokinins for immunity beyond growth, galls and green islands. *Trends Plant Sci.* 19, 481–484. doi: 10.1016/j.tplants.2014.04.001
- Nieminen, K., Imanen, J., Laxell, M., Kauppinen, L., Tarkowski, P., Dolezal, K., et al. (2008). Cytokinin signaling regulates cambial development in poplar. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20032–20037. doi: 10.1073/pnas.0805617105
- Osugi, A., and Sakakibara, H. (2015). Q&A: How do plants respond to cytokinins and what is their importance? *BMC Biol.* 13:102. doi: 10.1186/s12915-015-0214-5
- Papadopoulos, J. S., and Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* 23, 1073–1079. doi: 10.1093/bioinformatics/btm076
- Pertry, I., Václavíková, K., Depuydt, S., Galuszka, P., Spíchal, L., Temmerman, W., et al. (2009). Identification of *Rhodococcus fascians* cytokinins and their modus operandi to reshape the plant. *Proc. Natl. Acad. Sci. U.S.A.* 106, 929–934. doi: 10.1073/pnas.0811683106
- Raines, T., Shanks, C., Cheng, C. Y., McPherson, D., Argueso, C. T., Kim, H. J., et al. (2016). The cytokinin response factors modulate root and shoot growth and promote leaf senescence in *Arabidopsis*. *Plant J.* 85, 134–147. doi: 10.1111/tpj.13097
- Rambaldi, D., and Ciccarelli, F. D. (2009). FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics* 25, 2281–2282. doi: 10.1093/bioinformatics/btp381
- Riefler, M., Novak, O., Strnad, M., and Schmülling, T. (2006). *Arabidopsis* cytokinin receptor mutants reveal functions in shoot growth, leaf senescence, seed size, germination, root development, and cytokinin metabolism. *Plant Cell* 18, 40–54. doi: 10.1105/tpc.105.037796
- Romanov, G. A., Lomin, S. N., and Schmülling, T. (2006). Biochemical characteristics and ligand-binding properties of *Arabidopsis* cytokinin receptor AHK3 compared to CRE1/AHK4 as revealed by a direct binding assay. *J. Exp. Bot.* 57, 4051–4058. doi: 10.1093/jxb/erl179
- Sakakibara, H. (2006). Cytokinins: activity, biosynthesis, and translocation. *Annu. Rev. Plant Biol.* 57, 431–449. doi: 10.1146/annurev.arplant.57.032905.105231
- Schäfer, M., Meza-Canales, I. D., Brüttig, C., Baldwin, I. T., and Meldau, S. (2015). Cytokinin concentrations and CHASE-domain containing His Kinase 2 (NaCHK2)- and NaCHK3-mediated perception modulate herbivory-induced defense signaling and defenses in *Nicotiana attenuata*. *New Phytol.* 207, 645–658. doi: 10.1111/nph.13404
- Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., et al. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* 3, 265–274. doi: 10.1093/bib/3.3.265
- Snapp, E. L., Hegde, R. S., Francolini, M., Lombardo, F., Colombo, S., Pedrazzini, E., et al. (2003). Formation of stacked ER cisternae by low affinity protein interactions. *J. Cell Biol.* 163, 257–269. doi: 10.1083/jcb.200306020
- Spíchal, L. (2012). Cytokinins - recent news and views of evolutionarily old molecules. *Funct. Plant Biol.* 39, 267–284. doi: 10.1071/FP11276
- Spíchal, L., Rakova, N. Y., Riefler, M., Mizuno, T., Romanov, G. A., Strnad, M., et al. (2004). Two cytokinin receptors of *Arabidopsis thaliana*, CRE1/AHK4 and AHK3, differ in their ligand specificity in a bacterial assay. *Plant Cell Physiol.* 45, 1299–1305. doi: 10.1093/pcp/pch132
- Steklov, M. Y., Lomin, S. N., Osolodkin, D. I., and Romanov, G. A. (2013). Structural basis for cytokinin receptor signaling: an evolutionary approach. *Plant Cell Rep.* 32, 781–793. doi: 10.1007/s00299-013-1408-3
- Stolz, A., Riefler, M., Lomin, S. N., Achazi, K., Romanov, G. A., and Schmülling, T. (2011). The specificity of cytokinin signalling in *Arabidopsis thaliana* is mediated by differing ligand affinities and expression profiles of the receptors. *Plant J.* 67, 157–168. doi: 10.1111/j.1365-313X.2011.04584.x
- Tirichine, L., Sandal, N., Madsen, L. H., Radutoiu, S., Albrechtsen, A. S., Sato, S., et al. (2007). A gain-of-function mutation in a cytokinin receptor triggers spontaneous root nodule organogenesis. *Science* 315, 104–107. doi: 10.1126/science.1132397
- To, J. P., Haberer, G., Ferreira, F. J., Deruère, J., Mason, M. G., Schaller, G. E., et al. (2004). Type-A *Arabidopsis* response regulators are partially redundant negative regulators of cytokinin signaling. *Plant Cell* 16, 658–671. doi: 10.1105/tpc.018978

- Tran, L.-S. P., Urao, T., Qin, F., Maruyama, K., Kakimoto, T., Shinozaki, K., et al. (2007). Functional analysis of AHK1/ATHK1 and cytokinin receptor histidine kinases in response to abscisic acid, drought, and salt stress in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20623–20628. doi: 10.1073/pnas.0706547105
- Ueguchi, C., Koizumi, H., Suzuki, T., and Mizuno, T. (2001). Novel family of sensor histidine kinase genes in *Arabidopsis thaliana*. *Plant Cell Physiol.* 42, 231–235. doi: 10.1093/pcp/pce015
- Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., et al. (2012). Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.* 158, 590–600. doi: 10.1104/pp.111.189514
- Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nat. Genet.* 42, 833–839. doi: 10.1038/ng.654
- von Schwartzenberg, K., Lindner, A. C., Gruhn, N., Šimura, J., Novák, O., Strnad, M., et al. (2016). CHASE domain-containing receptors play an essential role in the cytokinin response of the moss *Physcomitrella patens*. *J. Exp. Bot.* 67, 667–679. doi: 10.1093/jxb/erv479
- Waadt, R., and Kudla, J. (2008). In planta visualization of protein interactions using bimolecular fluorescence complementation (BiFC). *CSH Protocol.* 2008:db.rot4995. doi: 10.1101/pdb.prot4995
- Wulfetange, K., Lomin, S. N., Romanov, G. A., Stolz, A., Heyl, A., and Schmülling, T. (2011). The cytokinin receptors of *Arabidopsis* are located mainly to the endoplasmic reticulum. *Plant Physiol.* 156, 1808–1818. doi: 10.1104/pp.111.180539
- Yonekura-Sakakibara, K., Kojima, M., Yamaya, T., and Sakakibara, H. (2004). Molecular characterization of cytokinin-responsive histidine kinases in maize. Differential ligand preferences and response to *cis*-zeatin. *Plant Physiol.* 134, 1654–1661. doi: 10.1104/pp.103.037176
- Zhang, H., de Bernonville, T. D., Body, M., Glevarec, G., Reichelt, M., Unsicker, S., et al. (2016). Leaf-mining by Phyllonorycter blancardella reprograms the host-leaf transcriptome to modulate phytohormones associated with nutrient mobilization and plant defense. *J. Insect Physiol.* 84, 114–127. doi: 10.1016/j.jinsphys.2015.06.003
- Zürcher, E., and Müller, B. (2016). Cytokinin synthesis, signaling, and function—advances and new insights. *Int. Rev. Cell Mol. Biol.* 324, 1–38. doi: 10.1016/bs.ircmb.2016.01.001
- Zwack, P. J., and Rashotte, A. M. (2015). Interactions between cytokinin signalling and abiotic stress responses. *J. Exp. Bot.* 66, 4863–4871. doi: 10.1093/jxb/erv172

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

Copyright © 2017 Daudu, Allion, Liesecke, Papon, Courdavault, Dugé de Bernonville, Mélin, Oudin, Clastre, Lanoue, Courtois, Pichon, Giron, Carpin, Giglioli-Guivarc'h, Crèche, Besseau and Glévarec. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## “Coupable par association”

Exploitation de ressources transcriptomiques pour la construction de réseaux de co-expression de gènes dédiés à l’élucidation de voies cellulaires

## Résumé

Avec l’essor de technologies transcriptomiques à haut débit, une grande quantité de données a été générée. Ce travail est axé sur la réutilisation de ces données publiquement disponibles, pour la construction de réseaux de co-expression de gènes et leur exploitation dans l’élucidation de voies métaboliques et de signalisation. Ce travail, dont l’objectif a été de fournir une méthodologie pour l’identification de gènes candidats, est axé autour de trois aspects : (i) le choix d’une distance appropriée pour évaluer la similarité de profils d’expression entre gènes, (ii) l’identification du nombre d’échantillons minimal à inclure dans la matrice d’expression, et (iii) la comparaison de réseaux, de type *Pathway Level Co-expression* (PLC) de différentes espèces et construits, avec les gènes codant les acteurs de la voie *Multi Step Phosphorelay* (MSP) comme guides.

**Mots-clés:** Réseau de co-expression, transcriptomique, voies métaboliques, voies de signalisation, larges données

## Abstract

With the rise of high throughput technologies able to provide a large-scale view of transcriptomes, a high amount data has been produced. This work focuses on publicly available data reuse to construct gene co-expression networks for metabolic or signalling pathways elucidation. The final aim of this work, was to provide a methodology for candidate gene identification and thus focuses on (i) the choice of an appropriated distance to evaluate similarity between gene expression profiles, (ii) the identification of a minimal number of samples to be included in the expression matrix in order to construct robust co-expression networks, and finally (iii) the comparison of targeted co-expression networks built with the *Pathway Level Co-expression* (PLC) approach and using guide genes coding actors of the *Multi Step Phosphorelay* (MSP) among different species.

**Keywords :** Co-expression network, transcriptomics, metabolic pathways, signaling pathways, large-scale data