Inférence des réseaux de régulation de la synthèse des protéines de réserve du grain de blé tendre (Triticum aestivum L.) en réponse à l'approvisionnement en azote et en soufre

Jonathan Vincent

▶ **To cite this version:**

Jonathan Vincent. Inférence des réseaux de régulation de la synthèse des protéines de réserve du grain de blé tendre (Triticum aestivum L.) en réponse à l'approvisionnement en azote et en soufre : Inférence des réseaux de régulation de la synthèse des protéines de réserve du grain de blé tendre (Triticum aestivum L.) en réponse à l'approvisionnement en azote et en soufre. Sciences agricoles. Université Blaise Pascal - Clermont-Ferrand II, 2014. Français. NNT : 2014CLF22485 . tel-01170075

HAL Id: tel-01170075

https://theses.hal.science/tel-01170075

Submitted on 1 Jul 2015

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

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

Université Blaise Pascal **Université d'Auvergne**

N° D.U. 2485

THESE DE DOCTORAT

Présentée à l'Université Blaise Pascal pour l'obtention du grade de

DOCTEUR D'UNIVERSITE

Spécialité : Physiologie et génétique moléculaires

Ecole Doctorale Sciences de la Vie, Santé, Agronomie, Environnement
N° d'Ordre : 637

Soutenue le 10 septembre 2014 par :

Jonathan VINCENT

Inférence des réseaux de régulation de la synthèse des protéines de réserve du grain de blé tendre (*Triticum aestivum* L.) en réponse à l'approvisionnement en azote et en soufre.

Directeur de thèse : Pierre Martre

Co-encadrante de thèse : Marie Pailloux

Membres du Jury

TATOUT Christophe	Professeur Université Blaise Pascal, Clermont-Ferrand	Président
BUITINK Julia	Directeur de Recherche INRA, Angers	Rapporteur
ROTHAN Christophe	Directeur de Recherche INRA, Bordeaux	Rapporteur
DUBREUCQ Bertrand	Directeur de Recherche INRA, Versailles	Examinateur
PAILLOUX Marie	Maître de Conférences Université Blaise Pascal, Clermont-Ferrand	Examinateur
MARTRE Pierre	Directeur de Recherche INRA, Clermont-Ferrand	Examinateur

GDEC UMR 1095 INRA-UBP, LIMOS UMR 6158 CNRS-UBP

“Génétique, Diversité et Ecophysiologie des Céréales”

5, chemin de Beaulieu, F-63 039 Clermont-Ferrand, France

“Laboratoire d’Informatique, de Modélisation et d’Optimisation des Systèmes”

Complexe Scientifique des Cézeaux, F-63 173 Aubière, France

AVANT-PROPOS

Le travail de thèse présenté dans ce manuscrit est multidisciplinaire, il s'inscrit dans plusieurs axes majeurs de recherche : la physiologie, la génétique moléculaire et la biologie computationnelle. Le projet en partie porté par cette thèse est le fruit d'une collaboration entre l'équipe Biologie Intégrative du Grain du GDEC (Génétique, Diversité et Ecophysiologie des Céréales, UMR 1095 INRA-UBP) et le LIMOS (Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, UMR 6158 CNRS-UBP). L'essentiel de la thèse a été financé par une allocation de recherche du Ministère de l'Enseignement Supérieur et de la Recherche. Les trois derniers mois de travaux ont été soutenus financièrement par le département Biologie et Amélioration des Plantes de l'INRA. La thèse est officiellement labellisée depuis le 14 avril 2011 par le pôle de compétitivité Céréales Vallée.

La démarche avait été initiée au cours de mon stage de Master Recherche en génétique et physiologie. Le travail de ce stage avait permis d'adapter une approche d'inférence de réseaux de gènes développée au LIMOS à des données d'expression de gènes. La méthode avait été appliquée à des données de puces ADN obtenues au laboratoire et avait permis de reconstruire partiellement et d'étendre un schéma d'interactions de facteurs de transcription établi chez l'orge. Ces résultats exploratoires avaient encouragé le GDEC et le LIMOS à proposer un sujet de thèse à l'Ecole Doctorale des Sciences de la Vie de Clermont-Ferrand. Le sujet proposé et les travaux initiés visent à explorer les réseaux de régulation et les bases moléculaires impliquées dans le contrôle de l'accumulation des protéines de réserve dans le grain de blé tendre (*Triticum aestivum* L.).

Ce manuscrit est présenté sous la forme d'une première partie très générale présentant les principales caractéristiques et l'histoire de la culture du blé, d'une synthèse et d'un recueil de quatre des publications en cours de préparation, soumises et/ou publiées au cours de la thèse suivis d'une discussion. La première partie permettra aux lecteurs peu familiers avec le blé de découvrir son développement et son histoire. La synthèse se compose de trois grandes parties, **(I)** une présentation méthodologique du processus d'annotation fonctionnelle dans le contexte des séquences de blé, **(II)** une présentation méthodologique de l'inférence de réseaux dans le cadre de la biologie des systèmes et **(III)** une synthèse de la physiologie du grain dans le contexte des métabolismes de l'azote (N) et du soufre (S). La synthèse fait parfois directement mention aux différentes figures de ces publications. Ces figures sont appelées par le numéro de la publication auxquelles elles font référence (I, II, III, IV selon leur ordre de présentation) et par leur numérotation respective dans chacune des publications (e.g. I-1, I-

2, II-S1 etc.). Les figures spécifiques à la synthèse sont nommées par une simple numérotation. Les transitions entre les articles synthétisent les principaux résultats obtenus dans les différentes publications. Un cinquième article de recherche auquel j'ai participé mais qui sort néanmoins du cadre de cette thèse est présenté en annexe.

PUBLICATIONS DANS DES REVUES SCIENTIFIQUES AVEC COMITE DE LECTURE

Vincent,J., Dai,Z-W., Ravel,C., Choulet,F., Mouzeyar,S., Bouzidi,M-F., Pailloux,M. and Martre,P. (2013) dbWFA: a web-based database for functional annotation of Triticum aestivum transcripts. Database, 2013 May 9; 2013:bat014. doi: 10.1093/database/bat014

Vincent,J., Martre,P., Gouriou,B., Ravel,C., Dai,Z-W., Petit,J-M. and Pailloux,M. **(2014)** RulNet: A web-oriented platform for regulatory network inference, application to wheat –omics data. *PLoS one*. **[En révision]**

Dai,Z-W., Plessis,A., Vincent,J., Duchateau,N., Besson,A., Dardevet,M., Duyen,P., Gibon,Y., Hilbert,G., Pailloux,M., Ravel,C. and Martre,P**. (2014)** Nitrogen to sulfur supply ratio modulates the composition and accumulation kinetics of grain storage protein in association with metabolism rebalancing and transcriptome reprogramming for wheat. *The Plant Cell.* **[Soumis]**

Vincent,J., Martre,P., Besson,A., Dai,Z-W., Romeuf,I., Rustenholz,C., Paux,E., Dardevet,M., Pailloux,M. and Ravel,C. **(2014)** Transcriptome analysis of wheat transcription factors reveals key regulators of grain development and their response to nitrogen and sulfur. *The New Phytologist*. **[En préparation]**

Ravel,C., Fiquet,S., Boudet.J., Dardevet,M., Vincent.J., Merlino.M., Michard.R. and Martre,P. **(2014)** Characterization of non-coding cis-regulatory elements in the promoter region of high-molecular weight glutenin subunit genes for *Triticum aestivum* L. *Frontiers in Plant Science*. **[En révision]**

COMMUNICATIONS DANS DES SEMINAIRES OU DES CONGRES IMPORTANTS

Vincent,J., Dai,Z-W., Ravel,C., Choulet,F., Mouzeyar,S., Bouzidi,M-F., Agier,M. and Martre,P**. (2011)** dbWFA: a web-based database for functional annotation of wheat transcripts. *Journées Ouvertes en Biologie, Informatique et Mathématiques.* Du 28 juin au j^{er} Juillet 2011. Paris, FRANCE. [POSTER]

Vincent,J., Martre,P., Ravel,C., Baillif,A. and Agier,M. **(2011)** A web-oriented platform for gene regulatory network inference. Application to seed storage proteins in wheat. *Journées Ouvertes en Biologie, Informatique et Mathématiques.* Du 28 juin au 1er juillet 2011. Paris, FRANCE. [POSTER]

Vincent,J., Romeuf,I., Besson,A., Martre,P., Dai,Z-W., Agier,M. and Ravel,C**. (2013)** Unravelling transcriptional regulators of seed development and their response to nitrogen and sulfur availability in wheat. 7th European Plant Science Organization conference. *Plants for a Greening Economy*. Du 1er au 4 septembre 2013. Porto Heli, GREECE. [POSTER]

Vincent,J., Romeuf,I., Besson,A., Martre,P., Dai,Z-W., Agier,M. and Ravel,C. **(2013)** Mise en évidence des régulateurs transcriptionnels clé du développement du grain de blé et leur réponse à la disponibilité en azote et en soufre. *4ème colloque national du réseau français de biologie des graines*. Du 30 au 31 octobre 2013. Dijon, FRANCE. [ORAL]

REMERCIEMENTS

Le mémoire de thèse représente l'aboutissement de trois années (ou un peu plus) de travail collectif, de sacrifices mais aussi d'échanges et de partage. L'exercice des remerciements s'impose donc naturellement dans le travail de rédaction et aussi peu scientifique soit-il, je ne le négligerai pas pour autant. Ces remerciements aussi succincts soient-ils, permettent malgré le caractère nominatif du manuscrit, de rendre le mérite à toutes les personnes ayant contribué, d'une façon professionnelle et/ou plus personnelle à l'achèvement de ce travail.

Je me dois de remercier le Ministère de l'Enseignement Supérieur et de la Recherche pour avoir subventionné 3 années de recherche ainsi que le département Biologie et Amélioration des Plantes de l'INRA pour les quelques mois supplémentaires.

Je me dois également de remercier le directeur du LIMOS, Alain Brelurut, et les directeurs successifs de l'UMR GDEC, dans l'ordre Gilles Charmet et Thierry Langin. Je remercie tout particulièrement Gilles qui m'a suivi depuis mes premiers pas au GDEC. Suite à mon stage de Master tu as suivi Pierre et tu as soutenu ce projet de thèse, tu étais là lorsque j'ai présenté mes travaux afin de « glaner » un financement lors du concours pour les bourses ministérielles. Tu as temporairement prêté ton nom, en tant que directeur de thèse, jusqu'à la soutenance de l'HDR de Pierre et je te remercie pour cela et pour ta gentillesse.

Je suis sensible à l'honneur que me font les membres du Jury, Julia Buitink, Christophe Tatout, Christophe Rothan, et Bertrand Dubreucq de juger le travail effectué au cours de ces trois années.

Je tiens aussi à remercier infiniment les personnes à l'origine de mon travail et qui en leur qualité de directeur de thèse et encadrants ont su me supporter. Pierre tout d'abord, tu m'as guidé tout au long de cette aventure, je retiens ton perfectionnisme, ton excellence scientifique, tu m'as appris à toujours tirer le maximum des données que nous avions, à voir plus loin que l'évidence dans les résultats que nous obtenions. J'espère que ces qualités déteindront un peu sur moi bien que la route soit encore très longue. J'ai toujours aimé travailler avec la plus grande autonomie mais j'ai apprécié que tu sois à mon écoute quand j'en avais besoin. Marie ensuite, ton dynamisme, ton aide et ton soutien m'ont été indispensables. Je te souhaite le meilleur dans ta vie de famille et dans ta carrière professionnelle. Je sais que ton voyage a été enrichissant et j'espère que tu auras l'occasion de renouveler cette expérience ! Catherine aussi tu as passé avec succès ton HDR l'an dernier et je suis certain que tu seras une excellente directrice de thèse pour les prochains étudiants qui prendront ma place. Je te remercie pour ton aide précieuse et pour ton immense

gentillesse. Bien que nous ayons parfois pris du retard, et je prends ma part de responsabilités la dedans, j'espère que nous sommes un exemple de collaboration réussie entre des domaines de recherche bien différents.

Je remercie Bruno Moulia, Jean-Marc Petit, Florence D'Alche-Buc et une nouvelle fois Bertrand Dubreucq pour les conseils éclairés qu'ils m'ont prodigués lors des deux comités de suivi de thèse.

Je remercie tous les collègues de l'UMR GDEC pour les échanges, les différents moments de vie que nous avons partagés, les discussions scientifiques, je pense notamment à Fred Choulet et Etienne Paux. J'ai aussi une reconnaissance sincère pour le travail de Valérie, Patricia et surtout Karine qui ont su me guider dans toutes les démarches administratives.

Je tiens à remercier toute l'équipe BIG, équipe à laquelle je me suis greffé pendant un peu plus de trois années. La qualité de leur travail ainsi que l'émulation qui règne même lorsque les réunions d'équipe s'éternisent sont admirables! J'avoue aujourd'hui, mais vous l'aurez bien remarqué, n'avoir pas toujours été le plus attentif aux problèmes « de paillasse » dont nous avons souvent discuté en réunion. J'espère que vous ne m'en tiendrez pas rigueur.

Je remercie aussi tout particulièrement Anne Plessis et Zhanwu Dai, anciens post-docs de l'équipe BIG qui ont collaboré avec moi sur la plupart des travaux. Pour leur gentillesse aussi !

Je remercie aussi tous les joueurs de l'équipe de volley de l'INRA de Clermont-Ferrand qui m'ont accepté parmi eux, cette petite famille qui permet à tous les amateurs du volley de se retrouver une fois par semaine, cela m'a souvent permis de me défouler et dieu sait qu'il y en avait besoin.

J'ai une pensée particulière pour tous les membres de l'équipe ABC avec qui j'ai partagé des cafés (trop ?), mais surtout de très nombreux moments de détente (jamais trop) si importants pour garder les idées claires. Nicole, Joëlle, David, Vincent, Jacques, Christine et tous les jeunes qui sont passés par là. Mon autre équipe d'adoption au GDEC, j'ai moins souvent traversé l'Artière ces derniers temps mais j'espère que vous aurez compris que j'étais très occupé. J'ai une pensée particulière pour Bernard…

Je remercie chaleureusement Sibille et David mes collègues de bureau, qui ont su me supporter et m'apporter au quotidien toute leur bonne humeur et partager avec moi leur humour qui a souvent rattrapé le mien.

J’ai aussi une pensé pour les jeunes copains du GDEC, Josquin, Delphine et Natasha pour ne pas les citer. Ça fait toujours du bien de discuter autour d’un café, d’un ballon de volley, d’un duel quizz ou même d’un verre d’eau.

Passons enfin aux incontournables, ceux qui ont tout partagé avec moi, ceux qui se sont investis autant que moi dans cette aventure. Remi, pendant trois ans tu as su tenir tous les rôles que peut avoir un ami, une oreille à mon écoute, un soutien sans faille, un troubadour ou même un joueur de banjo quand il le fallait. Tu as même su me faire croire que j’étais un golfeur talentueux ! J’espère que comme moi tu réalises à quel point les pauses café ensemble vont nous manquer, elles manqueront un peu moins à ta carte nespresso. Après ces 3 années de CDD tu as gagné ta titularisation. Je n’oublie pas une bise pour Marion et Loulou. Et comme tu dis si bien, bientôt l’Espagne !

Enfin Anne Marie ma moitié, plus que tout autre, je t’exprime mon plus grand amour, pour tant de choses, en ce qui concerne cette thèse ? Ton soutien quotidien, dans les bons moments comme dans les moments de doutes. Tu as su me supporter, me soutenir, m’apporter tout ton amour qui a été un vrai moteur pour moi, il m’a permis de tout donner, sans rien lâcher… Ta présence, pour faire simple, a fait de moi un étudiant de thèse comblé, et grâce à toi je suis l’une des rares personnes qui gardera un bon souvenir de cette fameuse troisième année fatidique. Je n’oublie pas tous les beaux moments que l’on a passé, et je me délecte déjà de nos futurs voyages. Il n’y a pas besoin de s’éterniser, je ne te remercierai de toute façon jamais assez, simplement, cette aventure de trois années, nous l’avons partagée ensemble, cette réussite, c’est la tienne.

Pour en finir avec les personnes les plus chères à mon cœur, je dédie ce travail à mes parents, Bachir et Sophie et mes frères et sœurs, Karim, Sarah, Tom, Mila ainsi qu’à Pépé. De simples mots sur une page blanche ne suffiront jamais pour exprimer tout l’amour que je vous porte, je ressens une immense joie de faire partie d’une famille métissée si formidable. Qui aurait pu penser que nous en serions là quand je vous cachais mes bulletins de notes ? Vous m’avez toujours rassuré, sachez simplement que je ne serais rien du tout sans vous.

Merci.

LISTE DES MATIERES

Avant-propos ____ *3*

Publications dans des revues scientifiques avec comité de lecture ____ **5**

Communications dans des séminaires ou des congrès importants ____ **5**

Remerciements ____ *6*

Liste des Matières ____ *1*

Table des Figures ____ *4*

Introduction générale ____ *13*

Origine, culture et utilisation des blés ____ *17*

Développement d'une plante de blé ____ **18**

Structure d'une plante de blé à floraison ____ **19**

Les phases clés du développement du grain de blé ____ **21**

Aspects paléogénomiques ____ **22**

Domestication du blé ____ **24**

Utilisation des blés ____ **25**

Amélioration moderne des variétés et des techniques de culture ____ **26**

La qualité dans un contexte de diminution des intrants ____ **26**

Synthèse ____ *29*

Annotation fonctionnelle ____ **30**

Problématique ____ 32
A l'origine, le système E.C. ____ 33
Le système Gene Ontology. ____ 33
La collection des voies métaboliques PlantCyc ____ 35
Classification fonctionnelle du MIPS ____ 35
MapMan ____ 37
Comparaison de l'annotation d'une phytoène synthétase dans les différents systèmes ____ 37
L'outil d'annotation développé au cours de la thèse : dbWFA ____ 38

Inférence de réseaux et biologie des systèmes ____ **39**

Contexte ____ 39
Biologie des systèmes ____ 39
Les méthodes d'inférence ____ 44
Les approches informatiques ____ 46
L'approche d'inférence de réseaux développée : RulNet ____ 50

Cycle de l'azote et du soufre dans le sol ____ **53**

Cycle de l'azote ____ 53
Cycle du soufre ____ 54
Assimilation et métabolismes de l'azote et du soufre dans la plante ____ 56
Interaction entre l'azote et le soufre dans la plante ____ 59
Remobilisation de l'azote et du soufre vers le grain ____ 60

Accumulation des protéines de réserve et nutrition azotée et soufrée 63
Développement du grain et accumulation des PR chez le blé 63
Azote, soufre et composition en protéines de réserve 65
Contrôle transcriptionnel de l'expression des gènes des protéines de réserve 68
Autres niveaux de régulation de la synthèse des protéines de réserve 72
Conclusion de la synthèse et stratégie d'étude 75
Références 79
Publication 1 ***99***
Conclusion Publication 1 ***112***
Publication 2 ***115***
ABSTRACT 116
INTRODUCTION 117
METHODS 119
Multipredicate rules and RQL language 119
Rule generation and interest measures 121
Characteristics and features 124
Computational performances 124
RESULTS 126
Topological comparison of correlation-based and RulNet regulatory networks 127
133
Identification of transcription factors involved in the transcriptional reprogramming occurring during grain development and in response to N and S supply 130
DISCUSSION 136
REFERENCES 138
Supplementary information 143
Conclusion Publication 2 ***152***
Références 153
Publication 3 ***155***
Abstract 156
Introduction 157
Materials and Methods 159
Plant growth conditions and experiments setup 159
Microarray hybridization, image analysis and data availability 162
Data analysis 163
Results 165
Organs transcriptional signature and global TFs expression 165
Roots, leaves, stem, ears, and grain have a strong transcriptional signature 168
Development has a stronger transcriptional signature than nitrogen and sulfur supply for grains 170
Differentially expressed TFs in the developing endosperm 172
174
Revealing TFs involved wheat grain response to N and S supply through a network inference approach 174
Discussion 177

References 180

Conclusion Publication 3 ***191***

Réferences 192

Publication 4 ***193***

ABSTRACT 194

INTRODUCTION 195

Materials and methods 198

Plant Material and Growth Conditions 198

Statistical Analysis 201

RESULTS 203

Nitrogen and Sulfur Supply Causes Profound Modifications in Mature Grains 203

The Effect of Nitrogen and Sulfur Supply on Grain Protein Composition Can Be Broken Down into Modifications of Rate and Duration of Accumulation 205

Sulfur-rich and Sulfur-poor Grain Storage Proteins Follow Distinct Allocation Relationships 208

S-poor Grain Storage Proteins are the Most Transcriptionally Regulated Grain Storage Proteins 209

Nitrogen and Sulfur Availabilities Affect Free Amino Acids and their Precursors 209

Sulfur Deficiency Induces Transcriptome Reprogramming in Metabolic and Transport Genes and Regulates Non-Prolamin and Grain Storage Protein Genes 212

Regulatory Network Analysis Uncovers Novel Protein-Metabolite-Gene Linkages 213

DISCUSSION 216

Allometric Allocation of Nitrogen and Sulfur to Grain Storage Proteins is Mainly Dependent on Transcriptional Regulation 216

Adjustments of Amino Acid Transport and Metabolism Are Necessary to Regulate Grain Storage Protein Accumulation in Response to Sulfur Deficiency 217

Sulfur Supplied at Mid-Grain Filling Can Compensates for Grain Sulfur Deficiency 219

REFERENCES 221

Supplementary Information 228

Conclusion publication 4 ***245***

References 246

Discussion ***249***

Conclusion 250

Perspectives 253

Références 253

Annexe : Publication 5 ***257***

259

Résumé ***276***

Abstract ***276***

TABLE DES FIGURES

Figure 1 – Développement d'une plante de blé représentée sur l'échelle de Zadoks (Large et al., 1974). 19
Figure 2 - Représentation d'une talle de blé après la floraison. 20
Figure 3 – Schéma de la structure d'un épi et d'un épillet de blé. 21
Figure 4 – Variations des masses (traits pleins) de matière fraîche (vert), matière sèche (noir) d'eau (bleu) et protéines (rouge) et teneurs (lignes pointillées) en eau (bleu) et protéines (rouge) du grain de blé. 22
Figure 5 - Une représentation graphique du système Gene Ontology.. 34
Figure 6 - Représentation de réseaux. 41
Figure 7 - Classification des méthodes d'inférence de réseaux 44
Figure 8 – Représentation (a) et distribution (b) des degrés de connectivité des nœuds pour trois modèles de réseaux. (A) Modèles de graphe aléatoire proposé par Erdos-Renyi. (B) Réseau scale-free proposé par Barabasi et Albert. (C) Réseau hiérarchique et scale-free. 48
Figure 9 – Origine de la topologie scale-free dans les réseaux biologiques. 49
Figure 10 – Schéma du cycle de l'N dans les écosystèmes terrestres et aquatiques. 53
Figure 11 – Schéma du cycle du S dans les écosystèmes terrestres et aquatiques. 55
Figure 12 – Représentation schématique des enzymes clés impliquées dans l'assimilation de l'azote dans les jeunes feuilles. 57
Figure 13 – Schéma du métabolisme du S chez les plantes. 58
Figure 14 – Schéma du transport du sulfate dans la plante. 59
Figure 15 - Quantité de protéines métaboliques-structurales (a), gliadines (b), et gluténines (c) en fonction de la quantité totale d'N par grain. 67
Figure 16 – Relations entre les différentes familles de FT identifiées chez A. thaliana. 70
Figure 17 – Réseau hypothétique impliqué dans la régulation de la synthèse des protéines de réserve du grain chez les céréales. 76

INTRODUCTION GENERALE

Au cours de la seconde moitié du XX^{ème}e siècle, la production agricole a été plus que doublée à l'échelle mondiale, essentiellement en raison de l'augmentation des rendements des grandes cultures comme le riz (*Oryza sativa* L.), le maïs (*Zea mays* L.) et le blé (*Triticum aestivum* L.) dans les pays industrialisés (Aceituno *et al.*, 2008). En France plus précisément, et sur cette même période, les rendements moyens observés sur les cultures de blé sont passés de 2 tonnes par hectare à plus de 7 tonnes par hectare (Brancourt-Hulmel *et al.*, 2003) Les trois espèces de blé les plus cultivées sont le blé tendre (*T. aestivum* L.), le blé dur (*T. turgidum* L. ssp. *Durum*) et l'engrain (*T. monococcum* L.). Le rendement en grains est depuis longtemps le principal caractère de sélection visé par les programmes d'amélioration variétale. On considère aujourd'hui que le progrès génétique d'une part, la mécanisation, les engrais et les traitements phytosanitaires d'autre part sont responsables chacun pour moitié dans l'augmentation des rendements observée depuis plus d'un demi-siècle. Au cours de cette période, le doublement de la production alimentaire mondiale s'est accompagné d'une importante augmentation de la nutrition azotée (Cassman *et al.*, 2003; Hirel *et al.*, 2011). Cependant, la surutilisation de fertilisants azotés provoque un déséquilibre environnemental qui est néfaste pour la santé des écosystèmes (Hirel *et al.*, 2007). D'une part, cette fertilisation, peut entrainer des émissions gazeuses d'ammoniac et de gaz azotés qui ont des impacts très important sur la santé de par leurs effets polluants, irritants voire toxiques. Le protoxyde d'azote (N_2O) par exemple est un gaz à effet de serre au pouvoir radiatif 310 fois plus élevé que celui du CO_2. En outre, le surplus de fertilisants azotés qui n'est pas fixé par les cultures ou par les micro-organismes des sols peut être lessivé, entrainant ainsi une eutrophisation des systèmes aquatiques environnants (Giles, 2005; Michael Beman *et al.*, 2005).

Ainsi, face aux changements climatiques et en interaction avec les nouvelles contraintes agricoles, notamment la limitation des intrants, un phénomène de stagnation des rendements est actuellement observé (Graybosch & Peterson, 2010; Brisson *et al.*, 2010) L'application avec discernement et parcimonie d'engrais minéraux est un enjeu majeur dans les systèmes de culture modernes, notamment à l'échelle des états parfois soumis à des politiques agricoles strictes quant à la quantité d'engrais pouvant être apportée à un sol.

Plus récemment, l'amélioration variétale a également eu pour objectif d'optimiser la valeur boulangère des grains, étroitement liée à la teneur en protéines. Bien qu'impactée par la variabilité génétique, et différents facteurs environnementaux, la relation entre le rendement

en grain et la teneur en protéines du blé demeure immuablement négative et profite à des composés issus du métabolisme du carbone (C), amidon ou huile (Triboi et Triboi blondel, 2002). Cette corrélation négative entre les quantités d'N et de C des grains repose sur un déterminisme d'ordre génétique (Oury & Godin, 2007; Bogard *et al.*, 2011). Malgré cet antagonisme depuis longtemps connu des scientifiques, le processus d'optimisation des rendements s'est effectué au détriment de la teneur protéique du grain de blé, indicateur majeur de sa qualité nutritionnelle et technologique.

Les échanges commerciaux de blé se pratiquent aujourd'hui à l'échelle mondiale, obligeant les producteurs à maintenir des critères de qualité et de rendement stricts en dépit des nouvelles contraintes. La diversité des utilisations du grain découle des propriétés rhéologiques uniques des protéines de réserve (PR) du grain qui mélangées à l'eau forme un réseau viscoélastique : le gluten. Ce sont les propriétés de ce réseau qui sont exploitées dans la plupart des utilisations industrielles du gluten. Les propriétés technologiques du gluten sont déterminées d'une part par le type des PR (caractéristique des variétés ; Branlard *et al.*, 2001) leur composition et leur état d'agrégation (déterminés principalement par les conditions agro-climatiques), d'autre part. Le maintien d'une valeur d'utilisation optimale du grain de blé sans impacter le rendement en grains constitue ainsi un enjeu essentiel à l'échelle mondiale, notamment en France qui en est le 5ème producteur mondial et 1er producteur européen. Le progrès génétique que peut apporter la sélection variétale se place donc comme un levier majeur pour les scientifiques. Il s'agit d'identifier les bases moléculaires sous-jacentes à la mise en place de la teneur en PR du grain, et ce dans un contexte de limitation des intrants. L'effet de différents facteurs environnementaux tels que l'N, le S ou la température et l'eau sur la qualité des grains matures est plutôt bien caractérisé, de même que pour les interactions génotype x environnement (Graybosch *et al.*, 1996). Néanmoins les bases moléculaires et physiologiques sous-jacentes à la synthèse des PR restent encore largement inconnues.

Afin de comprendre le comportement d'unités biologiques, il est profitable de conceptualiser celles-ci comme des systèmes régis par des interactions entre éléments biologiques. Dans cette optique, il est nécessaire (1) d'identifier les éléments impliqués dans le comportement du système et (2) de comprendre le sens des interactions qui les lient. La finalité d'une approche systémique en biologie est donc de synthétiser le comportement du système biologique étudié et des interactions qui le composent. Les méthodes statistiques et computationnelles utilisées dans le cadre d'une approche de « biologie des systèmes »

permettent d'élucider la structure des réseaux de régulations moléculaires sous-jacents aux processus biologiques étudiés.

Dans ce contexte, la thèse avait pour objectif d'étendre une approche d'inférence de réseaux basée sur la découverte de règles développée en 2007 (Agier *et al.*) dans l'optique de fournir un outil fini permettant d'inférer des réseaux de régulations à partir de données multi-échelles. L'extension de l'approche permet en outre, de diversifier les sémantiques des règles, et ainsi, d'étendre leur recherche à de nouvelles significations biologiques. La finalité consistait à déterminer les acteurs potentiels et les interactions moléculaires impliquées dans la synthèse des PR dans le grain de blé tendre. Une stratégie d'étude combinant des jeux de données transcriptomiques, protéomiques, et métabolomiques a été mise en place afin d'explorer les réseaux impliqués, à différents niveaux, dans la régulation de la synthèse des PR et d'observer le comportement du système en fonction des approvisionnements en N et en S. Le travail de thèse est appliqué au blé tendre (*Triticum aestivum* L.) et s'appuie sur des connaissances accumulées tant au niveau physiologique que génétique.

ORIGINE, CULTURE ET UTILISATION DES BLES

Le blé est la 3ème culture mondiale de par l'importance de sa récolte (695 Millions de tonnes en 2012), et avec le riz et le maïs, la céréale la plus consommée par l'homme (Organisation des Nations Unies pour l'alimentation et l'agriculture). Les plus grands pays producteurs de blé sont dans l'ordre la Chine, l'Inde, les Etats-Unis, la Russie et la France. La France en est donc le 1er producteur de l'Union Européenne. Actuellement environ 50% de la production française de blé est destinée à l'exportation, il joue donc un rôle majeur dans l'équilibre de notre balance commerciale. Plus de 90% de la production de blé dans le monde est destinée à l'alimentation, humaine et animale. On observe des variations importantes en matière de rendement (quintaux de blé produits par hectare de culture) entre les pays, traduisant des contraintes climatiques et des stratégies de culture agricole très différentes. L'Europe prône une agriculture intensive du fait d'une surface de culture réduite. Cette culture est synonyme de hauts rendements tandis qu'aux Etats-Unis par exemple, la culture du blé est dite extensive et produit des rendements bien plus faibles. Les échanges commerciaux de blé se pratiquent à l'échelle mondiale et l'importance de certains pays parfois soumis à des intempéries ou des changements de politique d'échange se traduit par une volatilité importante du prix d'échange. En 2010 par exemple, année de sécheresse en Russie (4ème pays exportateur) et en Ukraine (7ème pays exportateur), la spéculation sur la récolte annuelle a entrainé une augmentation des prix de l'ordre de 70%.

DEVELOPPEMENT D'UNE PLANTE DE BLE

La croissance du blé est généralement décrite par les systèmes de Zadoks (Zadoks *et al.,* 1974 ; Figure 1) ou de Feekes (Large, 1954 ; non présenté). Les deux systèmes sont utilisés dans des contextes différents, parfois conjointement ; l'échelle de Zadoks possède une résolution supérieure pouvant être avantageuse pour les scientifiques. Le développement et l'extension du blé (tallage, élongation de la tige, épiaison) s'accompagnent d'une augmentation de sa masse qui se prolonge même après la fin de l'extension de la plante. Une fois la taille finale de la plante acquise, les constituants qui la composent sont en équilibre entre import et export de la plante. C'est à ce moment que la plante peut remobiliser des constituants des organes plus anciens vers les entités végétatives en croissance. Cette étape appelée senescence aboutit à la mort des tissus. A maturité, la sénescence foliaire et la remobilisation de constituants vers les grains aboutissent à la mort quasi-complète des structures végétatives au moment de la récolte.

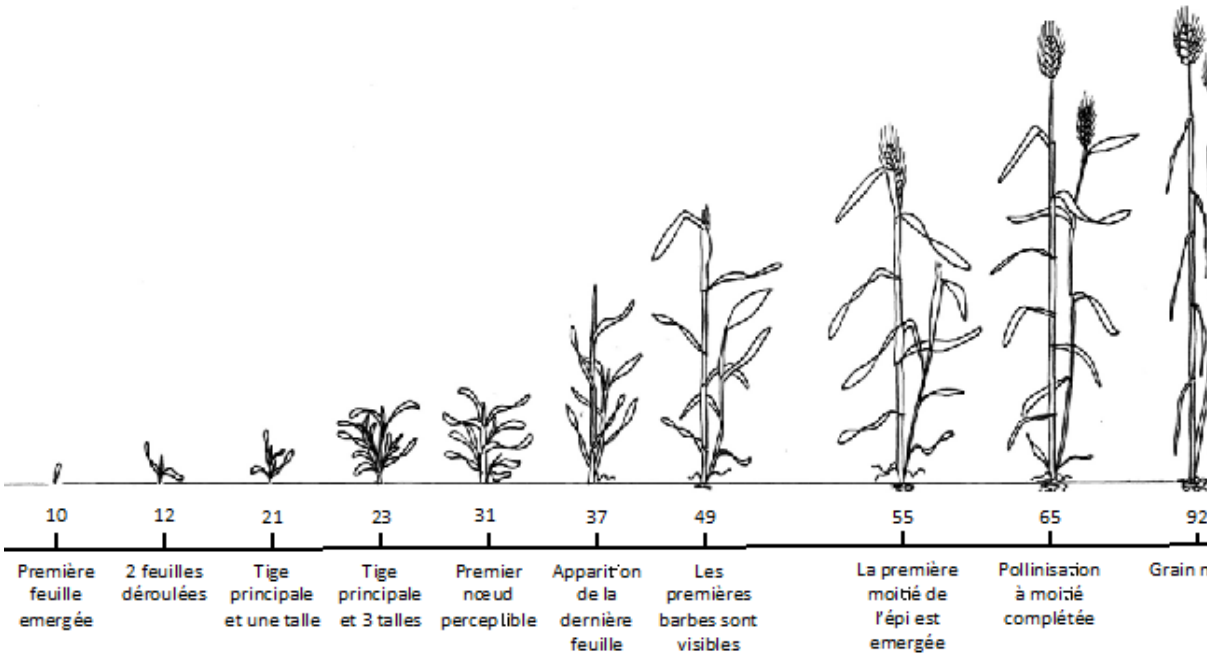


Figure 1 – Développement d'une plante de blé représentée sur l'échelle de Zadoks (Large et al., 1974).

STRUCTURE D'UNE PLANTE DE BLE A FLORAISON

Morphologiquement, une plante de blé à floraison est l'aboutissement de tous les processus de développement et de croissance qui surviennent lors de son cycle de vie. Toutes les structures telles que les feuilles, les talles et les épis ont atteint leur taille maximale. La plante de blé à floraison est constituée d'un ensemble de brins (aussi appelés talles) comportant chacun une tige feuillée (partie végétative), un épi qui dérive de la tige (partie reproductrice) et un système racinaire. Les brins sont tous issus de la ramification du brin maître, déjà présent dans la graine à l'état embryonnaire.

Le système racinaire du blé est de type fasciculé, avec quelques racines primaires d'égale importance et des racines secondaires qui en dérivent. La structure végétative de la plante est constituée d'une succession d'unités morphologiques similaires, les phytomères (Figure 2A), dont le nombre peut varier assez largement, de 6 à 16 le plus communément. Chaque phytomère est composé d'un nœud, d'un entre-nœud et d'une feuille (elle-même composée d'une gaine et d'un limbe séparées par la ligule ; Figure 2B). Chaque tige est surmontée à

floraison par le pédoncule de l'épi et l'épi qui apparaissent lorsque le développement de la tige est terminé. Les phytomères situés à la base de la plante sont assez courts et les nœuds sont très difficiles à distinguer. Les entre-nœuds de la tige sont pour la quasi-totalité entièrement entourés de gaines foliaires, seul l'entre nœud le plus proche de l'épi est partiellement visible. Des racines dites nodales se développent à partir des nœuds de la base de la plante, elles ne se développent pas à partir des autres nœuds plus aériens puisque leur croissance est photo-inhibée. Les entre-nœuds sont plus allongés à mesure que l'on s'approche de l'épi.

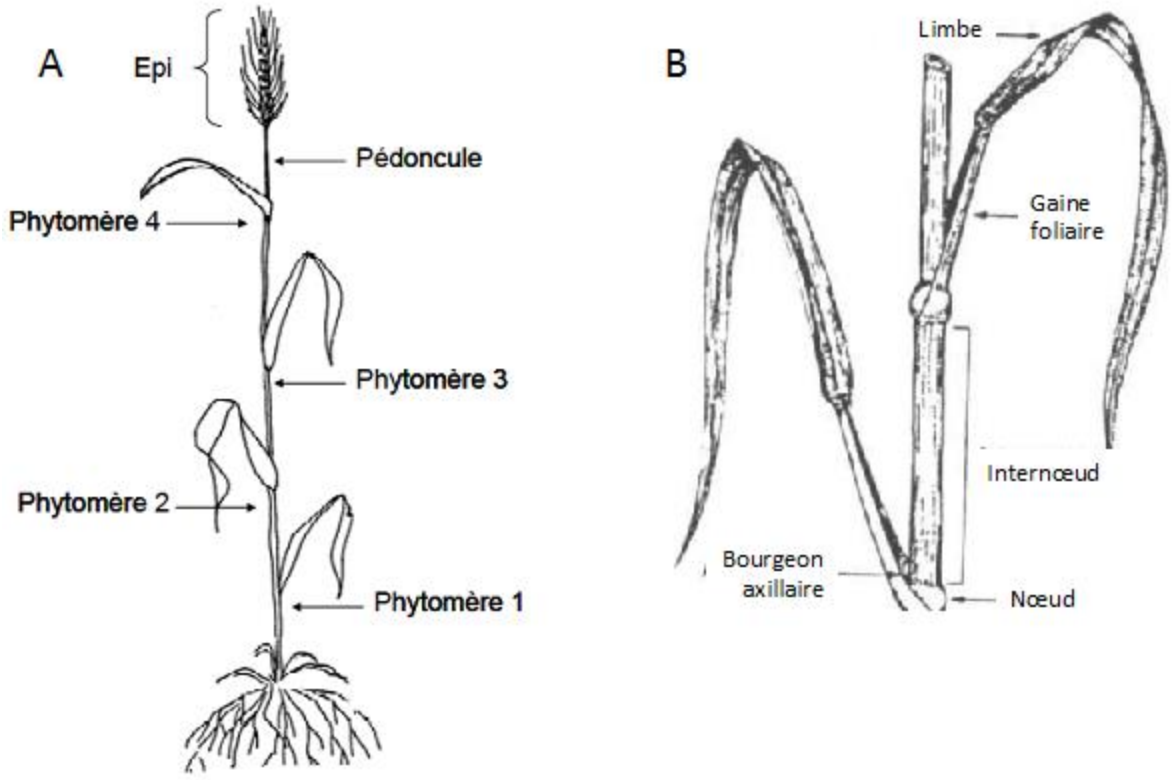


Figure 2 - Représentation d'une talle de blé après la floraison. *Une talle est constituée d'un assemblage d'unités morphologiques similaires, les phytomères, surmontés du pédoncule de l'épi et de l'épi (A). Chaque phytomère est potentiellement constitué, du bas vers le haut, d'un nœud portant un bourgeon axillaire et des racines nodales, d'un entre-nœud, d'une gaine foliaire et d'un limbe foliaire (B). Les phytomères inférieurs n'ont pas d'entre-nœuds allongés et les nœuds exposés à la lumière ne portent pas de racines. D'après Moore & Moser (1995).*

L'épi (Figure 3A) est l'inflorescence du blé, il est composé d'une succession d'épillets attachés à un rachis commun et séparés par des entre-nœuds. Chaque épillet (Figure 3B) est constitué de deux glumes qui renferment plusieurs fleurs sans pétales, chacune constituée de deux glumelles renfermant un grain après fécondation (généralement 3 ou 4 par épillet). Le grain

de blé (aussi appelé caryopse) est à la fois un fruit sec et une graine, sa face dorsale par rapport à l'épillet est arrondie tandis que le côté ventral comporte un sillon profond. Il est constitué du germe (2 à 3% du poids du grain) situé à la base du grain, au niveau de sa fixation sur l'épillet et des enveloppes (13 à 16% du poids du grain) qui renferment l'amende farineuse ou albumen (tissu de réserve triploïde ; 82 à 84% du poids du grain).

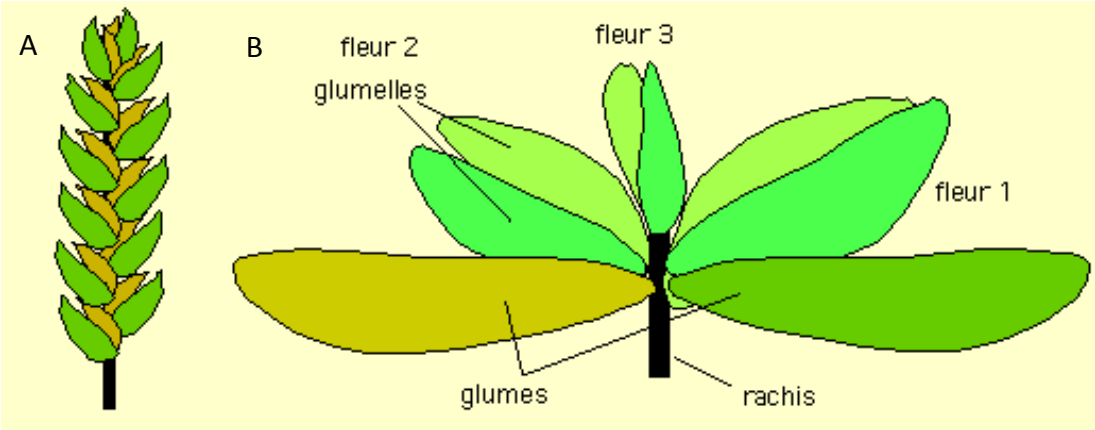


Figure 3 – Schéma de la structure d'un épi et d'un épillet de blé. *(A) Epi de blé composé d'un rachis central surmonté de plusieurs épillets séparés par des entre-nœuds. (B) Epillet composé de deux glumes renfermant les fleurs, glumelles et grains.*

LES PHASES CLES DU DEVELOPPEMENT DU GRAIN DE BLE

Le développement du grain de blé débute avec la formation de l'ovaire qui, après fécondation, va constituer les tissus externes (téguments) du grain d'origine maternelle. La double fécondation de l'ovule et de la cellule centrale donnent naissance respectivement à l'embryon et à l'albumen. Il s'en suit une phase dite syncitale (de 0 à 100 degré-jour[1] après floraison [°Cj], environ 0 à 5 jours après floraison [JAF] ; Figure 4) caractérisée par la multiplication rapide des noyaux de l'albumen. Cette étape est suivie par une phase de cellularisation de l'albumen (80-250°Cj, 5-14 JAF), caractérisée par une accumulation rapide d'eau. Au cours de ces deux premières phases le grain acquiert sa longueur finale. La troisième phase de développement du grain (250-750°Cj, 15-40 JAF) est caractérisée par

[1] La vitesse de développement des plantes dépend essentiellement de la température et est donc souvent exprimée en temps thermique (degré-jour, °Cj) calculé comme la somme des températures efficaces pour la croissance à partir d'une date de référence. La température efficace pour la croissance est égale à la température journalière moyenne moins la température de base (i.e. température à laquelle la croissance est égale à 0). La température de base pour le blé est de 0°C.

l'accumulation rapide de réserves carbonées (amidon) et azotées (PR). La dernière phase de développement (750-950°Cj, 40-50 JAF) correspond à l'acquisition de l'aptitude à la déshydratation du grain et à sa maturation.

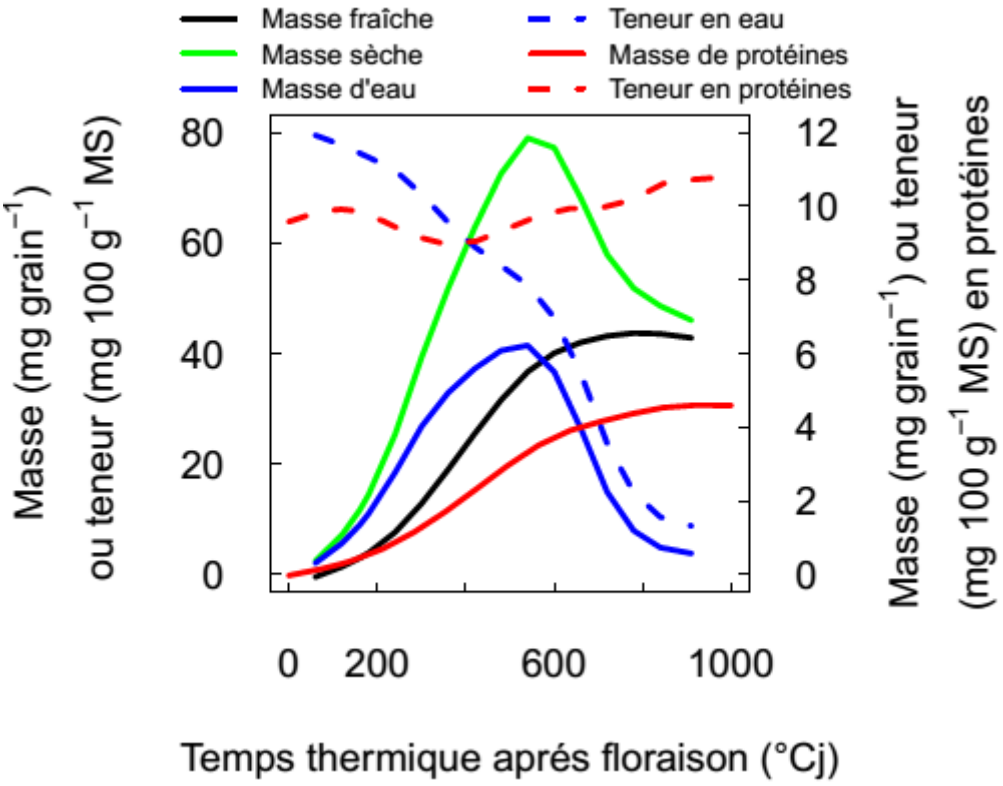


Figure 4 – Variations des masses (traits pleins) de matière fraîche (vert), matière sèche (noir) d'eau (bleu) et protéines (rouge) et teneurs (lignes pointillées) en eau (bleu) et protéines (rouge) du grain de blé. *MS : La matière sèche est ce que l'on obtient quand on retire l'eau du grain.*

ASPECTS PALEOGENOMIQUES

Le blé est une plante herbacée, monocotylédone du groupe des *Triticeae.* Le groupe des *Triticeae* est composé de cinq genres : *Aegilops*, *Elymus*, *Hordeum*, *Secale* and *Triticum.* Le blé est un terme générique qui regroupe plusieurs espèces du genre *Triticum* présentant une diversité morphologique et agro-écologique très importante. Le nombre de chromosomes chez les espèces du genre *Triticum* est x=7. Les espèces sauvages sont diploïdes (2n = 2x = 14) ou tétraploïdes (2n = 4x = 28). Enfin le blé tendre que l'on cultive massivement aujourd'hui est hexaploïde (2n = 6x = 42 ; Zohary & Hopf, 1993).

Les premières formes de blé cultivées étaient dipoïdes (*Triticum monoccocum*, constitution génomique AA) ou tetraploïdes (*Triticum dicoccum*, constitution génomique AABB). Les

analyses paléogénomiques permettent de retracer l'origine géographique de ces blés en Turquie (Heun *et al.*, 1997; Nesbitt & Samuel, 1998). Il apparaît que la polyploïdisation a joué un rôle essentiel dans l'évolution de la famille des Poacées (Salse *et al.*, 2008). Bien qu'il soit difficile de dater très précisément l'origine des évènements de polyploïdisation ainsi que les espèces impliquées, un consensus semble dégager une espèce étroitement liée à *Triticum urartu* (constitution génomique SS) et une seconde espèce fortement similaire à *Aegilops Speltoides* (constitution génomique BB) impliquées dans un premier événement de polyploïdisation naturelle remontant à près d'un demi-million d'années. Les blés dur (*Triticum turgidum* ssp *durum)* et amidonnier (*Triticum dicoccum*) semblent être des descendants de cette espèce tétraploïde fertile (Feldman, 2000). Les blés hexaploïdes tels que le blé tendre ou l'épeautre sont apparus il y a environ 10 000 ans lorsqu'une variété de blé tétraploïde domestiquée fut croisée avec *Triticum tauschii* (constitution génomique DD). La diversité des utilisations du blé tendre et son adaptabilité sont étonnantes malgré sa domestication récente et les événements de spéciation. Ces évènements défavorables à la diversité sont compensés chez le blé tendre par des évènements fréquents de translocation et de délétion de gènes qui génèrent de la diversité à une vitesse relativement élevée (Dubcovsky & Dvorak, 2007b). Le blé tendre tel qu'on le connaît aujourd'hui est caractérisé par la taille importante de son génome (17 000 Mb[2], soit environ 40 fois le génome du riz, organisme monocotylédone de référence). Plusieurs évènements successifs d'hybridation interspécifique sont à l'origine de la formation de son génome allohexaploïde. Ainsi chaque gène est en général présent en trois copies portées par chacun de ses trois génomes AA, BB et DD.

L'ancêtre du blé actuel était donc probablement difficile à cultiver, compliqué à récolter et tout aussi délicat à consommer. Néanmoins, la pression de la domestication, initiée il y a plus de 10 000 ans a permis de fixer des caractères avantageux notamment concernant la taille du grain, la facilité de battage et le maintien du grain sur l'épi et de l'épi sur la plante (Dubcovsky & Dvorak, 2007a). Des études récentes mettent en évidence l'importance de la plasticité du génome et de la ploïdie dans le succès de sa domestication jusqu'à la culture à grande échelle des variétés de blé modernes (Kareiva *et al.*, 2007).

[2] Millions de bases

DOMESTICATION DU BLE

La domestication du blé prend ses origines dans la zone géographique dite du croissant fertile située en Asie de l'Ouest (Smith, 1995). La domestication du blé fut avec celle de l'orge, un élément fondateur de la « révolution néolithique » voyant l'émergence des premiers fermiers. Cette période charnière de l'histoire vit une évolution drastique des coutumes alimentaires et bien que les blés de ce temps n'aient pas eu les propriétés de rendement et de qualité des blés modernes, une économie de production fondée notamment sur l'agriculture se mit en place progressivement. L'apparition de différents outils comme la faucille ou la meule en pierre permirent d'accélérer et d'étendre géographiquement l'adoption de cette économie de production. Cette période de domestication des céréales semble d'ailleurs coïncider avec la sédentarisation des sociétés alors nomades dans cette même zone géographique (Diamond, 2002). Les blés ont subi une première forme de sélection, naturelle, ainsi seuls les blés ayant résisté aux intempéries majeures ont pu prospérer. Une seconde forme de sélection a été appliquée. Les blés étaient essentiellement des variétés sélectionnées pour leurs caractéristiques avantageuses en condition de culture, les épis restant intacts au moment de la moisson, les grains étant plus gros ou les épillets étant plus solidaires de la tige lors de la croissance. Les fermiers de l'époque pratiquaient donc, bien qu'inconsciemment et de manière non scientifique, une premier forme de sélection variétale des blés, c'est ainsi qu'ont été stabilisés les premiers gènes dits de domestication. Cette domestication constitue la seconde grande phase de l'histoire de la culture des blés. La fixation ancienne d'allèles avantageux pour des gènes tels que *Tg*[3], *Q*[4] et *Br*[5] illustre parfaitement ce processus de sélection (Zohary & Hopf, 1988a). *Tg* est impliqué dans la ténacité des glumes. Ce sont des expansions foliaires dures et indigestes, ainsi lorsqu'elles adhèrent moins au grain, celui-ci se dénude plus aisément lors de la récolte, permettant de faire l'impasse sur une étape de décorticage. Le gène *q* quant à lui participe à l'apparition du caractère de non déhiscence de l'épillet, c'est à dire qu'un blé porteur d'un allèle *Q* que l'on considère avantageux aura un épi plus compact et non cassant. Ce caractère confère au blé une résistance de l'épi lors du battage. Enfin, de façon similaire au gène *q* un allèle avantageux pour le gène *Br* rend l'épi plus solide (Dubcovsky & Dvorak, 2007a).

[3] *Tg* : Tenacious glumes signifiant glumes adhérentes

[4] *Q* gene : free threshing signifiant grain libre au battage

[5] *Br* : Brittle rachis signifiant rachis cassant

Le processus de substitution des blés sauvages par les formes domestiquées s'est étalé sur environ un millénaire (Fuller, 2007). Le blé tendre n'a existé que sous une forme cultivée puisqu'il est le résultat de plusieurs évènements successifs de polyploïdisation. L'expansion de l'agriculture qui constitue la troisième phase de l'histoire de la culture des blés a ensuite suivi un axe est-ouest plus rapidement que nord-sud, les zones géographiques situées aux mêmes latitudes partageant des caractéristiques agro-climatiques identiques. Les échanges commerciaux vers l'Europe et l'Asie de l'Est permirent de propager la culture des blés (Balfourier *et al.*, 2007). Entre 7000 et 2000 avant JC elle fut adoptée de la France jusqu'à la Chine. Au cours de l'ère moderne, de nombreuses avancées technologiques et industrielles ont permis d'améliorer la récolte et la conservation des blés.

UTILISATION DES BLES

Les trois espèces de blé les plus cultivées sont le blé tendre (*T. aestivum* L.), le blé dur (*T. turgidum* L. subsp. *Durum)* et l'engrain (*T. monococum*). Le blé tendre, adapté à la culture en zones tempérées représente environ 95% de la production mondiale de blé, il est utilisé en l'état pour l'alimentation animal ou pour produire de la farine panifiable intervenant dans la confection du pain. Le blé dur quant à lui est plus adapté aux conditions agro-climatiques méditerranéennes et prend une part moins importante dans cette production, il est utilisé principalement dans la production de pâtes alimentaires ou de semoules. L'engrain enfin, est plus marginalement cultivé pour sa valeur nutritive supérieure sur certains aspects. L'importance de la production mondiale de blé tendre s'explique par sa grande plasticité génomique et phénotypique qui permet de le cultiver dans la plupart des régions agricoles mondiales, et par son fort potentiel de rendement (Shewry, 2009). Cette importance s'explique également par la multiplicité de ses utilisations alimentaires et industrielles

La tige des céréales, coupée lors de la moisson (paille) sert principalement de litière, à la base de la formation du fumier pour les animaux d'élevage. Le fumier peut également être utilisé dans la branche agricole comme fertilisant biologique ou dans l'alimentation animale en cas de pénurie de foin, on parle alors de blé fourrager. D'autres utilisations plus marginales existent e.g. combustible, matériaux de construction ou isolant thermique. Néanmoins, l'intérêt agronomique majeur du blé provient de son grain. La farine de blé, résultat de la mouture du grain est principalement constituée d'amidon (65–70%), d'eau (12–18%) et de protéines (8–15%). En outre, les polysaccharides non amylacés (2–3%), en particulier les arabinoxylanes et les lipides sont des constituants importants pour la qualité de la farine (Goesaert *et al.*, 2005). L'amidon du grain est un polysaccharide, composé de plusieurs

molécules de glucose. Il est utilisé essentiellement dans le domaine de l'agroalimentaire, à savoir l'industrie des boissons, des confiseries ou de la boulangerie. L'amidon est plus marginalement utilisé dans des domaines aussi variés que l'encapsulation de produits pharmaceutiques, la fabrication de cosmétiques, la papèterie, les biocarburants, les matières plastiques biodégradables… Les protéines du grain quant à elles, ont une importance primordiale, elles constituent la principale source de protéines dans l'alimentation humaine, on estime par ailleurs à plus de 20% la part représentée par le blé dans la consommation mondiale de calories (Zohary & Hopf, 1988b).

AMELIORATION MODERNE DES VARIETES ET DES TECHNIQUES DE CULTURE

Parmi les avancées les plus importantes du XX^{ème} siècle, les travaux de Norman Borlaug à l'origine de la « révolution verte » dans les années 1960-1970 ont eu un impact majeur en matière d'amélioration variétale. Ce sont ses travaux qui ont permis l'apparition de blés semi-nains, à haut rendement, insensibles à des maladies telles que la rouille du blé. Les premiers blés semi-nains ont été obtenus par croisement d'une variété américaine (Brevor 14) caractérisée par un haut rendement et une résistance à la maladie, avec une variété de blé naine japonaise (Norin 10). L'avantage du nanisme chez le blé s'explique par une propension moindre de la plante à plier ou se briser sous les contraintes physiques et par une allocation supérieure des différentes ressources au développement de la graine plutôt qu'à l'élongation de la tige. Il a ainsi permis de lutter contre la famine dans certaines régions du monde en étant à l'origine de variétés mieux adaptées à certaines contraintes agro-climatiques. Grâce et à cause de l'amélioration variétale qui s'applique toujours, les blés modernes présentent des caractères spécifiques au « syndrome de l'espèce cultivée ». L'évolution des engrais, d'abord organiques puis synthétisés a également permis d'augmenter considérablement les rendements des cultures agricoles mais les pays développés s'orientent désormais vers une limitation des intrants.

LA QUALITE DANS UN CONTEXTE DE DIMINUTION DES INTRANTS

L'homme s'intéresse à la nutrition des plantes depuis longtemps, Aristote au IV^{ème} siècle avant JC suggérait déjà qu'elles se nourrissaient de matières organiques qu'elles puisaient dans le sol. Au XVIII^{ème} siècle, une hypothèse plus précise est émise par Hassenfratz et Thaer, la « théorie de l'humus » selon laquelle les substances organiques nécessaires à la croissance d'une plante seraient extraites par cette dernière depuis le sol. Cette hypothèse est rejetée en 1840 par les travaux de Von Liebig qui établit le caractère purement minéral de l'alimentation des plantes. Plusieurs travaux, entre 1840 et 1870 permettent de mettre en

évidence l'absorption d'N notamment, par les plantes au niveau des racines. L'N est un élément indispensable à la croissance des plantes de culture (Sinclair & Horie, 1989), il constitue un facteur limitant de la croissance du blé lorsque sa disponibilité sous forme minérale dans le sol est insuffisante pour satisfaire les besoins de la plante. L'efficacité d'utilisation de l'N est un indicateur de la capacité de la plante à assimiler l'N disponible et à la traduire en rendement. Il existe deux composantes dans l'efficacité d'utilisation de l'N. Le RE (Recovery Efficiency) i.e. le rapport entre l'N assimilé et l'N apporté. L'IE (Internal Efficiency) qui correspond au rapport entre l'N assimilé et le rendement en grain. Ainsi, l'efficacité d'utilisation de l'N inclut différents processus physiologiques qui affectent l'utilisation de l'N tels que l'absorption de l'N du sol, son transport et sa distribution dans la plante (Moll *et al.*, 1982; Paponov *et al.*, 1996). Différents facteurs ont une influence sur l'efficacité d'utilisation de l'N, par exemple la rotation des espèces cultivées sur un même sol (Johnston *et al.*, 1994), le système de labour utilisé (Sieling *et al*., 1998) ou la disponibilité en eau (Campbell *et al.*, 1993). L'efficacité de l'apport d'N augmente avec l'augmentation du niveau d'N apporté (kg ha^{-1}) jusqu'à une valeur seuil au-dessus de laquelle l'efficacité commence à diminuer (Doyle & Holford, 1993). Ce niveau d'apport d'N optimal est dépendant de la variété de blé cultivée. Au-delà de cette valeur seuil, ce n'est plus l'N mais d'autres facteurs qui deviennent limitants. L'efficacité d'utilisation de l'N peut alors être rétablie grâce à un nouvel apport du nutriment limitant (Balasubramanian *et al.*, 2004).

Le S est également un élément essentiel dans la nutrition des plantes. Des conditions particulières telles qu'un faible taux de matière organique dans les sols, leur érosion et leur appauvrissement par les cultures successives peuvent se traduire par une carence en S dans les plantes. Des travaux ont montré que l'apport de S entraine une augmentation de la biomasse et du rendement en grain (Salvagiotti & Miralles, 2008), traduisant une interaction positive entre l'N et le S sur l'efficacité d'utilisation de l'N. Le métabolisme du S est lié au métabolisme de l'N dans de nombreuses fonctions physiologiques. Différentes études ont mis en évidence les effets de cette interaction N x S sur le rendement en grain chez le blé (Randall *et al.*, 1981; Reneau *et al.*, 1986). Il ne semble pas y avoir d'étude permettant de caractériser précisément les effets de l'apport de S, au niveau du RE ou de l'IE. Le S est un constituant essentiel d'enzymes impliquées dans le métabolisme de l'N telles que des nitrate réductases ou des nitrite réductases (Mendel, 1997; Campbell, 1999). La carence en S pourrait éventuellement expliquer un ralentissement du métabolisme de l'assimilation de l'N. Pour étayer cette hypothèse, des études ont montré l'accumulation de nitrates dans des plantes carencées en S (Zaroug & Munns, 1980). Des résultats suggèrent qu'il existe un synergisme

entre l'N et le S puisque l'apport de S permet d'augmenter le rendement dans le contexte d'une fertilisation azotée élevée, en revanche, le S apporté lors d'une fertilisation azotée faible n'apporte pas de changement significatif quant au rendement en grain (Salvagiotti *et al.*, 2009). En plus de l'effet de l'N et du S sur le rendement en grain, ces deux éléments ont surtout un impact sur la quantité et la composition du grain en PR.

SYNTHESE

ANNOTATION FONCTIONNELLE

L'élan principal qui a été insufflé au domaine de la bioinformatique provient sans doute de la croissance des quantités de données biologiques. Trouver les homologues d'une séquence inconnue est devenu une opération de routine dans le domaine de la biologie moléculaire. Avec l'augmentation exponentielle du nombre de gènes identifiés par les projets de génomiques, les biologistes sont confrontés à un océan de données à partir desquelles des connaissances nouvelles doivent être extraites. La question principale étant simple : que font ces gènes ? Il est techniquement impossible de mettre en place des expériences fonctionnelles pour chacun des gènes dont la fonction est inconnue. Compte-tenu de ces observations, les biologistes se tournent vers des méthodes bioinformatiques.

La nécessité d'utiliser un système de classification structuré utilisant un vocabulaire contrôlé ne se limite pas au spectre de l'annotation fonctionnelle des génomes mais permet aussi de fournir des outils efficaces pour l'analyse de données transcriptomiques et protéomiques qui sont générées à grande échelle. Les nouvelles techniques d'acquisition de données transcriptomiques permettent désormais de surveiller le niveau d'expression de chaque gène du génome dans de nombreuses conditions expérimentales. Alors que les protéines codées par un génome représentent conjointement les capacités physiologiques d'un organisme, les niveaux d'expression des gènes sous certaines conditions physiologiques décrivent la réaction physiologique de l'organisme aux conditions spécifiques étudiées. La connaissance des niveaux de transcription des gènes au cours du développement et dans certaines conditions ajoute une dimension dynamique à l'information génomique intrinsèquement statique. Comme les expériences de puces à ADN génèrent de grandes quantités de données à analyser et interpréter, les produits de gènes (protéines) doivent être classés en groupes fonctionnels pour être en mesure de découvrir les interactions fonctionnelles entre les gènes impliqués dans une réponse cellulaire.

L'annotation fonctionnelle de séquences codantes demeure un facteur limitant dans l'analyse des données transcriptomiques à haut débit. Ce problème est d'autant plus récurrent chez les espèces comme le blé qui possèdent trois génomes capables de coder pour approximativement 50 000 gènes chacun (Choulet *et al.*, 2010). Identifier la fonction des gènes de tels organismes est généralement problématique. Pour compliquer le processus d'annotation, les informations fonctionnelles pertinentes concernant les espèces proches du blé sont généralement dispersées dans des bases de données biologiques sémantiquement hétérogènes. Il existe pourtant des standards d'annotation fonctionnelle très répandus dans

le milieu de la biologie. Combiner et extraire l'information de ces données et les rendre accessibles dans un format unifié pourraient donc améliorer nos capacités de prédiction de la fonction des gènes de blé.

La définition même d'une fonction biologique est ambiguë et dépend du contexte de recherche. Il semble important de relever que la fonction biologique d'une protéine a plusieurs aspects. Si l'on choisit l'exemple d'une protéine kinase, dans un contexte biochimique, la fonction d'une kinase est la phosphorylation d'un groupe hydroxyle d'un substrat spécifique. Toutefois, les protéines remplissent leur fonction au sein d'un organisme, et cela a des conséquences au niveau cellulaire comme au niveau de l'organisme étudié dans son ensemble. Dans un contexte physiologique, cette même kinase peut faire partie d'une voie de signalisation, dans laquelle elle phosphoryle un ou plusieurs partenaires d'interaction. Une mutation dans cette kinase pourrait faire naitre l'apparition d'un phénotype particulier. Par conséquent, lorsque l'on parle de la fonction d'une protéine, il faut toujours spécifier le ou les contextes de la description fonctionnelle.

Actuellement, les séquences et les informations structurales sont analysables par des méthodes informatiques, les séquences génomiques, transcriptomiques ou protéiques sont retranscrites sous forme de chaines de caractères qui sont adaptées à de nombreux processus tels que l'alignement de séquences, la recherche de motifs et toute autre étape d'extraction d'information biologique. Cette propriété des séquences permet donc de traiter les séquences biologiques par des algorithmes informatiques de manière relativement directe.

Cependant, contrairement à la séquence et à la structure, l'annotation fonctionnelle dépend du langage humain et du contexte de recherche. Ces deux vecteurs d'information apportent donc un niveau de complexité supérieur et des subtilités difficilement analysables par des méthodes computationnelles. Le vocabulaire humain évolue et il existe de nombreux termes synonymes, cette observation est d'autant plus vérifiée dans le domaine de la biologie ou le vocabulaire évolue rapidement. Cette synonymie introduit d'avantage de confusion, que ce soit pour les scientifiques mais encore plus pour les machines et leurs algorithmes.

Une caractéristique fondamentale d'un système d'annotation est l'équilibre entre l'utilisabilité par l'homme, la spécificité des catégories et l'exigence de structures standardisées indispensables dans le cadre de l'utilisation d'applications informatiques. Pour rendre l'annotation fonctionnelle à la fois intelligible pour l'homme et accessible à un traitement informatique, il est donc nécessaire de la rendre accessible sous un format structuré, utilisant un vocabulaire standardisé et si possible hiérarchisé. Ce besoin a été

identifié en premier lieu par les biochimistes et a vu la genèse du système EC (*Enzyme Commission Classification*). Le système E.C. répond aux exigences à la fois d'un vocabulaire contrôlé et d'une relation bien définie entre les termes. Il existe cependant d'autres fonctions non abordées par le domaine de l'enzymologie et d'autres aspects fonctionnels en plus de la fonction biochimique. En 1993, le premier programme génomique pour classer les protéines a été suggéré pour *Escherichia coli* (Riley, 1993). L'objectif principal de cette classification était la description de l'ensemble connu des protéines d'*E. coli*. A l'origine, cette classification n'autorisait que l'affectation d'une catégorie fonctionnelle par protéine, mais cela s'est avéré insuffisant pour l'annotation complète de protéines. Par conséquent, la classification imaginée a évolué, rendant possible l'affectation de plusieurs catégories fonctionnelles par protéine (Serres & Riley, 2000)

D'autres systèmes d'annotation ont été proposés, principalement suite à la nécessité d'annoter des projets de génomique émergents (Rison *et al.*, 2000; Whisstock & Lesk, 2003). Le point commun entre ces régimes est la création d'un vocabulaire contrôlé et dans la plupart des cas une classification qui débute par une annotation fonctionnelle générale pour aboutir à une fonction plus spécifique.

Problématique

Le problème qui a motivé la première étape de ce travail de thèse est l'analyse de données transcriptomiques de blé tendre obtenues grâce à une puce à ADN NimbleGen® (Rustenholz *et al.*, 2011). Cette puce contient des sondes permettant de suivre l'expression de 40 642 gènes dont 2 891 facteurs de transcription. Le génome du blé n'étant pas encore totalement séquencé, il était nécessaire de connaître de façon efficace la fonction des gènes qui pouvaient apparaitre dans les réseaux inférés par la suite. Bien qu'il subsiste encore une grande quantité d'informations fonctionnelles stockées sous forme de textes non structurés, il existe un certain nombre de systèmes, nomenclatures et standards d'annotation fonctionnelle en biologie. Des solutions existent afin d'extraire de l'information de données non structurées, l'utilisation de l'ingénierie linguistique par exemple, une discipline à l'interface de l'informatique et de la linguistique est envisageable. Cette discipline est aujourd'hui communément utilisée afin d'identifier dans les données non structurées des références à des standards d'annotation. Ces méthodes sont néanmoins difficiles à implémenter. Une solution plus triviale mais tout aussi efficace est l'utilisation des standards d'annotation, les plus répandus qui ont été utilisés dans ce travail de thèse, sont le système G.O. (Gene Ontology ; NC-IUBMB, 1999), le système PlantCyc, le système FunCat (Munich

Information Center for Protein Sequences Functional Catalogue ; Ruepp *et al.*, 2004) et la classification hiérarchique de MapMan (Thimm *et al.*, 2004). Il existe également des correspondances entre ces différents standards d'annotation même si tous ne suivent pas les mêmes principes de classification ou de hiérarchisation.

A l'origine, le système E.C.

Le système E.C. a d'abord été publié en 1961, a évolué et a été présenté en version imprimée pour la dernière fois en 1992 (IUBMB 1992). E.C. est depuis maintenu à jour en version web et contient actuellement plus de 3 700 identifiants. E.C. propose une nomenclature hiérarchisée permettant de distinguer les enzymes selon les réactions qu'elles catalysent. Il existe dans ce système trois niveaux de hiérarchie, le plus haut étant composé de six classes d'enzymes (oxydoréductases, transférases, hydrolases, lyases, isomérases et ligases), chacune étant ensuite sous-divisée en deux niveaux hiérarchiques inférieurs, le dernier identifiant de l'enzyme étant le numéro de série de l'enzyme. Ainsi l'identifiant E.C. d'une enzyme est composé de 4 nombres séparés par des points (e.g. 3.4.13.7 pour une Glu-Glu dipeptidase). Un exemple de décomposition d'un identifiant E.C. et de l'identification d'une Glu-Glu dipeptidase est illustré par la table 1.

Table 1 - Décomposition hiérarchique de la classification E.C. d'une enzyme de type Glu-Glu dipeptidase (3.4.13.7).

Identifiant E.C.	Spécificité de la réaction	Exemple
3.-.-.-	Type de réaction	Hydrolase
3.4.-.-	Classe du substrat	Peptidase
3.4.13.-	Sous-classe du substrat	Dipeptidase
3.4.13.7	Substrat	Glu-Glu-dipeptidase

Ce système n'est pas utilisé dans notre travail car il prend une orientation purement biochimique, néanmoins il représente la genèse des standards d'annotation fonctionnelle.

Le système Gene Ontology.

Le système G.O. est actuellement le standard dominant dans le domaine de l'annotation fonctionnelle lisible par les machines. G.O. propose un cadre composé d'un vocabulaire contrôlé décrivant trois aspects de la fonction d'une protéine : la fonction moléculaire, le processus biologique et la localisation cellulaire.

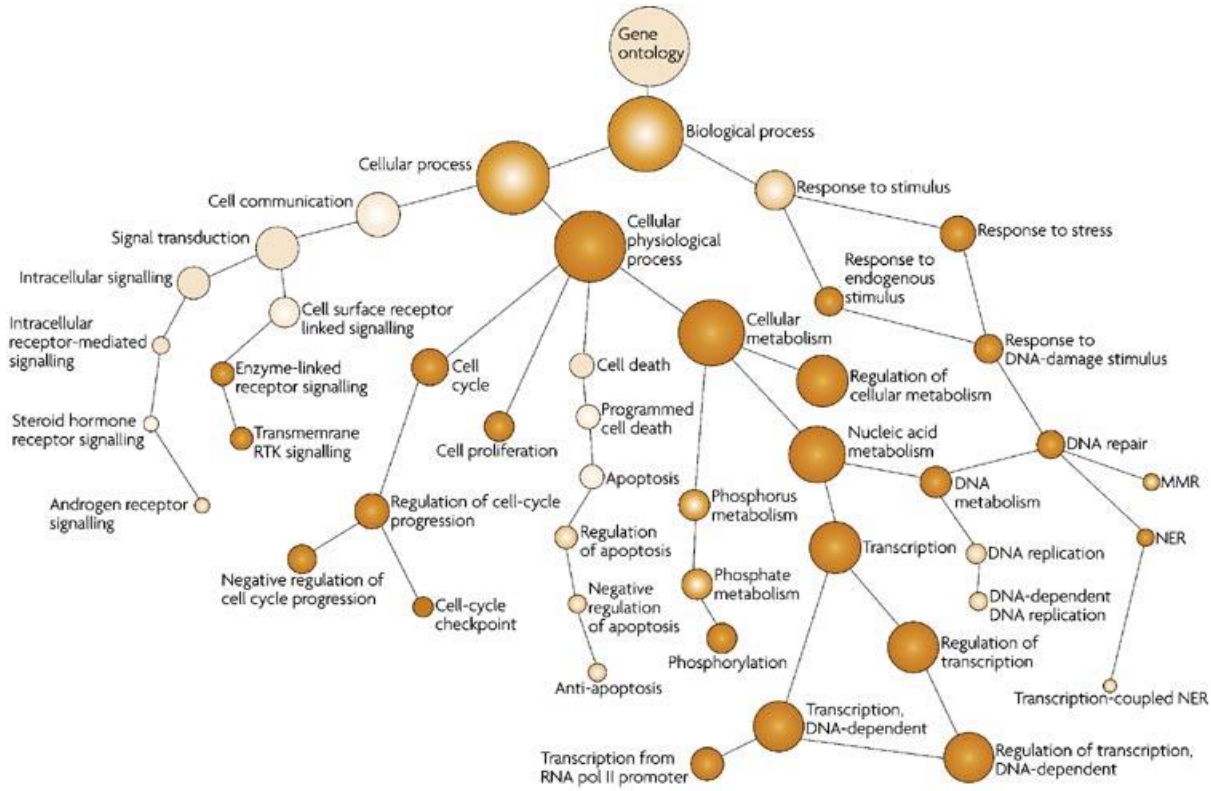


Figure 5 - Une représentation graphique du système Gene Ontology. *La classification imbriquée montre les annotations fonctionnelles attribuées à des gènes impliqués dans le cancer chez l'homme. La taille des nœuds du graphe est proportionnelle à l'adhésion totale de gènes, et la couleur indique le degré de signification statistique (du plus clair pour les moins significatifs au plus foncé pour les plus significatifs). D'après Hu et al. (2007).*

La catégorie *cellular component* fait référence à des constituants de la cellule ou de son environnement extérieur, et permet de localiser le produit du gène dans la cellule (e.g. GO:0016020 : membrane). Cette catégorie, bien que n'étant pas un aspect fonctionnel en tant que tel, est jugée importante pour l'annotation fonctionnelle puisque les protéines fonctionnent dans un contexte bien défini de la cellule vivante. La catégorie *molecular function* est une annotation fonctionnelle qui décrit l'activité biochimique du produit du gène (e.g. GO:0005215 : transporter activity). Enfin, la catégorie *biological process* permet de décrire le processus biologique, c'est à dire la réaction biochimique dans laquelle la protéine est impliquée. Elle résume l'objectif biologique qui est visé par l'association ordonnées de plusieurs fonctions moléculaires (e.g. GO:0050896 : response to stimulus). Chaque ontologie est implémentée comme un graphe orienté acyclique (Figure 5) où les termes sont représentés comme des nœuds dans le graphe et sont disposés hiérarchiquement du plus général au plus spécifique. Cette structure signifie que chaque nœud peut avoir plus d'un seul parent, afin de décrire les fonctions qui sont impliquées dans plusieurs processus biologiques,

compartiments cellulaires, ou fonctions moléculaires. Les identifiants des G.O. sont écrits de manière standardisée, « GO: » suivi de 7 chiffres (e.g. GO:0007049 Cycle cellulaire).

La collection des voies métaboliques PlantCyc

MetaCyc propose une classification de voies métaboliques et d'enzymes provenant de tous les domaines du vivant. Elle se matérialise par plus de 1 800 voies métaboliques sélectionnées à partir de plus de 30 000 publications de recherche. La plupart des voies métaboliques de MetaCyc sont liées à une ou plusieurs enzymes caractérisées fonctionnellement. MetaCyc se ramifie en ressources spécifiques à chacun des organismes de leur catalogue, AraCyc par exemple représente une ressource spécifique à *Arabidopsis thaliana* tandis que RiceCyc est spécifique au riz. La classification des enzymes de MetaCyc présente une structure en arbre avec 6 branches de niveau 1 (Biosynthèse, Dégradation/Utilisation/Assimilation, Formation de métabolites précurseurs et énergie, Détoxification, Activation/Inactivation/Inter-conversion et Clusters métaboliques). Les trois premières branches citées représentent la grande majorité des annotations tandis que les trois autres, moins importantes en contiennent moins de 5%.

Classification fonctionnelle du MIPS

En comparaison avec d'autres systèmes d'annotation fonctionnelle, la classification fonctionnelle du MIPS nommée FunCat se veut être un compromis équilibré entre profondeur, largeur et précision, mais sans être trop granuleuse. Comme le progrès scientifique est constamment révélateur de nouvelles connaissances de la biologie des organismes, le développement de la classification fonctionnelle FunCat est un processus continu et est mis à jour régulièrement.

La classification du MIPS permet la description fonctionnelle de protéines d'origine eucaryote et procaryote. C'est une classification en arbre composée de 28 branches principales (Table 2) qui couvrent des fonctions générales telles que le transport cellulaire, le métabolisme ou la communication cellulaire et la transduction du signal. Chaque branche est ensuite sous-divisée jusqu'à six fois pour identifier une fonction de plus en plus spécifique. Au total, 1 363 catégories fonctionnelles sont fournies par ce catalogue. Ces catégories fonctionnelles ne sont pas spécifiques d'une espèce mais permettent l'annotation fonctionnelle d'un large éventail d'organismes. La classification contient toutefois des sous-catégories dédiées à des groupes spécifiques d'organismes. En outre, une des 28 catégories principales est dédiée aux transposons, aux protéines virales et aux protéines plasmidiques. Chacune des catégories

fonctionnelles est affectée à un numéro unique à deux chiffres. Le cadre vers le haut de l'arbre hiérarchique se compose du préfixe des nœuds précédents, situées dans les niveaux supérieurs de la hiérarchie. Les niveaux de catégories sont séparés par des points, par exemple 01 métabolisme est un représentant de plus haut niveau, et 01.01.03.02.01 biosynthèse du glutamate appartient au niveau le plus spécifique de FunCat. Pour la majorité des protéines, la fonction cellulaire ne peut pas être entièrement décrite avec une seule catégorie fonctionnelle, mais doit être considérée comme la somme des propriétés différentes.

Table 2 – Catégories principales utilisées par la classification FunCat. D'après Ruepp et al. (2004).

Functional classification catalogue (FunCat) version 2.0	
Metabolism	
	1 Metabolism
	2 Energy
	4 Storage protein
Information pathways	
	10 Cell cycle and DNA processing
	11 Transcription
	12 Protein synthesis
	14 Protein fate (folding, modification and destination)
	16 Protein with binding function or cofactor requirement (structural or catal
	18 Protein activity regulation
Transport	
	20 Cellular transport, transport facilitation and transport routes
Perception and response to stimuli	
	30 Cellular communication/signal transduction mechanism
	32 Cell rescue, defense and virulence
	34 Interaction with the cellular environment
	36 Interaction with the environment (systemic)
	38 Transposable elements, viral and plasmid proteins
Developmental processes	
	40 Cell fate
	41 Development (systemic)
	42 Biogenesis of cellular components
	43 Cell type differentiation
	45 Tissue differentiation
	47 Organ differentiation
Localization	
	70 Subcellular localization
	73 Cell type localization
	75 Tissue localization
	77 Organ localization
	78 Ubiquitous expression
Experimentally uncharacterized proteins	
	98 Classification not yet clear-cut
	99 Unclassified proteins

MapMan

MapMan est avant tout un outil permettant de visualiser des données d'expression sur des schémas picturaux qui représentent symboliquement des domaines de fonction biologique. Mais MapMan propose sa propre ontologie, conceptualisée par une structure en arbre comprenant 35 branches principales pour un total de 1 854 catégories fonctionnelles appelées *bins*. Dans la classification de MapMan, les classes peuvent être des secteurs de métabolisme, des fonctions cellulaires spécifiques (par exemple la synthèse de protéines), une réponse biologique (par exemple des gènes impliqués dans le métabolisme et/ou la réponse à une hormone). En utilisant des catégories hiérarchiques et des diagrammes, les différents secteurs fonctionnels peuvent être analysés à différents niveaux de résolution, en fonction de la question scientifique et de la quantité d'information disponible. MapMan suit un paradigme visant à assigner le minimum de *bins* possible à un gène ou une protéine. Les diagrammes proposés sont ensuite comparables entre espèces.

Comparaison de l'annotation d'une phytoène synthétase dans les différents systèmes

L'exemple d'un gène codant pour une phytoène synthétase (Table 2) chez *A. thaliana* (ATG17230) et son annotation dans les différents systèmes d'annotation présentés permet de mettre en évidence la consistance et les spécificités des différents systèmes.

Table 3 – Annotation fonctionnelle d'un gène (ATG17230) codant pour une protéine phytoène synthétase dans différents systèmes.

Système	Identifiant	Annotation fonctionnelle
G.O.		
	GO:0009507	*Composant cellulaire* : chloroplaste
	GO:0009536	*Composant cellulaire* : plaste
	GO:0016117	*Processus biologique* : synthèse de caroténoïdes
	GO:0016767	*Fonction moléculaire* : activité phytoène synthétase
	GO:0046905	*Fonction moléculaire* : activité phytoène synthétase
E.C.		
	2.5.1.32	*rephytoene diphosphate = all-trans-phytoene + diphosphate*
MIPS		
	01.06.06.13	Métabolisme des caroténoïdes
	70.26.03	Chloroplaste
MapMan		
	16.1.4.1	Métabolisme des caroténoïdes, phytoène synthétase

L'outil d'annotation développé au cours de la thèse : dbWFA

L'outil d'annotation qui a été développée au cours de la thèse s'inscrit dans ce challenge d'annotation des données transcriptomiques haut débit. Afin de combler la carence en ressources pour l'annotation fonctionnelle efficace des transcrits, nous avons développé dbWFA, une base de données basée sur l'homologie des séquences de l'UniGene du NCBI et des ADNc pleine longueur de TriFLdb (Mochida *et al.*, 2009). Nous avons utilisé pour homologues de référence les données d' *Arabidopsis thaliana* (Lamesch *et al.*, 2011), du riz (Ouyang *et al.*, 2007) et du maïs (Schaeffer *et al.*, 2011). Au moment du développement de cet outil et encore à l'heure actuelle, peu de génomes de plantes sont bien annotés. L'annotation du génome de *Brachypodium distachyon* (The International Brachypodium Initiative, 2010) ne fait que croitre en qualité et en quantité et pourrait aujourd'hui représenter une intéressante source de données à intégrer à l'outil. Il en est de même pour le *T. urartu* dont le génome a été récemment séquencé (Ling *et al.*, 2013).

Nous avons également lié l'outil avec la base de données de facteurs de transcription (FT) de blé wDBTF (Romeuf *et al.*, 2010). Les caractéristiques précises de cette base de données ainsi que son application dans le cadre d'une analyse transcriptomique sur le blé sont présentés dans la ***publication 1***.

INFERENCE DE RESEAUX ET BIOLOGIE DES SYSTEMES

Contexte

La biologie systémique est une approche interdisciplinaire de la biologie qui élargit le champ de la biologie moléculaire pour étudier des ensembles d'éléments en interaction les uns avec les autres. L'ensemble des interactions concertées entre gènes, ARNm, protéines et métabolites constitue des réseaux de régulations biologiques dont l'élucidation est un enjeu important de la biologie systémique. Les méthodes d'analyse du transcriptome permettent l'analyse globale de l'expression des gènes en quantifiant l'expression de milliers de gènes simultanément. La complexité du vivant est ainsi projetée dans un espace mesurable sur lequel les conséquences de modifications génétiques ou de l'environnement peuvent être étudiées par des analyses informatiques, mathématiques et statistiques. L'analyse des profils d'expression des différents gènes, protéines et métabolites comme l'ensemble des composants d'un système et non comme une masse de données indépendantes permet de mettre en évidence des relations qui existent au sein de réseaux de régulations globaux. Le principe de l'inférence de réseaux de gènes est d'extraire des interactions entre les entités moléculaires à partir de données d'expression, en utilisant des algorithmes mathématiques.

Biologie des systèmes

Afin de comprendre l'organisation de systèmes biologiques complexes, il est souvent avantageux de les conceptualiser comme des systèmes d'éléments en interaction. Dans le cadre d'une telle approche, qui représente l'essence même de la biologie des systèmes, les objectifs sont **(I)** d'identifier les constituants du système biologique étudié, **(II)** de déterminer le comportement dynamique de ces constituants (i.e. observer leurs changements d'activité au cours du temps et/ou dans différentes conditions) et **(III)** d'identifier les interactions entre les constituants du système.

La biologie des systèmes trouve ses origines dans la théorie générale des systèmes apparue progressivement dans la seconde partie du XX^{ème} siècle. C'est une vision holistique qui permet d'aborder des systèmes complexes difficilement conceptualisables par les approches parcellaires largement utilisées en ce temps. La théorie des systèmes repose sur l'hypothèse que tout processus peut être visualisé comme un réseau d'interactions entre éléments et que tout système, quelle que soit sa nature, peut être exploré par le biais des mêmes méthodes (Von Bertalanffy, 1969; Voit, 2000). Elle se détache des approches plus réductionnistes et est fondée sur l'hypothèse que les éléments observables ne fonctionnent pas de manière isolée

et additive mais agissent de manière coordonnée et en interaction au sein d'un réseau. La théorie des systèmes prend aujourd'hui une importance capitale dans de nombreux domaines de recherche grâce aux nouvelles technologies qui permettent d'accumuler des masses de données très importantes.

La structure de certains réseaux peut être triviale tandis que pour certains systèmes biologiques, l'application de méthodes issues de la théorie des graphes est indispensable. La représentation la plus simple pour un système biologique est de réduire chaque unité du système à un nœud (aussi appelé sommet) et chaque interaction entre paire de nœuds à un lien (aussi appelé arête). Dans les systèmes biologiques qui nous intéressent ici les nœuds peuvent représenter une protéine, un transcrit, un métabolite ou toute autre entité moléculaire. Les arêtes dirigées aussi appelées arcs possèdent un nœud source et un nœud destination, et sont plus adaptées à la représentation de transformations biochimiques ou de liens de régulation tandis que les arêtes non orientées sont plus adaptées pour représenter des interactions protéine/protéine ou des interactions dont la source et la destination ne sont pas identifiables. En fonction de la profondeur des connaissances que l'on peut extraire des données, il est possible d'établir un réseau avec des propriétés d'arcs plus ou moins précises (signe, poids, confiance, vitesse de réaction etc.) et des états de nœuds variables dans des réseaux dynamiques par exemple.

L'aspect de la biologie des systèmes abordé dans ce travail de thèse est la construction de réseaux à partir de méthodes permettant d'inférer des interactions ou des relations fonctionnelles entre éléments du système à partir de données accumulées sur l'état des éléments du système dans différentes conditions. Les liens inférés à partir des données permettent de construire le réseau sous-jacent au système étudié. L'analyse du réseau permet ainsi d'obtenir de nouvelles connaissances et prédictions biologiques. La modélisation dynamique de réseaux est une étape supplémentaire dans l'inférence de réseaux, elle permet de décrire le comportement des nœuds et des interactions du système dans différentes conditions. Connaître la dynamique d'un réseau permet aux scientifiques de prédire son comportement face à des perturbations. Quel que soit le type de réseau inféré, l'objectif est toujours de comprendre, prédire le comportement, et si possible d'être capable de contrôler le système étudié.

L'utilisation la plus répandue des approches d'inférence de réseaux en biologie s'applique aux données d'expression de gènes, autrement dit le domaine de la transcriptomique. Il s'agit dans ce cas de prédire quels gènes/protéines influent quels gènes/protéines par le biais de

régulations transcriptionnelles, post-transcriptionnelles, traductionnelles ou post-traductionnelles. Les liens de régulation inférés sont ensuite visualisés au sein d'un réseau de régulations et doivent par la suite être validés expérimentalement.

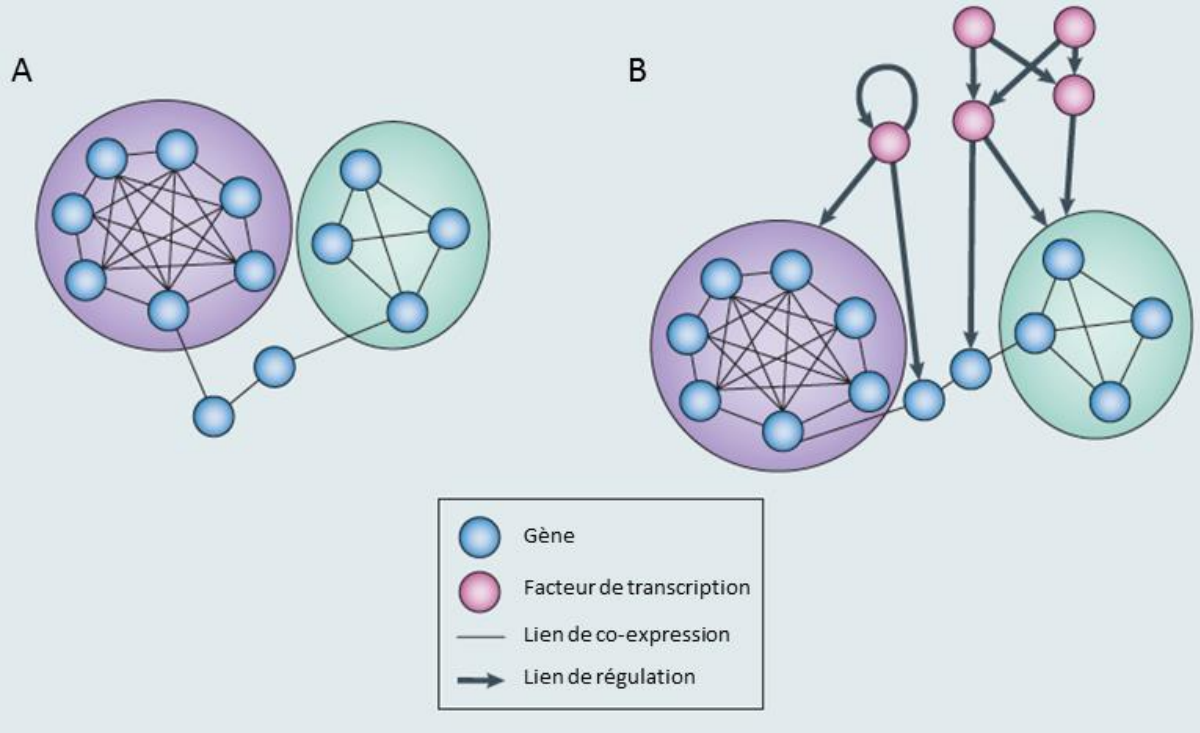


Figure 6 - Représentation de réseaux *(A) Réseau de co-expression, dans lequel chaque nœud représente un gène, chaque arête représente un certain degré de similarité entre les profils d'expression de deux gènes. Les modules entourés en violet et vert représentent des sous-réseaux fortement connectés. Les arêtes ne sont pas orientées et ne représentent pas des liens de causalité. (B) Réseau de régulations transcriptionnelles, dans lequel les nœuds représentent des gènes ou des facteurs de transcription qui ont un effet sur des gènes ou des modules entiers. Les modules entourés en violet et vert représentent des sous-réseaux hautement connectés. Les arêtes représentent des liens de causalité. Adapté de De Smet & Marchal (2010).*

La majeure partie des approches utilisées reposent sur un postulat de départ identique, à savoir que des gènes ayant des profils d'expression similaires (quelle que soit la condition) sont susceptibles d'être fonctionnellement liés, c'est le principe de « guilt by association » (Oliver, 2000). Les méthodes basées essentiellement sur ce principe permettent de construire des réseaux de co-expression qui sont non orientés. Les méthodes capables d'inférer des réseaux de régulations transcriptionnelles vont un niveau plus loin dans l'analyse puisqu'ils permettent de caractériser le sens et le type d'interaction entre deux entités moléculaires (Figure 6).

L'application de telles méthodes d'inférence à des données publiques concernant des organismes modèles a permis d'améliorer considérablement notre compréhension des réseaux de régulations transcriptionnelles. Les régulons de bactéries par exemple, qui n'impliquent que quelques opérons, possèdent une structure modulaire (des groupes de gènes co-exprimés appelés modules). L'organisation modulaire des opérons est cruciale et conservée dans certaines conditions tandis qu'elle peut être brisée par la présence de promoteurs intra-opéroniques dans d'autres conditions (Fadda *et al.*, 2009; Mendoza-Vargas *et al.*, 2009). Par ailleurs, des régulations complexes, impliquant plusieurs régulateurs entrainent souvent l'expression différentielle d'un seul gène, ce qui limite la portée des méthodes d'inférence basées sur la co-expression (Lemmens *et al.*, 2009). En outre, des études sur *E. Coli* ont montré que de nombreux FT possèdent des profils d'expression similaires et régulent des groupes de gènes similaires, suggérant ainsi une activité redondante de certain FT et/ou une coopération complexe entre différents FT déclenchant une réponse unique.

Un certain nombre de résultats marquants ont ouvert la voie à l'adoption des méthodes d'inférence de réseaux dans le domaine de la biologie. La mise en évidence du lien entre les erreurs de traduction et la réponse au stress oxydatif chez *E. coli* a été mise en évidence par le biais d'une approche combinant l'inférence de réseaux et des étapes de validation fonctionnelle (Kohanski *et al.*, 2008). Grâce à l'utilisation de données d'expression publiques, Bassel *et al.* (2011) ont établi un réseau d'interactions transcriptomiques impliquées dans la reprogrammation qui survient chez la graine d'*A. thaliana* entre les phases de dormance et de germination. Le réseau ainsi établi met en lumière des groupes de gènes impliqués dans cette reprogrammation et leurs liens avec des master-régulateurs connus. Par le biais d'une approche similaire, Verdier *et al.* (2013) ont découvert un certain nombre de gènes impliqués dans la tolérance à la dessiccation des graines de *Medicago truncatula* en fin de développement. Leur approche a également permis de mettre en évidence des liens entre certains gènes alors inconnus et des master-régulateurs du développement de la graine. Récemment, l'analyse transcriptomique de lignées transgéniques de tomate (*Solanum lycopersicum*) combinée à une approche d'inférence de réseaux de neurones artificiels (Pan *et al.*, 2013) a permis de mettre en évidence un gène impliqué dans l'accumulation de pigments dans les tissus du fruit.

L'inférence de réseaux constitue donc une ressource permettant d'extraire des connaissances à partir des informations expérimentales mais avant de l'utiliser dans le cadre de la biologie des systèmes, il est nécessaire d'en connaître les différentes méthodes et leurs avantages et

inconvénients. Il est possible de distinguer les différentes méthodes d'un point de vue purement informatique, il est néanmoins plus intéressant, dans le cadre de la biologie, d'utiliser le point de vue du biologiste qui cherche à déterminer le type de méthode le plus adapté à l'objet de sa recherche.

En considérant qu'une espèce au génome complexe comme le blé tendre possède environ 150 000 gènes, on pourrait envisager d'étudier une à une toutes les interactions possibles, soit en théorie plus de 22 milliards de possibilités. En prenant en compte le fait qu'il existe des boucles de rétroaction et des régulations au caractère combinatoire, on réalise rapidement la complexité difficilement abordable sans l'utilisation d'algorithmes adaptés et optimisés. Le grand champ de possibilités, conjointement avec le nombre limité de points de données indépendants, font de l'inférence de réseaux biologiques un challenge ambitieux.

Une première stratégie de simplification vise à inférer des réseaux modulaires, ce qui revient à rassembler des gènes aux profils d'expressions similaires dans des modules, qui seuls seront pris en compte pour l'inférence de liens de causalité (Figure 7). Cette stratégie implique ainsi une perte d'information mais un gain de temps computationnel et une conceptualisation simplifiée de la réalité biologique. Cette stratégie peut s'appliquer de la même façon aux régulations combinatoires, pour lesquelles les méthodes d'inférence actuelles peuvent très difficilement élucider la complexité de la combinaison de gènes impliqués mais surtout les interactions sous-jacentes. Les méthodes basées sur la découverte de modules n'ont pas pour but d'assigner un programme transcriptionnel à chaque gène du jeu de données mais plutôt à des groupes de gènes aux profils d'expression similaires. Une seconde stratégie se rapporte à la combinaison des données d'expression avec différentes sources d'information telles que des données concernant des motifs de régulation (cibles des FT) ou des connaissances précédemment acquises sur les interactions. Il est par exemple possible de donner plus grande importance aux interactions validées indépendamment par différents types de données. Il peut alors s'agir de méthodes supervisées ou semi-supervisées. Une autre stratégie qui regroupe les méthodes dites dirigées (*query-driven*) permet de limiter l'espace de recherche en limitant de manière intentionnelle et à priori du processus d'inférence le nombre d'interactions à évaluer (Figure 7).

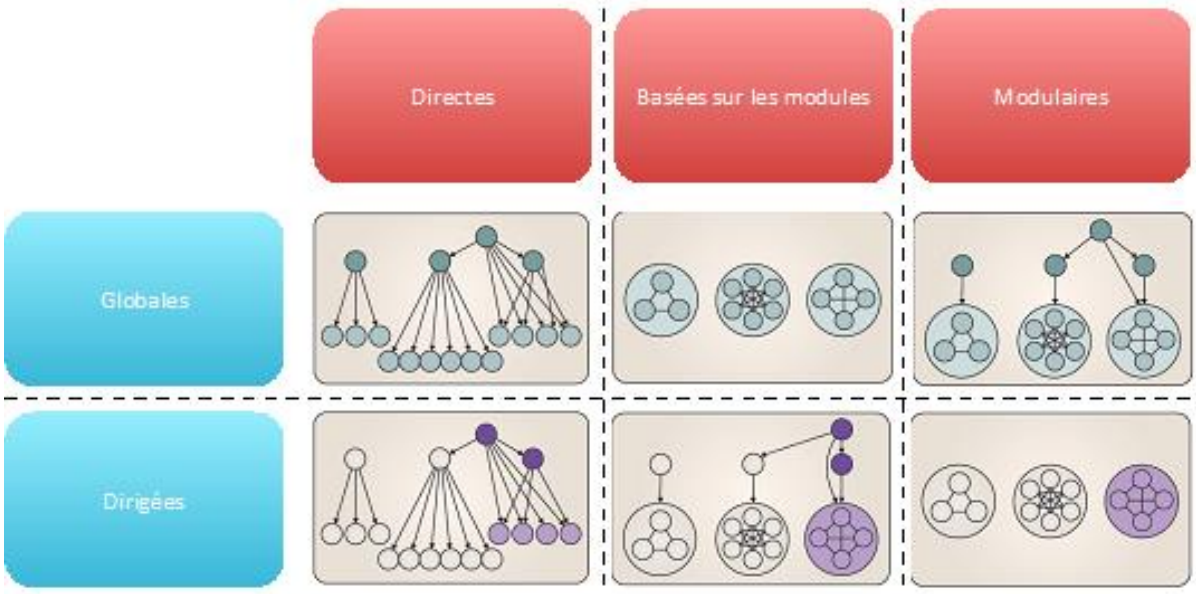


Figure 7 - Classification des méthodes d'inférence de réseaux. *Les méthodes d'inférence sont classées par les stratégies d'approche qu'elles implémentent. Les méthodes directes prennent en compte chaque gène individuellement tandis que les méthodes modulaires ou basées sur la découverte de modules traitent des groupes de gènes aux profils d'expression similaires comme une entité unique (modules). Le résultat obtenu par chaque type de méthode peut être global, si la totalité des données est traitée sans discrimination, ou dirigé (en violet) s'il existe à priori de l'inférence, une hypothèse prédéfinie prise en compte dans la recherche de liens. D'après De Smet & Marchal (2010).*

Les méthodes d'inférence

Les méthodes directes opposées aux méthodes modulaires

Puisque chaque stratégie adoptée repose sur différentes hypothèses, et que chacune pose différentes contraintes, la stratégie à adopter détermine à priori, le type de réseaux qui peut être élucidé. De par les stratégies sur lesquelles elles reposent, on peut opposer les méthodes d'inférences « directes » aux méthodes basées sur la découverte de modules. L'exploitation du concept de modularité offre des avantages tant du point de vue biologique que statistique. La plupart des méthodes basées sur la découverte de modules ne prédisent pas seulement des interactions mais apportent aussi une information concernant les conditions expérimentales dans lesquelles ces modules sont impliqués. Cette information est essentielle puisqu'elle va permettre de préparer des conditions expérimentales adaptées lors de la validation fonctionnelle des résultats obtenus (Bonneau *et al.*, 2006 ; Lemmens *et al.*, 2009). L'approche modulaire permet également de lier des gènes possédant des profils d'expression peu similaires, ce qui représente à la fois un avantage et un inconvénient. Cela permet en effet de regrouper les gènes impliqués dans des voies métaboliques impliquant un grand nombre de gènes par exemple. En revanche, chez les bactéries, chez qui l'expression des

régulateurs opéroniques est très similaire à celle de leurs cibles, le concept de modularité entraine des erreurs de prédiction (Michoel *et al.*, 2009).

Les méthodes globales opposées aux méthodes dirigées

Les méthodes globales sont généralement utilisées pour identifier des processus biologiques impliquant un grand nombre de gènes et qui sont impliqués dans des réponses métaboliques majeures ou dans des changements de conditions expérimentales importants. Ces méthodes fournissent une vue globale des réseaux de régulations impliqués dans le changement de l'état physiologique de l'objet d'étude. Les méthodes dirigées sont intentionnellement biaisées puisqu'elles permettent de définir à priori de l'inférence, un espace de recherche limité afin de générer un sous-réseau spécifique du problème de recherche. Les approches dirigées sont donc typiquement utilisées afin d'étudier une voie métabolique particulière ou découvrir des gènes inconnus liés à des gènes aux fonctions connues.

Les méthodes supervisées opposées aux méthodes non-supervisées

Les méthodes supervisées se basent sur des groupes d'interactions connues afin de prédire de nouvelles interactions. Elles utilisent un jeu de données d'entrainement (interactions connues) avant de dériver les caractéristiques des interactions connues pour chercher de nouvelles interactions similaires. La fiabilité de telles méthodes dépend fortement de la significativité du jeu de données d'entrainement. Encore une fois, si ces méthodes basées sur l'apprentissage peuvent donner des résultats satisfaisants, il est nécessaire de posséder un jeu de données d'entrainement de qualité et de taille conséquente. Cette nécessité représente un inconvénient majeur dans le cadre de l'étude d'organismes aux bases moléculaires peu connues comme le blé.

Les méthodes intégratives opposées aux méthodes non-intégratives

Les méthodes intégratives sont basées sur l'hypothèse que toute l'information n'est pas portée par l'expression des gènes, au contraire des méthodes non-intégratives qui font abstraction de l'importance des régulations non transcriptionnelles. De plus, les méthodes non-intégratives ne permettent pas de distinguer les réelles interactions des simples similarités entre profils d'expression. L'utilisation de séries temporelles permet d'atténuer cette incertitude puisqu'elles contiennent intrinsèquement une information pour les régulations FT-Gène cible par exemple, si l'on considère que l'effet du changement d'expression d'un FT sera observé aux temps suivants sur le gène cible. Néanmoins, inférer de tels réseaux nécessite l'utilisation de méthodes capables de saisir cette dynamique. En pratique, il reste difficile de tirer parti de séries de données temporelles puisqu'il est difficile

techniquement d'obtenir des intervalles de temps suffisants et tout aussi difficile de distinguer les réels changements dynamiques d'expression du bruit de fond. En outre, les changements d'expression transitoires qui peuvent être observés pour certains gènes sont difficilement identifiables.

Les approches informatiques

L'inférence de réseaux suscite un intérêt croissant auprès de la communauté scientifique et notamment dans le domaine de la biologie. Si une première distinction permet de classer ces méthodes par la stratégie proposée, il est également intéressant d'observer ce qu'il se passe dans la « boite noire » que peut représenter une méthode d'inférence du point de vue d'un biologiste. On pourra distinguer les méthodes les plus communément implémentées telles que les réseaux booléens, l'analyse de corrélation, les Relevance Networks, les réseaux bayésiens et la découverte de règles.

Le principe des **réseaux booléens** est de binariser des données d'expression temporelles, ainsi, à un instant donné, un gène est soit dans un état actif soit dans un état inactif. Il s'agira de fixer des seuils permettant de déterminer les gènes exprimés qui seront représentés par l'état actif et les autres qui seront représentés par l'état inactif. La finalité étant de déterminer des règles logiques permettant d'expliquer l'état d'un gène à un instant donné lorsque l'on connaît l'état d'autres gènes à un instant donné (instant identique ou non ; Liang *et al.*, 1998). L'approche des réseaux booléens a évolué et il s'agit plus souvent de réseaux booléens probabilistes (Zhou, 2010).

Les **Relevance Networks** quant à eux sont basés sur la comparaison exhaustive de toutes les paires de gènes par le biais d'un calcul de corrélation. L'ensemble des coefficients de corrélation doit ensuite être filtré afin de ne conserver que les connexions les plus fortes. Pour aller plus loin, il est possible d'utiliser le calcul d'une pente métrique afin d'élucider non seulement la présence d'une relation entre deux gènes, mais aussi son orientation (Klie *et al.*, 2012).

Les **réseaux bayésiens** sont des modèles basés sur des distributions de probabilités conditionnelles entre les variables (Friedman *et al.*, 2000). Ces modèles sont intéressants pour leur capacité à décrire les processus complexes. Plusieurs modèles sont déterminés à partir des données. A chaque modèle correspond un score permettant de choisir celui qui a la plus grande probabilité d'être correct. L'approche bayésienne se décline aujourd'hui sous

plusieurs formes comme les réseaux bayésiens dynamiques par exemple (Kaderali *et al.*, 2009).

Les **réseaux de Petri** regroupent des méthodes graphiques et de modélisation mathématique applicables à de nombreux systèmes. Le concept de ces réseaux est apparu pour la première fois au début des années 1960 (Petri, 1962). Selon la définition d'origine, les réseaux de Petri représentent des processus simultanés. Les transitions entre les états des nœuds se déclenchent comme une réaction en chaine, et de manière non déterministe. La transition entre chaque état peut être définie par une fonction probabiliste faisant du réseau de Petri un réseau stochastique (Heiner *et al.*, 2009). Ces réseaux sont particulièrement adaptés pour la représentation de systèmes tels que les voies de signalisation (Hamed *et al.*, 2010; Durzinsky *et al.*, 2011).

La **découverte de règles d'association** est une méthode d'apprentissage non supervisée, qui permet de découvrir à partir d'un ensemble de données, un ensemble de règles qui exprime une possibilité d'association entre différents items (exemple : Si le gène G1 est surexprimé alors le gène G2 est également surexprimé. Où G1 et G2 sont des items). Il existe des mesures de qualité qui permettent d'évaluer la significativité des règles inférées, les principales mesures de qualité d'une règle sont le support qui permet d'exprimer la fréquence de la règle dans les données et la confiance qui permet de juger la probabilité que lorsque l'antécédent de la règle est vérifié, alors le conséquent est également vérifié. Cette méthode a été appliquée pour la première fois à l'inférence de réseaux de gènes en 1997 (Brazma *et al.*, 1997) et est présentée dans plusieurs articles plus récents chez l'homme et *Saccharomyces cerevisiae* (Morgan *et al.*, 2007; Martinez *et al.*, 2008) .

Structure des réseaux

La théorie des graphes a permis d'orienter les approches d'inférence de réseaux dans le cadre de la modélisation des réseaux biologiques. Traditionnellement, les réseaux complexes étaient étudiés grâce à la théorie des graphes aléatoires (Erdos & Renyi, 1960). Le modèle d'Erdos-Renyi propose un graphe aléatoire (Figure 8Aa) construit à partir d'un ensemble de *N* nœuds et pour lequel l'ajout de nouvelles arêtes se produit aléatoirement. Ainsi, à partir des *N* nœuds, chaque arête est ajoutée indépendamment avec une probabilité *p.* La distribution du degré de connectivité des nœuds suit ainsi une loi de poisson (Figure 8Ab). Un tel type de réseau possède un faible coefficient de clusterisation, c'est à dire qu'il a tendance à ne pas former des groupes de nœuds fortement connectés. Or, la plupart des réseaux complexes, notamment les réseaux biologiques sont invariants d'échelles (*scale-free*), ce qui

signifie que leur degré de distribution suit une loi puissance (Figure 8Bb). C'est un terme emprunté au vocabulaire de la physique statistique. Les mécanismes à la base de la construction de tels réseaux sont appelés *avantage cumulatif* et *attachement préférentiel* (Price, 1976). Barabási & Albert (1999) rapportent qu'il existe un fort pouvoir d'auto-organisation qui caractérise la topologie *scale-free* des réseaux complexes. Leurs travaux ne se concentrent pas sur les réseaux biologiques, et ils démontrent qu'un grand nombre de réseaux, qu'ils soient sociaux, technologiques ou biologiques, partagent les caractéristiques des réseaux *scale-free*. L'attachement préférentiel stipule que la probabilité pour qu'un nouveau nœud ajouté au réseau se connecte à un autre nœud dépend directement de la connectivité de ce nœud. Une caractéristique importante de ce modèle de réseau est qu'il possède des *hubs* i.e. des nœuds très connectés qui possèdent un rôle essentiel pour le réseau et dont la suppression peut détruire la topologie du réseau. L'observation des systèmes biologiques complexes permet de proposer un modèle dérivé du modèle *scale-free* plus proche de la réalité biologique.

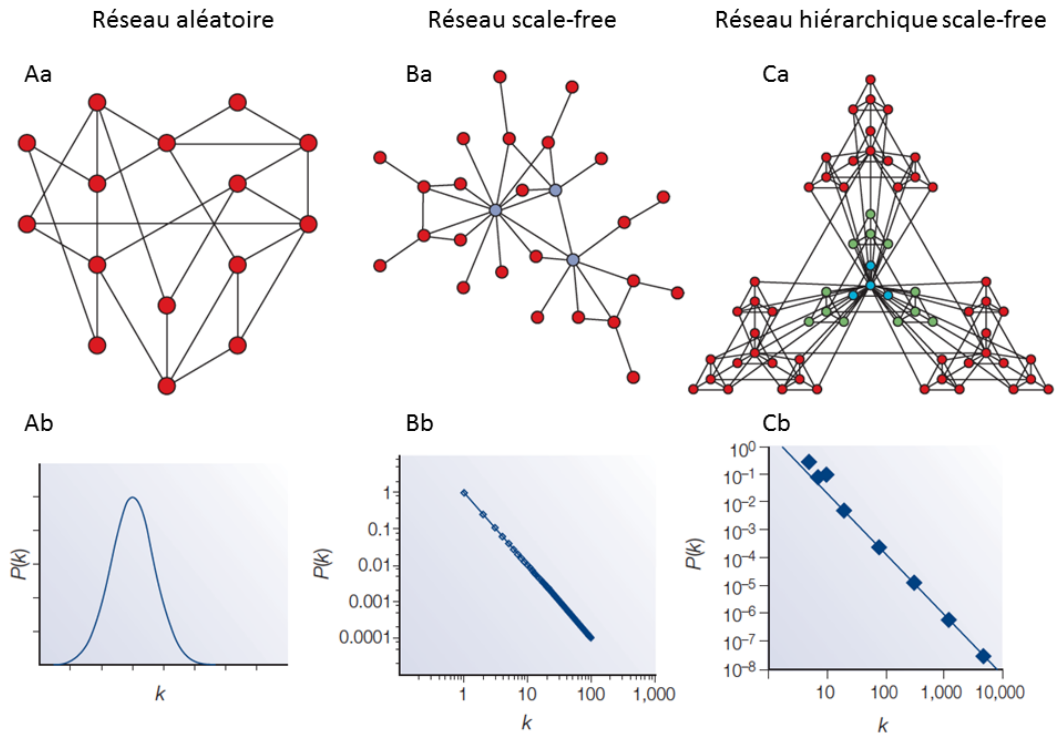


Figure 8 – Représentation (a) et distribution (b) des degrés de connectivité des nœuds pour trois modèles de réseaux. *(A) Modèles de graphe aléatoire proposé par Erdos-Renyi. (B) Réseau scale-free proposé par Barabasi et Albert. (C) Réseau hiérarchique et scale-free. Modèle prenant en compte la coexistence de la modularité, du regroupement local et de la topologie scale-free. D'après Barabasi & Oltvai (2004).*

Ce modèle dit *hiérarchique* et *scale-free* (Figure 8Ca ; Barabási et Oltvai, 2004) prend à la fois en compte la topologie *scale-free* mais aussi la modularité et le regroupement local. Concernant la distribution du degré de connectivité des nœuds, il suit la loi puissance caractéristique des réseaux *scale-free* (Figure 8Cb). Une structure hiérarchique implique que les nœuds faiblement connectés font partie de sous-réseaux fortement connectés, reliés entre eux par le biais des *hubs*, moins nombreux.

L'observation du modèle *scale-free* dans des réseaux très différents (sociaux, biologiques etc.) requiert une explication logique. Il est apparu que deux processus fondamentaux sont à la base de cette structure. Tout d'abord, la plupart des réseaux sont le résultat d'un processus de *croissance* au cours duquel de nouveaux nœuds et arêtes sont régulièrement ajoutés au réseau. C'est par exemple le cas du réseau Internet qui est passé d'une page web à plus de 3 milliards de pages web sur une période de 10 ans. Le second processus repose sur le fait que les nœuds qui apparaissent dans le réseau préfèrent se lier à des nœuds déjà fortement connectés. Ce processus appelé ***attachement préférentiel,*** se vérifie sur le réseau internet, ou les pages web nouvellement créées auront tendance à se connecter aux pages web déjà populaires. Une page web populaire gagnera donc plus rapidement de nouveaux liens que les pages web moins populaires.

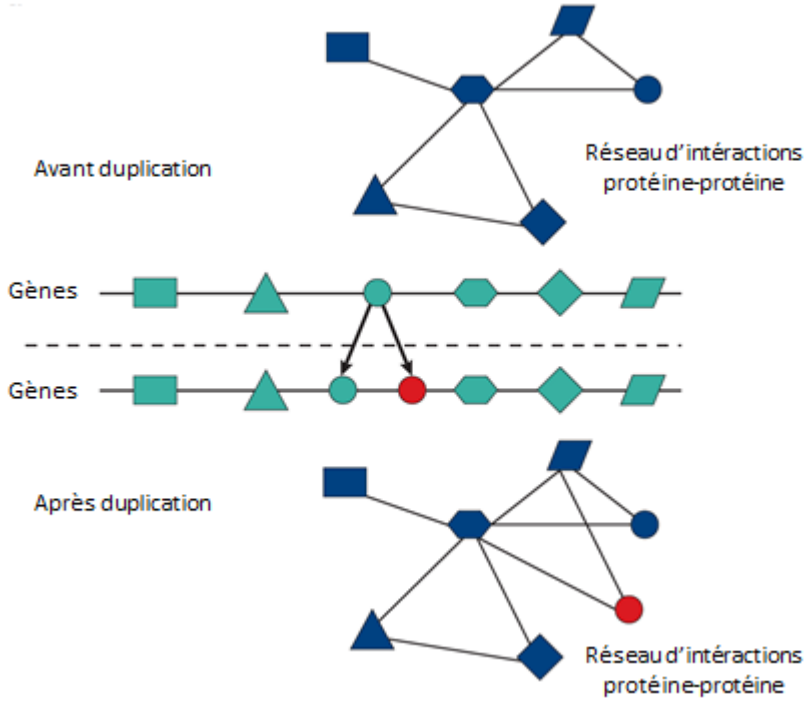


Figure 9 – Origine de la topologie scale-free dans les réseaux biologiques. *Réseau d'interactions protéine-protéine. Lors de la division cellulaire, un ou plusieurs gènes (vert) peuvent être dupliqués (rouge). La nouvelle protéine (rouge) possède la même structure que la protéine d'origine et interagi avec les mêmes protéines. Dans le réseau présenté, le gène dupliqué n'a pas d'importance, dans tous les cas le hub au centre du réseau gagnera une interaction. D'après Barabási et Oltvai (2004)**.***

La croissance et l'attachement préférentiel ont une origine commune dans les réseaux d'interactions protéine-protéine : la duplication des gènes (Figure 9 ; Rzhetsky et Gomez, 2001; Qian *et al.*, 2001). Des gènes dupliqués codent pour des protéines identiques qui vont interagir avec les mêmes partenaires. Ainsi, chaque protéine en contact avec une protéine dupliquée gagne un nouveau lien. De cette façon, les protéines déjà fortement connectées possèdent un avantage naturel. Bien qu'elles n'aient pas plus ou moins tendance à être dupliquées, en considérant que le phénomène de duplication intervient au hasard, elles ont une probabilité plus grande d'être liée à la protéine qui va être dupliquée. Un tel biais représente une version subtile du processus d'attachement préférentiel. Bien que le rôle de la duplication des gènes ne s'explique directement que pour les réseaux d'interactions protéine-protéine, il explique aussi probablement une topologie *scale-free* dans les réseaux métaboliques et réseaux de régulations transcriptionnelles. Bien que la duplication des gènes puisse expliquer l'apparition de la topologie *scale-free* dans les réseaux biologiques, il n'est pas impossible que d'autres processus biologiques soient également impliqués (Wagner, 2003).

Des résultats prouvent directement que la croissance est à la base de l'apparition de la topologie *scale-free* dans les réseaux biologiques. Par définition, le modèle *scale-free* implique que les nœuds apparus tôt dans l'histoire du réseau sont les nœuds les plus fortement connectés (Barabási & Albert, 1999). Or, il a été observé que les *hubs* métaboliques de la régulation de la transcription, de la glycolyse ou du cycle de Krebs (Wagner & Fell, 2001) sont les substrats les plus fortement connectés des réseaux métaboliques. De plus, des analyses comparatives inter-espèces ont montré que les protéines les plus anciennes sont plus fortement connectées que les protéines nouvelles (Wagner, 2003; Eisenberg & Levanon, 2003), validant ainsi également le processus d'attachement préférentiel.

L'approche d'inférence de réseaux développée : RulNet

La méthode d'inférence RulNet proposée dans le cadre de cette thèse est une approche intégrative et dirigée, basée sur des algorithmes de fouille de données (data mining) et plus particulièrement la découverte de règles entre attributs. La notion de règle est très répandue dans les domaines des bases de données, de la fouille de données et de l'intelligence artificielle. Des exemples connus sont les dépendances fonctionnelles, les implications ou les règles d'association. Une règle reflète une observation dans les données et malgré la diversité des domaines, la syntaxe reste inchangée, à savoir $X \rightarrow Y$ qui se lit « *X* implique *Y* ». Une règle signifie que lorsqu'une propriété, appelée *prédicat* est observée sur *X* (antécédent), alors un

prédicat est observé sur *Y* (conséquent)*, l*es prédicats pouvant être identiques ou non. Dans le cadre de l'inférence de réseaux de régulation, *X* et *Y* peuvent représenter différentes entités moléculaires : gènes, protéines, métabolites mais aussi certaines caractéristiques phénotypiques.

L'originalité première de la méthode développée dans ce travail est d'étendre la notion de règles en proposant plusieurs sémantiques pour les règles et d'utiliser des méthodes identiques de découverte, de post-traitement et de visualisation des règles pour toutes ces sémantiques. L'idée est qu'à chaque sémantique peut correspondre une signification biologique, il est donc possible d'analyser des données biologiques selon plusieurs perspectives.

La méthode initialement proposée par Agier *et al.* (2007) a été étendue d'une part par la mise en place d'un nouveau langage logique permettant de caractériser formellement les différentes sémantiques (Chardin *et al.*, 2013) et d'autre part dans le cadre de cette thèse, pour être adaptée aux problématiques de l'inférence de réseaux de régulations et aux données omiques. Ces travaux ont conjointement abouti à la mise en place d'une plateforme web dédiée, nommée RulNet.

En fonction des différents objectifs biologiques, il peut être utile de découvrir différents types de liens entre entités moléculaires et ne pas se restreindre à une seule signification biologique. L'idée de l'approche est alors d'offrir la possibilité à l'utilisateur de choisir la signification biologique des liens à inférer et d'inférer des règles à partir de données hétérogènes. L'originalité de l'approche est entre autre de permettre la superposition des liens inférés par les différentes sémantiques sur un même réseau global. L'application permet par ailleurs à l'utilisateur de spécifier des gènes « centraux » qu'il souhaite à priori voir apparaître dans les réseaux générés. Ceux-ci permettent de limiter la génération des règles aux seules règles impliquant les entités d'intérêt.

Le langage de requêtes nommé RQL qui a été développé (Chardin *et al.*, 2013) et étendu dans le cadre de cette thèse, permet de définir des types de règles déjà connus comme les règles d'association ou les dépendances fonctionnelles mais permet surtout de définir des sémantiques spécifiques aux jeux de données utilisés et particulièrement adaptées par exemple aux données omiques qualitatives et quantitatives.

Lorsqu'un utilisateur infère des réseaux sur la plateforme RulNet, celle-ci suit plusieurs étapes (Figure II-2). Tout d'abord, l'utilisateur spécifie où se trouvent les données à analyser, soit il

entre les paramètres pour que RulNet accède à une base de données, soit il spécifie un fichier texte classique. La seconde étape consiste à établir les sémantiques des règles à inférer. Cette étape peut se faire directement (rédaction de requêtes RQL) ou par le biais d'un assistant de création de requêtes. L'utilisateur peut définir ou non des entités d'intérêt pour la recherche de règles. Enfin il peut définir des seuils pour les mesures de qualités, permettant ainsi de ne sélectionner que les règles les plus pertinentes et de mettre un poids sur chaque lien du réseau. Il visualise ensuite le réseau obtenu directement sur la plateforme ou bien choisit de télécharger un fichier contenant toutes les règles inférées qu'il pourra ensuite visualiser via d'autres outils comme Cytoscape (Shannon *et al.*, 2003; Smoot *et al.*, 2011).

Ce travail permet de fournir un outil d'inférence de réseaux adapté à nos objectifs de recherche et utilisable par une large partie de la communauté scientifique. L'approche, ici appliquée à des données de blé peut être adaptée à de nombreuses études, que ce soit dans le domaine de la biologie comme dans des domaines plus éloignés. Le développement de cette plateforme s'inscrit dans l'approche biologie des systèmes qui a été choisie afin d'étudier l'influence de l'approvisionnement en N et S sur la synthèse des PR du blé tendre. Nous verrons dans la partie suivante comment l'approche RulNet peut s'intégrer à une telle étude.

CYCLE DE L'AZOTE ET DU SOUFRE DANS LE SOL

Cycle de l'azote

L'N est un élément essentiel des protéines qui forment les composants cellulaires et les tissues de la plante. Malgré le fait que l'atmosphère terrestre soit principalement constituée d'N (N_2 ; 78%), seule une catégorie de végétaux comme les légumineuses sont capables de l'assimiler sous cette forme et de le rendre disponible à la plante via un processus symbiotique avec des bactéries du genre *rhizobium* et la formation de nodules fixateurs d'N au niveau des racines. Les formes minérales d'N disponibles pour la plante sont le nitrate (NO_3^-), le nitrite (NO_2^-) et l'ammonium (NH_4^+).

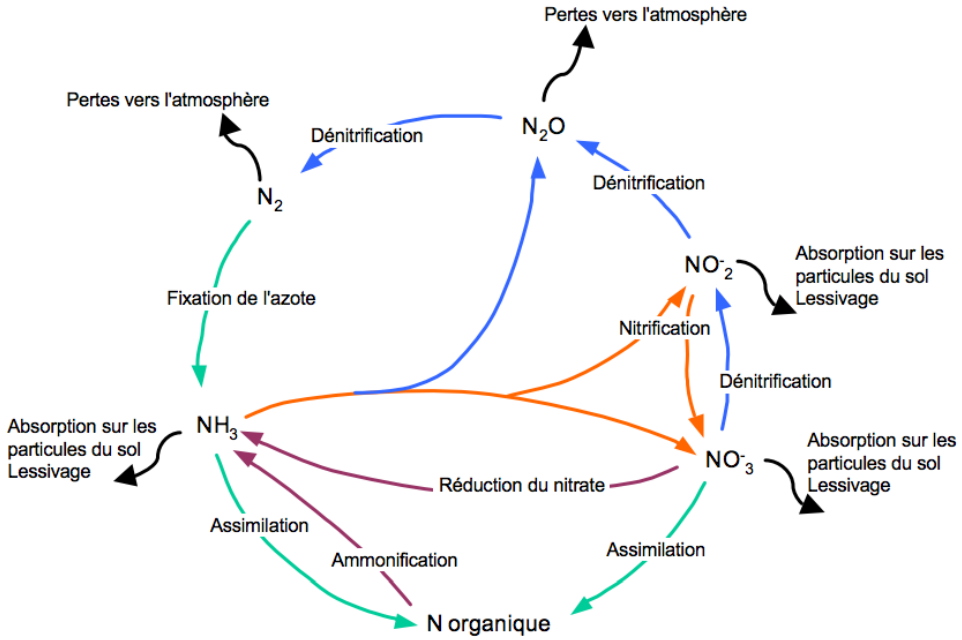


Figure 10 – Schéma du cycle de l'N dans les écosystèmes terrestres et aquatiques. *Flèches bleues : réactions biochimiques par lesquelles l'N a tendance à quitter le sol ou les eaux. Flèches vertes : processus par lesquels l'N est assimilé par la matière organique. Flèches rouges : réactions biochimiques par lesquelles l'N reste dans l'écosystème. Flèches noires : processus géochimiques par lesquels l'N quitte l'écosystème. D'après Baretta-Bekker et al. (1992).*

Ces différentes formes d'N (Figure 10) sont absorbées par les racines et les feuilles des plantes. Les taux d'absorption par les feuilles sont difficilement quantifiables, et dépendent du taux de pollution et des concentrations relatives de ces différentes formes d'N (captées plus ou moins efficacement selon la forme) dans l'air. Les feuilles peuvent également absorbées et assimiler l'urée qui peut être apporté aux cultures sous forme liquide par

pulvérisation. Dans les sols, ces trois formes d'N peuvent être adsorbées par des particules du sol ou lessivées et emportées. L'N réduit sous forme d'ammonium peut être nitrifié, et transformé en nitrate ou nitrite, libérant ainsi 2 protons qui vont participer l'acidification des sols. A l'inverse, l'N oxydé peut être dénitrifié et transformé en diazote ($N_{2)}$ ou protoxyde d'N (N_2O), un gaz à effet de serre. Les vitesses relatives de ces différentes réactions dépendent de nombreux facteurs comme la biodiversité, les taux de dépôts, le passé des sols et ses propriétés (teneur en matières organiques, compaction, humidité, température, pH etc.).

Cycle du soufre

De la même façon que pour l'N, le cycle du S est en partie dépendant des conditions environnementales (humidité, température, pH etc.) et de la biomasse microbienne. Le S sous ses formes organiques représente jusqu'à 99% du S total disponible dans les sols. Le profil de disponibilité en S dans le sol suit généralement celui de la concentration en matière organique (Probert, 1980). Les formes organiques de S peuvent être regroupées en deux classes, avec d'un côté l'atome de S sous sa forme oxydée et de l'autre l'atome de S sous sa forme réduite. Seulement 1,5 à 5% du S organique des sols est représenté par la biomasse microbienne (Wu *et al.*, 1993; Banerjee & Chapman, 1996).

Les protéines et les acides aminés représentent la plus grande partie du S de cette biomasse (Banerjee & Chapman, 1996). La biomasse microbienne est relativement instable, et son évolution dépend de la température et de l'humidité du sol ainsi que de sa compaction (Gupta & Germida, 1989). Néanmoins, la biomasse microbienne représente la principale source de renouvellement du S du sol. Généralement, l'application de matière organique sur les terres cultivables augmente la quantité de biomasse et parallèlement le S d'origine microbienne.

Le S sous forme minéral est donc généralement moins abondant dans les sols que le S organique. Il est disponible sous forme de dihydrogène de S (H_2S) provenant des éruptions volcaniques, des échanges de gaz avec l'océan et de la décomposition de matière organique. Il faut en outre prendre en compte les rejets industriels (SO_2 et SO_4^{2-}), générés par la combustion d'énergies fossiles. Le dioxyde de S est généralement oxydé en dioxyde de S (SO_2) puis en acide sulfurique (H_2SO_4) qui va être déposée sur les sols par les pluies. De même que pour les réactions de nitrification, l'acide sulfurique possède un fort pouvoir acidifiant. Le sulfate est la forme de S minéral majoritaire dans les sols. Le S peut précipiter sous forme de sulfate de calcium, de magnésium ou de sodium. Une quantité maximum de SO_4^{2-} peut être obtenue dans les couches supérieures des sols par application de fertilisants contenants du S. Les couches inférieures des sols ont une capacité supérieure à adsorber les sulfates (Barton

et al., 1999), sans doute à cause de la concentration plus forte en matière organique des couches supérieures. Il reste néanmoins difficile d'évaluer l'adsorption et la désorption du sulfate par les sols mais une saturation des sols peut être observée. Lorsque cette saturation est atteinte, le sulfate excédentaire est lessivé sous forme ionique, pouvant emporter avec lui d'autres éléments nutritifs vers les eaux de surface.

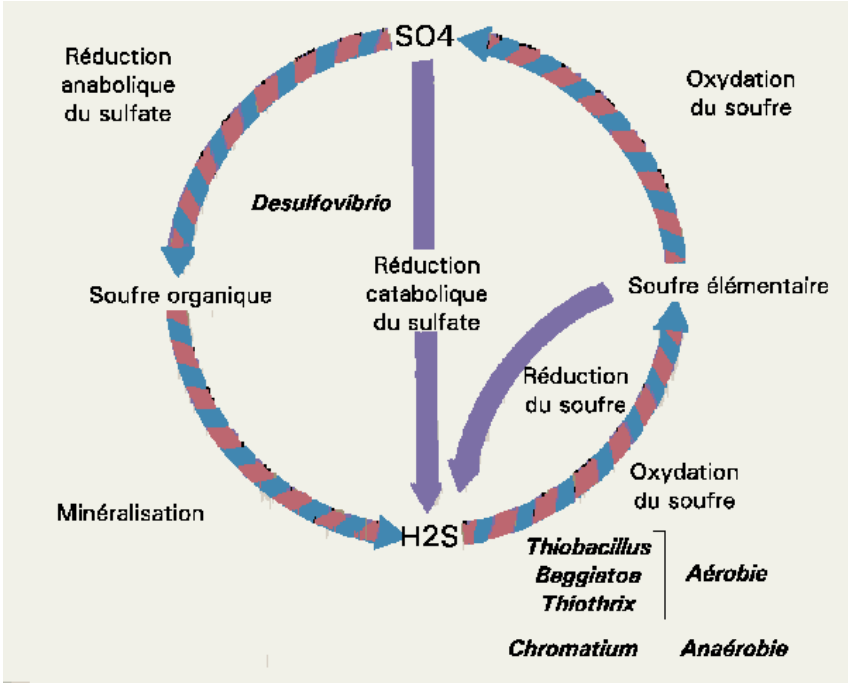


***Figure 11 – Schéma du cycle du S dans les écosystèmes terrestres et aquatiques**. Flèches violettes : émissions vers l'atmosphère. Flèches hachurées : réactions biochimiques. D'après Prescott et al. (2002).*

De la même façon que pour l'N, le S peut être absorbé par la plante, sous forme de dioxyde de S par les feuilles et sous forme de sulfate par les racines (Figure 11). Le S est indispensable à la synthèse des acides aminés soufrés (cystéine et méthionine). Un excès de S peut être toxique pour les plantes mais à l'heure actuelle et avec les mesures de réduction d'émissions, la toxicité du S est principalement indirecte et due à son pouvoir acidifiant. Tandis que la réduction importante des dépôts atmosphériques de S en Europe a permis de revenir à des valeurs préindustrielles (Eriksen, 2009), une carence en S peut être observée dans les sols de certaines zones agricoles. Pour la première fois dans l'histoire de l'agriculture, des symptômes de carence en S sont observés au niveau des cultures (Pedersen *et al.*, 1998). On observe donc depuis les années 1990 et en parallèle de la diminution des émissions de S, une augmentation

constante de l'utilisation d'engrais à base de S dans ces mêmes pays européens. La déficience en S des plantes de culture n'a été mise en évidence que récemment (Scherer, 2001). Des quantités de S suffisantes aux besoins des plantes de culture étaient généralement apportées de manière fortuite, lors de l'utilisation de fertilisants azotés par le biais du sulfate d'ammonium par exemple. La pollution industrielle fournissait également une quantité substantielle de S aux plantes de culture par le biais des retombées atmosphériques. L'augmentation des rendements (qui ont parfois doublé lors des deux dernières décennies) entraine un appauvrissement des sols en nutriments, notamment en S (Scherer, 2001). De plus, les fertilisants contenant du S ont progressivement été remplacés par des fertilisants à haute teneur en N, P. Au cours de la dernière décennie, l'augmente des surfaces de culture de colza (*Brassica napus* L. ; espèce de la famille des crucifères ayant de forts besoins en S) contribue également à diminuer les teneurs en S minéral dans les sols.

Assimilation et métabolismes de l'azote et du soufre dans la plante

Bien que généralement faible, la disponibilité du sol en N peut considérablement fluctuer dans l'espace et dans le temps en fonction de facteurs tels que les précipitations, la température, le vent, le type et le pH du sol. Par conséquent, la forme sous laquelle l'N est absorbé dépend de l'adaptation de la plante aux conditions du sol. En général, les plantes adaptées à des pH faibles et des sols réducteurs que l'on trouve dans des forêts anciennes ou en toundra arctique ont tendance à prélever l'ammonium ou des acides aminés. Au contraire, les plantes adaptées à des pH plus élevés et des sols plus en aérobie absorbent préférentiellement le nitrate. Les mécanismes d'absorption de l'N interviennent au niveau du système racinaire où des transporteurs de nitrate agissent de manière coordonnée pour absorber et distribuer le nitrate au sein de la plante (Daniel-Vedele *et al.*, 1998). Durant la croissance, l'N est transporté entre les organes par le biais d'un nombre restreint de mécanismes. Une partie de l'N organique est transportée via l'activité de *transaminases* et de *glutamine-amide transférases,* tandis qu'une autre partie est libérée sous forme de NH_3 qui est réassimilé par le biais de la *glutamine-synthétase* (Miflin & Habash, 2002). L'assimilation de nitrate a principalement lieu au niveau des racines de la plante et dépend fortement de l'âge et de la limitation de l'espace disponible pour la croissance des racines (Marquez *et al.,* 2007). Le nitrate absorbé par les plantes est réduit en nitrite par des *nitrate réductase* (Rosales *et al.*, 2011). Etant donné que le nitrite est très réactif, les cellules végétales le transportent immédiatement du cytosol vers les chloroplastes des feuilles et les plastes des racines. C'est dans ces organelles que le nitrite est alors réduit en NH_4 par la *nitrite réductase* (Rosales *et al.*, 2011).

Chez les plantes, l'ammonium est originaire de la réduction des nitrates, de l'absorption directe, de la photo-respiration, de la fixation de l'N gazeux (N_2) ou la désamination de composés azotés tels que l'asparagine (Figure 12 ; Wickert *et al.*, 2007). L'N minéral est d'abord réduit en ammonium, qui est la forme qui entre les voies de biosynthèse des acides aminés (Ruiz-capillas & Moral, 2002). L'ammoniaque issu de groupes *amine* est ensuite assimilée à la glutamine et au glutamate, qui servent à déplacer l'N organique des sources vers les puits. L'ammoniaque peut également être incorporée dans les acides aminés en utilisant les squelettes carbonés produits par d'autres voies métaboliques, telles que la photo-respiration et la photosynthèse. L'ensemble de ces réactions catalysées par la *nitrate réductase*, sont mieux comprises en considérant cette enzyme comme un système redox avec une chaîne interne de transfert d'électrons (Campbell, 1999). L'urée étant l'engrais azoté le plus utilisé dans l'agriculture, elle joue un rôle en tant que source d'N primaire; elle est absorbée activement par les plantes au niveau des feuilles et des racines et est également un intermédiaire du catabolisme de l'arginine, responsable de la remobilisation de l'N à partir des tissus sources (Witte, 2011).

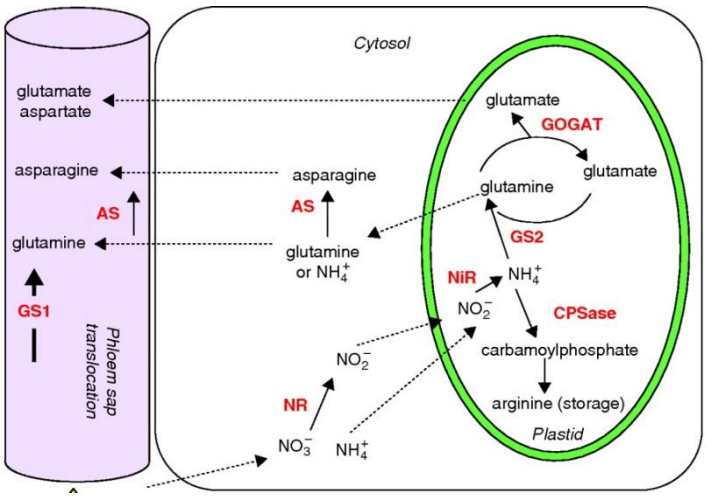


Figure 12 – Représentation schématique des enzymes clés impliquées dans l'assimilation de l'azote dans les jeunes feuilles. *NR : Nitrate Réductase (NR) ; AS : Asparagine Synthétase ; NIR : Nitrite Réductase ; GS1 et GS2 : Glutamine Synthétases ; GOGAT : Glutamate Synthase ; CPSase : Carbamoylphosphate Synthétase.D'après Masclaux-Daubresse et al. (2010).*

Le S est assimilé par la plante au niveau du système racinaire sous forme de sulfate. Il est alors transporté dans la plante majoritairement sous forme de sulfate mais aussi sous forme de cystéine qui est incorporée au glutathion. Son transport du système racinaire ou il est absorbé jusqu'à son assimilation dans les chloroplastes des feuilles ou des grains sous forme de protéines implique des enzymes de transport comme des transporteurs de sulfate et des transporteurs tans-membranaires. De plus le S en excès peut également être transitoirement stocké dans les vacuoles des cellules racinaires et foliaires sous forme de sulfate (Kataoka et al., 2004). Le S stocké sous forme de sulfate ou sous forme organique peut alors être remobilisé en cas de besoin. Les transporteurs de S sont exprimés différentièllement dans les différents organes de la plante et au cours de son développement et sont responsables des différentes étapes du transport du S dans la plante. La Figure 13 illustre les principales régulations du métabolisme du S dans la plante et les interactions avec les métabolismes du N dans la plante tandis que la figure 14 illustre le transport du S sous forme de sulfate dans la plante.

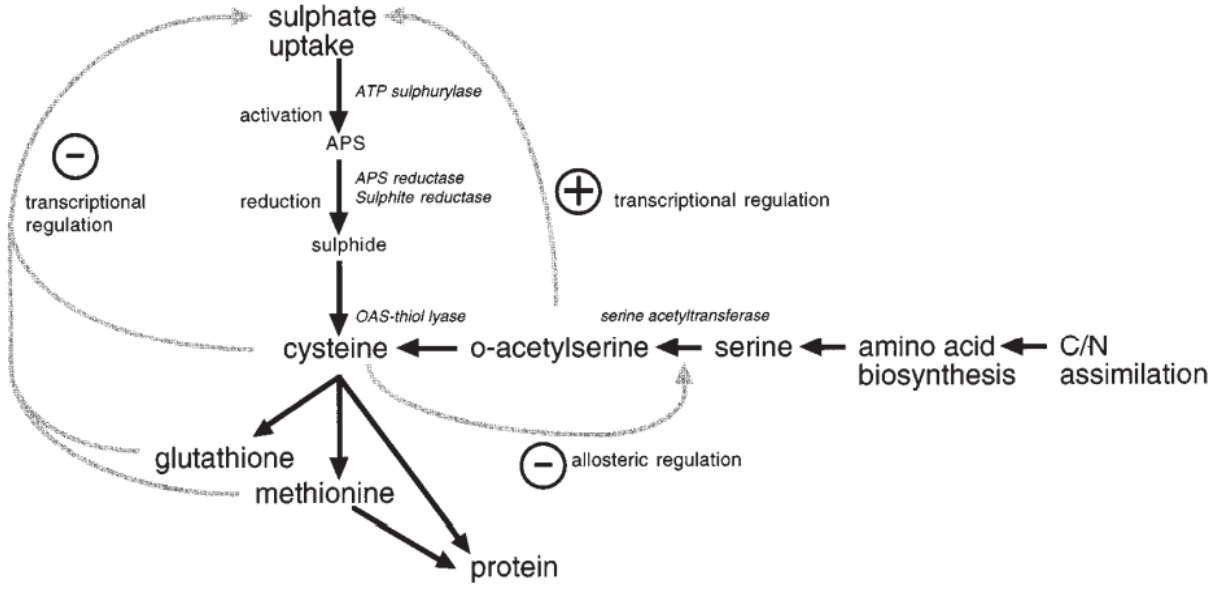


Figure 13 – Schéma du métabolisme du S chez les plantes. *Une combinaison de boucles de rétrocontrôles transcriptionnelles et la régulation allostérique de la voie d'assimilation du S entrainent une coordination des métabolismes du S et de l'N. Les transporteurs de S, l'ATP phosphorylase et l'APS reductase (adenosine-5'-phosphosulphate reductase) sont négativement régulées par le sulfate et les composés organiques soufrés. L'O-acétylsérine qui s'accumule lorsque le S est limitant, régule de manière positive la transcription des protéines impliquées dans l'absorption et l'assimilation du sulfate. Une rétro-inhibition allostérique négative de la synthèse de l'O-acétylsérine empêche l'accumulation de la cystéine potentiellement toxique. D'après Hawkesford (2000)*

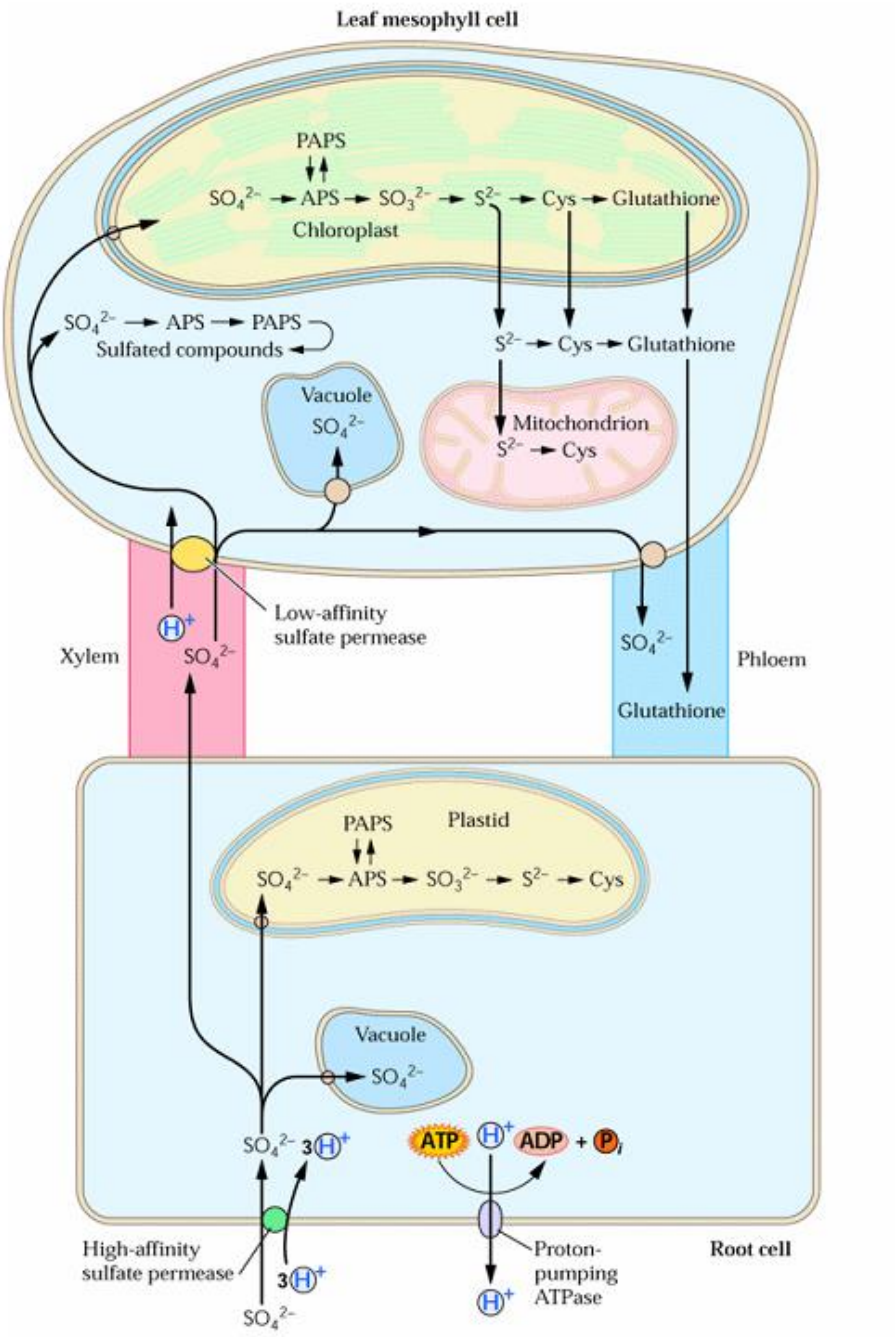


Figure 14 – Schéma du transport du sulfate dans la plante. *Comme pour le nitrate, l'absorption du sulfate à travers la membrane plasmique est entraînée par un gradient électrochimique qui est maintenue par une ATPase. Le sulfate peut être stocké dans les vacuoles. La réduction de sulfate et son assimilation sous forme de cystéine a lieu dans les plastes des cellules racinaires et foliaires. D'après Buchanan et al. (2000).*

Interaction entre l'azote et le soufre dans la plante

En raison des rôles centraux de l'N et du S dans la synthèse des protéines, les métabolismes de ces nutriments dans les plantes sont étroitement liés. Des interactions entre les métabolismes de l'N et du S ont été mises en évidences dans de nombreuses études (Zhao *et*

al., 1993; McGrath & Zhao, 1996). Un approvisionnement pauvre en S réduit l'utilisation du N par la plante (Lakkineni & Abrol, 1992), augmentant ainsi le lessivage des nitrates vers les écosystèmes aquatiques environnants.

Le S et l'N sont conjointement nécessaires à la synthèse des protéines, par conséquent, le rapport N:S dans les tissus de la plante peut refléter la capacité de la plante à synthétiser des protéines plus ou moins riches en S (Brunold & Suter, 1984). La carence en S entraine de profonds changements métaboliques, tels qu'une diminution de la synthèse de protéines et une accumulation de composés organiques et minéraux riches en N. Les teneurs observées en cystéine et méthionine chez le riz sont inférieures lors d'une carence en S, tandis que les métabolismes de l'asparagine et de l'arginine sont stimulés (Beaton, 1966). Une altération de l'activité *Nitrate Reductase* est observée lors d'une carence en nitrate mais également lors d'une carence en S dans les tissus, indiquant un rôle positif du S dans cette activité (Pal *et al.,* 1976). De même, un rôle de l'N a été mis en évidence (Smith, 1975) dans la régulation de l'assimilation du sulfate, par le biais de la régulation de l'activité de l'*ATP-sulphurylase*. La synthèse de la cystéine résultant de l'incorporation d'un groupement sulfure à l'*O-acétylsérine* semble être un point de rencontre entre les métabolismes de l'N et du S (Hesse *et al.*, 2004).

Remobilisation de l'azote et du soufre vers le grain

Remobilisation de l'azote vers le grain

Les protéines des feuilles et en particulier les protéines photosynthétiques de plastes sont dégradées au cours de la sénescence, offrant une source d'N que la plante peut utiliser pour compléter l'alimentation des organes en croissance tels que les nouvelles feuilles et les graines. La remobilisation de l'N vers le grain semble généralement être régulée par la source (Martre *et al.*, 2003). Chez le blé, la cinétique de la teneur en Rubisco (*ribulose-1,5-bisphosphate carboxylase/oxygénase* ; Irving & Robinson, 2006) et l'accumulation de l'N dans le grain (Bertheloot *et al.*, 2008) suggèrent que pendant le remplissage du grain, la translocation de l'N depuis organes végétatifs est principalement limitée par la disponibilité du substrat dans les organes sources.

La remobilisation de l'N a été étudiée dans plusieurs espèces végétales par la méthode de «remobilisation apparente». Cette méthode consiste à déterminer la quantité d'N totale présente dans les différents organes de la plante, à différents moments du développement, à l'aide d'un marquage au ^{15}N, ce qui permet l'étude des flux d'N (Gallais *et al.*, 2006 ; Schiltz *et al.*, 2005). Chez *A. thaliana* et le colza, il a été montré que l'N peut être remobilisé pendant

le stade végétatif, des feuilles sénescentes vers les feuilles en expansion, ou encore, pendant le stade reproductif des feuilles sénescentes vers les grains (Malagoli *et al.*, 2005). Chez le pois, Le remplissage des graines est une phase critique car l'absorption d'N et la fixation de N_2 diminuent au cours de la maturation de la plante et du remplissage des graines (Salon *et al.*, 2001) et les feuilles, les tiges, les épis, et les enveloppes des gousses servent successivement de sources d'N (Cliquet *et al.*, 1990). L'absorption et l'assimilation d'N pendant la phase de remplissage de la graine sont généralement insuffisantes pour satisfaire la forte demande de la graine en N et la remobilisation progressive qui se produit successivement dans les différents organes de la plante est nécessaire pour acheminer l'N vers la graine. La contribution du N des feuilles dans la remobilisation varie ostensiblement entre les espèces. Chez les céréales elle représente 50 à 90% de l'N des grains à maturité. Au cours de la phase linéaire de remplissage du grain la sénescence foliaire dépend principalement de la remobilisation de l'N vers les grains (Sinclair & de Wit, 1975; Martre *et al.*, 2006). Par ailleurs, Bertheloot *et al*. (2008) ont pu montrer, qu'en absence d'absorption d'N post-floraison, la remobilisation de l'N des différents organes végétatifs vers le grain suit une cinétique d'ordre un. Ces résultats vont également dans le sens d'un contrôle par les sources du flux d'N vers le grain. Des résultats expérimentaux suggèrent que le contrôle du flux d'N vers les grains par les sources s'exerce au niveau de l'accumulation des PR, qui représentent environ 60% du pool total des protéines du grain ; à l'inverse la synthèse des protéines métaboliques (environ 30% du total), dont le métabolisme est relativement indépendante de la disponibilité en N.

Remobilisation du soufre vers le grain

Chez les céréales l'N et le S requis pour la synthèse des PR du grain doivent être recrutés dans l'apoplasme de la cavité de l'albumen et dans le sac embryonnaire. L'apoplasme de la cavité de l'albumen est alimenté en métabolites par le phloème. De même, le grain en développement doit disposer de mécanismes de recrutement de composés soufrés et azotés depuis apoplasme. Les composés de l'apoplasme sont soumis aux évènements suivants: (1) synthèse au niveau d'un site source puis transfert vers le xylème ou le phloème; (2) transfert de métabolites d'abord transférés vers le xylème dans le phloème; (3) transfert du phloème vers l'apoplasme; (4) transfert de l'apoplasme vers le grain; (5) transformations biochimiques dans le grain pour former des acides aminés et (6) incorporation des acides aminés dans les PR du grain.

Les principales formes de S recrutées par le grain depuis l'apoplasme sont le glutathion (et/ou ses dérivés) et le sulfate. Cela implique l'existence de mécanismes d'absorption de ces

métabolites. Le transport transmembranaire de glutathion dans les plantes a fait l'objet de peu d'attention, mais des résultats suggèrent que les différentes membranes cellulaires diffèrent dans leur perméabilité à ce métabolite. Schneider *et al.* (1994) ont rapporté un système de basse et de haute affinité pour l'absorption de glutathion réduit (GSH) dans des cellules de tabac en culture. Cependant, Jamai *et al.* (1996) ont rapporté que la cinétique de l'absorption des formes oxydée (GSSG) et réduite de glutathion par les cellules du mésophile de haricot sont en accord avec un mécanisme de co-transport de protons capable de le transporter sous les deux formes mais qui est plus actif pour le GSSG.

Comme pour le transport du glutathion, peu d'études ont porté sur le transport du sulfate vers les grains en développement. Cependant, plusieurs transporteurs de sulfate ont été caractérisés chez les plantes et leur activité en relation avec la disponibilité en N et en S a été étudiée. Par exemple l'absorption sulfate chez *Stylosanthes hamata* est renforcée en condition de carence en S (Hawkesford, 2000). Ceci est associé à une augmentation de l'expression de deux transporteurs de sulfate de haute affinité (*SHST1* et *SHST2*). La plupart des organes de la plante sont impliqués à un certain stade de développement dans le chargement et/ou l'absorption du sulfate du phloème. Cela soulève la question de savoir si de tels transporteurs pourraient être impliqués dans le transfert du sulfate des compartiments apoplastiques internes jusqu'au grain en développement.

La redistribution du S des tissus végétatifs vers le grain est moins importante pour le S que pour l'N. Cette observation se reflète sur le *l'indice de récolte* (proportion d'un nutriment dans le grain par rapport aux tissus foliacés) qui est de 0.4 à 0.5 pour le S chez le blé contre 0.7 à 0.8 pour l'N. L'N remobilisé par le grain depuis les feuilles et la tige représente environ 70% du N total du grain tandis que cette proportion est inférieure à 50% pour le S (Hocking & Steer, 1983). Une étude plus récente (Monaghan *et al.*, 1999) montre que le S remobilisé dans le grain provient pour 14% de S absorbé au stade précoce d'extension de la tige (GS 12-32); pour 30% de S absorbé entre le stade précoce d'extension de la tige (GS 32-39); pour 6% de S absorbé entre le stade d'extension de la tige et l'anthèse (GS 39-69); et pour 50% de S absorbé après l'anthèse (GS 69-92). Les résultats de cette étude suggèrent donc que l'approvisionnement en S est important à la fois avant et après l'anthèse.

La majorité du S présent dans les grains matures est sous forme de glutathion, de cystéine et de méthionine. A la différence des tissus végétatifs dont le sulfate représente jusqu'à 50% de la teneur en S, le sulfate ne représente que 1 à 5% de la teneur en S du grain. Byers *et al.* (1979) ont étudié l'effet de la nutrition soufrée sur la composition en acides aminées du grain

de blé et montré qu'une carence en S diminuait la teneur en acides aminés soufrés (cystéine et méthionine) avec un effet plus important sur la cystéine que la méthionine. La déficience en S entraine également une augmentation de la teneur en asparagine et/ou en acide aspartique. La méthionine n'étant pas l'acide aminé le plus limitant, l'effet de la disponibilité en S sur la composition protéique du grain pourrait être moins important si la teneur en autres acides aminés n'était pas affectée. En outre, il a été mis en évidence que la carence en S diminuait également la teneur en lysine et en thréonine, particulièrement en condition de fort approvisionnement en N.

ACCUMULATION DES PROTEINES DE RESERVE ET NUTRITION AZOTEE ET SOUFREE

La germination du grain et la croissance de la plantule s'effectuent au dépend des composés stockés dans le grain jusqu'à ce que la plantule puisse subvenir à ses propres besoins en C et autres éléments par le biais de la photosynthèse et de l'absorption minérale au niveau du sol. Ainsi, une activité fondamentale du développement du grain est d'accumuler des éléments essentiels et de les incorporer à des composés de stockage. Quantitativement, le C et le plus important, les autres éléments comme l'N et le S sont requis en quantités inférieur mais ne sont pas moins indispensables.

La régulation de la synthèse des PR du grain chez les céréales dépend de nombreux facteurs tant génétiques qu'environnementaux. Il est aujourd'hui considéré que cette régulation s'effectue principalement au niveau transcriptionnel, et un certain nombre de facteurs cis- et trans- de cette régulation ont été décrits. L'interaction dans des réseaux de régulations transcriptionnelles de ces facteurs reste néanmoins peu connue. Nos connaissances sur les signaux moléculaires et environnementaux auxquels ces facteurs répondent restent limitées. Seule la régulation par la nutrition azotée des FT interagissant avec le motif GLM a été mise en évidence *in vitro* (Müller et Knudsen, 1993).

L'analyse des protéines, métabolites et transcrits à l'échelle du génome (approche –omique) permet aujourd'hui d'étudier ce type de réseaux de régulation de manière globale (approches systémiques) via la mise en œuvre d'outil mathématiques et informatiques d'inférence de réseaux de régulation. C'est ce type d'approche que nous avons mis en place.

Développement du grain et accumulation des PR chez le blé

L'art de la panification a été établi par l'homme il y a environ 4 000 ans et le pain est depuis devenu un aliment de base de l'alimentation dans de nombreuses cultures. La farine issue de

la mouture de grains de blé possède des caractéristiques de cuisson nettement supérieures à celles d'autres céréales comme l'orge ou le seigle (*Secale cereale* L. Dewettinck *et al.*, 2008).

Les PR s'accumulent de manière spécifique dans l'albumen du grain, à maturité elles représentent 60% à 80% des protéines totales du grain. L'accumulation des PR est en fait essentiellement déterminée par la disponibilité en N, C et S au cours de la phase de remplissage du grain (Shewry *et al.*, 2001; Triboï *et al.*, 2003a; Weichert *et al.*, 2010). La diversité des utilisations du grain découle des propriétés rhéologiques uniques de ces PR qui mélangées à l'eau forme un réseau viscoélastique : le gluten. Les propriétés de ce réseau qui sont exploitées dans la plupart de ses utilisations industrielles. Les propriétés technologiques du gluten sont déterminées d'une part par le type des PR (caractéristique des variétés ; Branlard *et al.*, 2001) leur composition et l'état d'agrégation **(**déterminés principalement par les conditions agro-climatiques ; Lafiandra *et al.*, 1999), d'autre part. L'étude scientifique des protéines du grain remonte à plus de 250 ans, avec les premières protéines du gluten isolées en 1745 par Beccari. Classiquement, on distingue trois classes de protéines chez le blé : (1) les albumines-globulines, protéines métaboliques et structurales solubles dans un tampon salin ; (2) les gliadines, PR monomériques solubles dans une solution alcoolique diluée ; et (3) les gluténines, PR polymériques solubles dans des solutions basiques diluées et le sodium dodecyl sulfate. L'analyse des PR en électrophorèse monodimensionnelle permet de distinguer quatre classes de gliadines (α/β, γ, ω1,2 et ω5) en fonction de leurs séquences d'acides aminés et leurs poids moléculaire (Wieser, 2007). On distingue également deux types de sous-unités de gluténines, de haute (HMW-GS) et faible (LMW-GS) masse moléculaire (Shewry & Halford, 2002). Les gluténines jouent un rôle dans l'acquisition des propriétés élastiques de la pâte, tandis que les gliadines contribuent à sa viscosité en lui conférant un caractère extensible. Les loci qui codent pour ces protéines ont été très largement étudiés. Des relations directes entre la composition allélique aux loci des PR et les propriétés fonctionnelles de la farine ont été décrites en détail (e.g. Branlard *et al.*, 2001; Eagles *et al.*, 2002; Békés *et al.*, 2006).

Ces différentes PR ont des teneurs en acides aminés soufrés très différentes et peuvent être classées en deux classes : les protéines riches et les protéines pauvres en S (Table 4). Cette classification est importante du fait du rôle des acides aminés soufrés (en particulier la cystéine qui permet la formation de ponts disulfures intermoléculaires) sur les propriétés technologiques des PR. Au cours de la germination du grain, elles fournissent les acides aminés essentiels au développement de la plantule. Les PR sont le produit de gènes de structure. Les gènes des HMW-GS ont été étudiés en détails (Shewry & Halford, 2002). Ils sont

portés par le groupe de chromosome 1 au locus *Glu-1*, chaque locus code pour 2 gènes de HMW-GS appelés x et y. L'organisation des gènes et loci codant pour les autres PR est beaucoup plus complexe et le nombre de gènes codant à chacun de ces loci n'est pas connu avec précision (Shewry *et al.*, 2003).

Table 4 – Composition en S et rapport N:S pour les différentes classes de PR. D'après Verlag et al. (2001).

Protéine	Acides aminés (nombre)			Rapport N:S	
	Total	Cystéine	Méthionine	mol/mol	Poids/poids
HMW-GS	842	4	3	170.1	74.4
LMW-GS	354	8	6	39.3	17.2
α/β-Gliadine	243	5	2	47.0	20.6
γ-Gliadine	292	9	6	27.3	11.9
ω-Gliadine	330	0	0	∞	∞

Les PR n'ont pas de fonction métabolique connue. Elles sont dégradées au moment de la germination et durant le développement précoce de la plantule par des protéases néo synthétisées ou activées au cours de la germination, puis utilisées comme source d'acides aminés (Wilson *et al.*, 1986). Les PR s'accumulent de manière spécifique dans l'albumen. Elles commencent à s'accumuler vers 250°Cj après floraison et jusqu'à la fin de la phase d'accumulation des réserves. Le stade 250°Cj apparait comme un point charnière dans le développement du grain car il correspond à la fin des divisions cellulaires dans l'albumen et au début de la synthèse des PR (Triboï *et al.*, 2003b). Toutes les PR ne s'accumulent pas au même moment lors de la phase de remplissage (Gupta *et al.*, 1996), en conséquence la composition en protéines du grain évolue au cours de cette phase. L'accumulation des PR est relativement peu affectée par la température (Daniel & Triboï, 2001) et la disponibilité en eau (Panozzo *et al.*, 2001) au cours du développement du grain. Ces facteurs de l'environnement modifient par contre significativement l'agrégation des PR au cours de la phase de dessiccation du grain. L'accumulation des PR est en fait essentiellement déterminée par la disponibilité en N et S au cours de la phase de remplissage du grain (Triboï *et al.*, 2003b; Wieser *et al.*, 2004).

AZOTE, SOUFRE ET COMPOSITION EN PROTEINES DE RESERVE

Il existe des relations allométriques entre la quantité de chaque classe de PR et la quantité d'N totale par grain (Figure 15), suggérant que leur synthèse est finement régulée par des

réseaux de régulation métaboliques et géniques (Triboï *et al.*, 2003b; Ravel *et al.*, 2009). Un tel réseau a d'ailleurs été élucidé chez l'orge (Rubio-Somoza *et al.*, 2006a,b). La teneur en S et le rapport N:S du grain sont des indicateurs de la distribution des macropolymères du gluten qui gouvernent les propriétés rhéologiques de la pâte (Zörb *et al.*, 2009).

La teneur et la composition en PR dans les grains matures sont particulièrement affectées par la disponibilité en N et S. L'N est le principal déterminant de la teneur en PR puisqu'il existe une corrélation positive entre la disponibilité en N et la teneur en PR (Shewry *et al.*, 2001; Triboï *et al.*, 2003b; Chope *et al.*, 2014). D'un autre côté, la carence en S aboutit à une diminution de la teneur en PR riches en S, tandis que par un effet de compensation, la teneur en PR pauvres en S augmente, n'influant ainsi pas sur la teneur totale en PR du grain (Higashi *et al.*, 2006; Zörb *et al.*, 2010). De plus, un rapport de disponibilité N:S trop élevé induit une réponse du grain similaire à une carence en S (Wieser & Seilmeier, 1998).

Malgré l'importance du rapport N:S sur la valeur d'utilisation du grain, nous disposons de peu de connaissances sur les mécanismes moléculaires sous-jacents à cette régulation. Au niveau de la régulation transcriptionnelle, l'apport d'N après anthèse aboutit à une augmentation des teneurs en ω-, α/β- gliadines et HMW-GS tandis que les teneurs en γ-gliadines et LMW-GS sont peu affectées (DuPont *et al.*, 2006; Altenbach & Kothari, 2007; Wan *et al.*, 2014). Il est admis aujourd'hui que l'N agit au moins en partie, au niveau transcriptionnel, dans une réponse relayée par des éléments de régulation présents dans les promoteurs des PR (Müller & Knudsen, 1993).

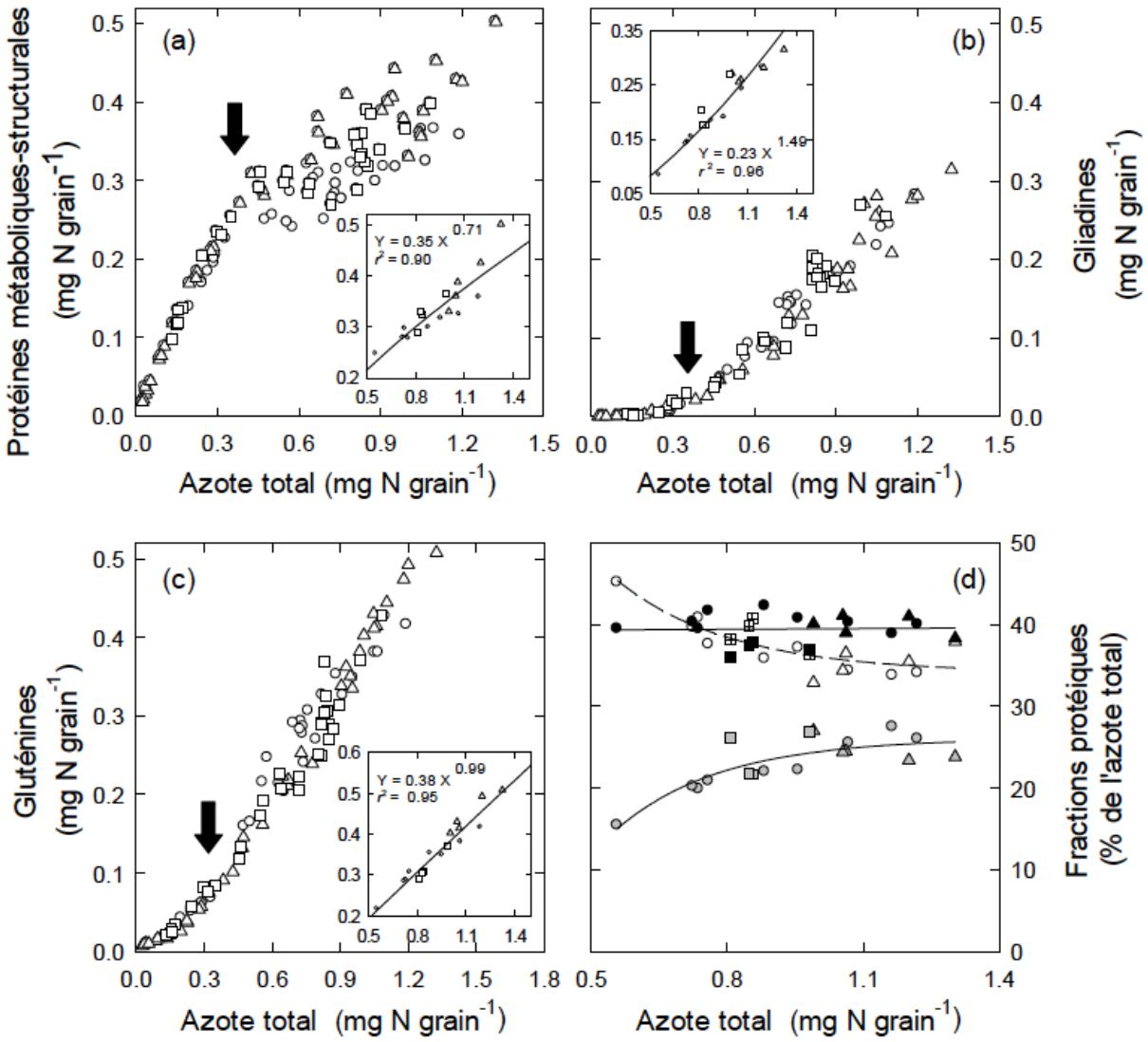


Figure 15 - Quantité de protéines métaboliques-structurales (a), gliadines (b), et gluténines (c) en fonction de la quantité totale d'N par grain. *Grains de blé tendre en cours de développement (cv. Thésée) cultivés au champ à Clermont-Ferrand à différents niveaux de nutrition azotée (ronds) et en conditions semi-contrôlées avec différentes températures (triangles) et régimes hydriques (carrés). Les inserts montrent les relations allométriques d'allocation pour les grains matures. (d) Pourcentage de protéines métaboliques-structurales (symboles ouverts), gliadines (symboles grisés), et gluténines (symboles fermés) en fonction de la quantité totale d'N par grain. Adaptée de Triboi et al. (2003).*

Peu de choses sont en revanches connues quant à la régulation transcriptionnelle des PR en fonction de la disponibilité en S. Chez le soja par exemple, une carence en S peut aboutir à une augmentation de la teneur en certaines PR pauvres en S par le biais d'une régulation transcriptionnelle (Hirai *et al.*, 1995). Chez *A. thaliana* en revanche, bien que l'expression des gènes codant pour des PR pauvres en S soit activée lors d'une carence en S, la composition finale du grain ne semble pas affectée, suggérant l'intervention de mécanismes de régulation

post-transcriptionnels ou post-traductionnels (Higashi *et al.*, 2006). Ce manque d'information est encore plus flagrant chez le blé, qui a pourtant un caractère d'intérêt majeur. Des études ont mis en évidence que la carence en S induit des changements d'expression de gènes impliqués dans le transport et l'assimilation du S (Buchner *et al.*, 2010; Ciaffi *et al.*, 2013). Chez le blé, un seul gène, nommé *sdi1* pour *S-deficiency induced* a été fonctionellement caractérisé (Howarth *et al.*, 2009). Il reste donc encore tout à découvrir concernant les mécanismes moléculaires impliqués dans la régulation de la composition des PR chez le grain de blé.

Contrôle transcriptionnel de l'expression des gènes des protéines de réserve

Chez les organismes eucaryotes, la régulation de l'expression des gènes se déroule à travers des mécanismes qui sont fondamentalement différents de ceux des procaryotes, ce qui explique à la fois le grand nombre et la diversité des protéines qui sont impliquées dans ce processus. Le processus de régulation de l'expression des gènes peut être régulé finement afin de faciliter la diversité des niveaux d'expression des gènes qui est nécessaire à la complexité biologique (Struhl, 1999). Chez un procaryote comme *E. coli,* la régulation de l'expression des gènes est non-restrictive, c'est à dire que l'accès de la *RNA polymérase* à l'ADN n'est pas limité. En outre, il a été estimé que la structure globale du réseau de régulation des gènes d'*E. coli* possède une faible complexité. Chez *E. coli*, un FT serait en moyenne impliqué dans la régulation de trois gènes tandis qu'un gène serait en moyenne sous le contrôle direct de deux FT (Thieffry *et al.*, 1998).

Chez les eucaryotes en revanche, la régulation de l'expression des gènes est restrictive, l'ADN est empaqueté dans la chromatine, qui bloque la reconnaissance des promoteurs par la machinerie de transcription (Struhl, 1999; Kornberg, 1999). La dynamique de la structure de la chromatine (plus ou moins condensée) est donc d'importance dans la régulation de la transcription. En outre, les régulateurs de la transcription chez les eucaryotes fonctionnent suivant une logique combinatoire (un moyen efficace d'augmenter le nombre et la diversité des activités de régulation des gènes) et la complexité des réseaux de régulation peut être grande. Cette logique combinatoire et synergique découle de la nature complexe des éléments cis- des promoteurs eucaryotes. La coexistence de plusieurs éléments cis-, souvent adjacent, entraine des interactions complexes entre les FT qui se lient à ces éléments. En outre, les interactions directes entre les FT ne sont pas le seul mécanisme moléculaire par lequel ils peuvent fonctionner de façon combinatoire pour réguler l'expression des gènes, car

ils peuvent également interagir avec d'autres composants de la machinerie transcriptionnelle ou même avec d'autres classes de protéines de régulation. C'est dans le cadre de cette logique de régulation combinatoire par les FT que nous avons mis en place une étude spécifique des séquences promotrices de certains gènes codants pour les PR (***publication 5***).

Globalement, les protéines impliquées dans la transcription chez les eucaryotes peuvent être classés en quatre groupes fonctionnels différents: (1) la machinerie transcriptionnelle de base et les facteurs intrinsèques associés (également connus en tant que facteurs de transcription généraux); (2) les co-activateurs et autres cofacteurs; (3) les FT spécifiques de domaines de liaison à l'ADN ; et (4) des protéines associées à la structure de la chromatine. Par contraste avec les composants de la machinerie transcriptionnelle de base, qui sont généralement hautement conservés, les co-régulateurs et FT ont divergé plus largement entre les eucaryotes.

L'expression des gènes est donc régulée par différents mécanismes, la régulation par les FT étant le plus étudié. Ces protéines de liaison à l'ADN contrôlent l'expression des gènes en interagissant avec des éléments cis-régulateurs dans les promoteurs de leurs gènes cibles. Les domaines de liaison à l'ADN sont en général fortement conservés entre les espèces de sorte que leurs caractéristiques sont utilisées pour classer les FT en familles (Figure 16). Le contrôle transcriptionnel de l'expression des gènes influe sur de nombreux processus biologiques. Les FT peuvent être classés en familles selon leur(s) domaine(s) de liaison à l'ADN. Chaque lignée eucaryote a inventé une fraction importante de ses propres FT. Les domaines de liaison d'ADN, qui sont conservés tant en terme de séquence que de structure, ont été réarrangés de différentes façons afin de créer de nouvelles protéines. Dans de nombreux cas, les familles qui sont communes aux trois règnes interviennent dans des processus nouveaux ou différents dans chacune des lignées.

Chez les plantes, la proportion des gènes codant pour des FT est d'environ 5 à 7% (Riechmann & Ratcliffe, 2000; Riechmann, 2000). Le séquençage des génomes végétaux modèles a permis d'identifier de nombreux gènes codant pour des FT. 1 533 FT ont par exemple été identifiés dans le génome d'*A. thaliana* (Riechmann, 2000) et classés en 34 familles. Environ 45% des FT d'*A. thaliana* appartiennent à des familles spécifiques des plantes (Riechmann, 2002). Chez le riz 1 611 FT appartenant à 37 familles ont été identifiés (Xiong *et al.*, 2005) La plupart des familles de FT sont de tailles similaires chez le riz et *A. thaliana* (Xiong *et al.*, 2005). Chez le blé, 3 820 FT ont été identifiés et classés en 40 familles et 84 sous-familles (Romeuf *et al*., 2010).

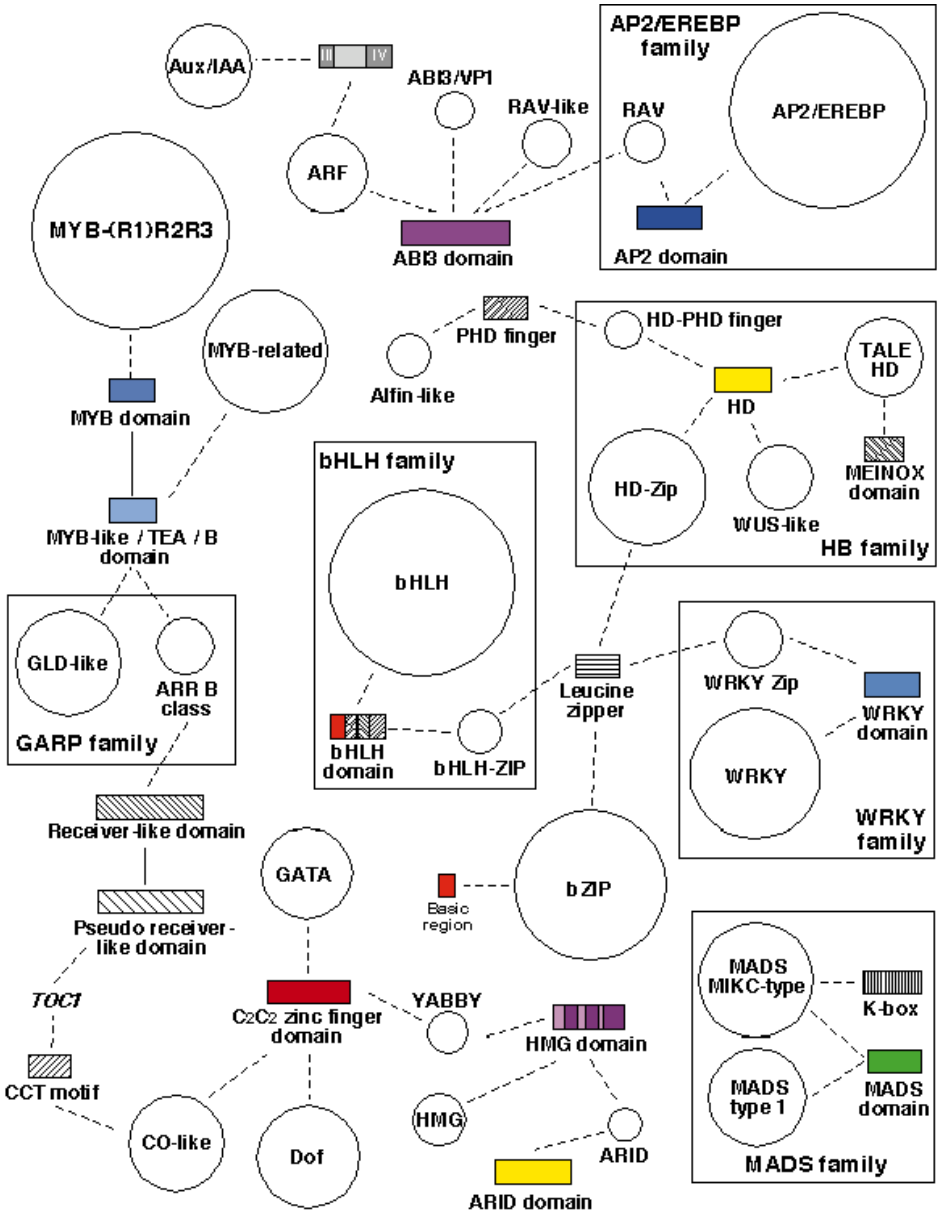


Figure 16 – Relations entre les différentes familles de FT identifiées chez A. thaliana. *Les familles sont représentées par des cercles dont la taille est proportionnelle au nombre de membres qu'elles comportent. Les domaines partagés entre les différentes familles et qui permettent donc de les relier sont représentés par des rectangles dont la taille est proportionnelle à la longueur du domaine. Les domaines de liaison à l'ADN sont colorés tandis que les autres domaines (domaines d'interaction protéine-protéine par exemple) sont représentés par des motifs hachurés. Les lignes en pointillés indiquent qu'un domaine est caractéristique d'une famille ou sous-famille. D'après Riechmann et al. (2000).*

Des analyses de transcription *in vitro* (Evans *et al.*, 1984; Walling, 1986; Nielsen & Dickinson, 1989) ont révélé que l'expression des PR est principalement contrôlée par la régulation transcriptionnelle au cours de la phase de remplissage du grain. L'analyse fonctionnelle de promoteurs de PR spécifiques de l'albumen chez l'orge et d'autres céréales a permis

d'identifier des cis-motifs conservés. Au moins six éléments cis conservés ont été identifiés au niveau des promoteurs. Le plus important est l'endosperm box (EB), un motif tri-partite constitué de l'élément GCN4-Like Motif (GLM), de la Prolamin Box (PB ; Colot *et al.*, 1987) et du motif AACA/TA ; le motif GARE (Gibberellic Acid Responsive Element) caractérisé par la séquence AACA (Gubler & Jacobsen, 1992; Zheng *et al.*, 1993) ; le motif GATA (Rubio-Somoza *et al*., 2006a) ; et le motif RY-element (Moreno-Risueno *et al.*, 2008) qui est un cis-motif impliqué dans diverses régulations au niveau du grain. Le motif GCNA semble être l'élément cis dominant mais des interactions synergétiques avec d'autres motifs est nécessaire pour une régulation positive par l'N. Un faible niveau d'N semble transformer le motif GCNA en motif répresseur (Müller & Knudsen, 1993).

Plusieurs FT ont été mis en évidence dans des interactions spécifiques avec ces séquences promotrices. Le premier de ces FT a été décrit chez le maïs, il s'agit d'*Opaque2* (Lohmer *et al.*, 1991), un FT de la famille des bZIP (basic leucine Zipper) se fixant de manière spécifique sur le motif GLM. Des orthologues à *Opaque2*, appelés *BLZ2* (Basic Leucine Zipper) et *SPA* (Storage Protein Activator) ont été mis en évidence chez l'orge et le blé, (Albani, 1997; Ravel *et al.,* 2009), respectivement. Des paralogues à *Opaque2* ont aussi été identifiés chez ces deux espèces : *BLZ1* chez l'orge (Vicente-Carbajosa *et al.*, 1998) et *SHP* chez le blé (Ravel *et al*., non publié). Deux FT de la famille DOF (DNA-binding with one finger) reconnaissent la PB chez le blé et l'orge, il s'agit de *PBF* (Prolamin box-Binding Factor ; Mena *et al.*, 1998) et de *SAD* (Scutellum et Aleurone-expressed DOF ; Isabel-LaMoneda *et al.*, 2003). Trois FT de la famille des MYB, *GAMYB* (Diaz *et al*., 2002), *MYBS3* (Rubio-Somoza *et al*., 2006a), et *MCB1* (Rubio-Somoza *et al*., 2006b), identifiés chez l'orge et le blé reconnaissent le motif GARE. Enfin, une interaction entre un FT de type B3 (*FUSCA3*), et le motif RY des PR a également été mise en évidence chez l'orge (Moreno-Risueno *et al.*, 2008). Les transcrits de *FUSCA3* s'accumulent dans l'albumen et atteignent leur maximum au milieu de la phase de maturation du grain (Moreno risueno 2008) et il interagit, au niveau de sa région N-terminale avec *BLZ2*. Ces résultats suggèrent qu'il existe un mécanisme de régulation de la synthèse des PR et que ce mécanisme est fortement conservé chez les angiospermes (Moreno-Risueno *et al.*, 2008). Des interactions bi- ou tri-partites ont été mises en évidence entre ces différents FT. Rubio-Somoza *et al.* (2006b) ont proposé un modèle d'interactions entre ces FT chez l'orge.

Des analyses transcriptionnelles de graines d'*A. thaliana* et d'orge (Ruuska *et al.*, 2002; Zhang *et al.*, 2004; Sreenivasulu *et al.*, 2006) révèlent des profils d'expression similaires pendant le développement du grain et ont permis d'identifier des FT co-régulés qui pourraient potentiellement contrôler des voies métaboliques spécifiques. Chez l'orge, plus de 1 200 FT

ont été identifiés, à partir de banques de données d'EST (Expressed Sequence Tags ; Zhang *et al.*, 2004), parmi eux, 214 sont différentiellement exprimés au cours du développement du grain (Sreenivasulu *et al.*, 2006). Dans cette étude 8 gènes codant pour des protéines impliquées dans le remodelage de la chromatine sont regroupées dans un cluster de co-expression avec des PR de l'orge, des inhibiteurs de protéases et 23 FT dont trois de la famille des NAC, deux de la famille des bHLH et des DOF. Chez le haricot, il a été constaté que la structure de la chromatine interagit avec la machinerie transcriptionnelle pour réguler l'expression des phaseolines (PR du haricot ; Li *et al.*, 2001).

Autres niveaux de régulation de la synthèse des protéines de réserve

D'autres facteurs déterminants dans la régulation du remplissage du grain ont été mis en évidence. La concentration de certains métabolites par exemple, comme les sucres, et plus précisément le rapport sucrose:hexose, qui agit comme un signal pour engager la phase de remplissage du grain (Weber *et al.*, 2005). La signalisation véhiculée par le sucrose contrôle l'accumulation des réserves et le processus de différenciation par la régulation de l'expression des gènes (Gibson, 2005; Rolland *et al.*, 2006). Les sucres et les acides aminés provenant des feuilles et de la tige ne servent pas seulement de substrat pour la synthèse de l'amidon et des protéines, ils interviennent aussi dans la régulation de l'expression des gènes et dans certaines voies métaboliques. Les sucres agissent comme des facteurs de croissance dans les graines (Finkelstein & Gibson, 2002 ; Rolland *et al*., 2006) bien que l'on ne connaisse pas encore précisément, leur rôle dans le développement de l'albumen. Un transporteur de sucrose, *HvSUT1* a été mis en évidence dans la régulation de l'expression des PR chez l'orge et le blé (Weichert *et al.*, 2010). La surexpression de ce transporteur augmente la capacité d'absorption de sucrose et les résultats suggèrent qu'un tel changement, par le biais d'un réseau de régulation impliquant la voie de biosynthèse des acides aminés et le métabolisme du carbone, conduit à une plus grande teneur en PR du grain. C'est une piste à considérer avec intérêt puisque l'on connait trois gènes de transporteurs de sucres qui sont exprimés dans le blé en cours de développement (Aoki *et al.*, 2002).

D'autres résultats suggèrent que les hormones jouent également un rôle. L'acide abscissique (ABA) est une hormone clé lors de la maturation (Finkelstein & Gibson, 2002), elle est nécessaire pour l'acquisition de tolérance à la dessiccation et l'entrée en dormance des graines. Son action dans ces processus est modulée par des effets antagonistes d'une deuxième hormone, l'acide gibbérellique (GA). Chez *A. thaliana*, le rôle central de la

signalisation par l'ABA se reflète dans ses interactions avec les masters régulateurs du développement du grain (*ABI3*, *FUS3*, *LEC2* et *LEC1*). Par exemple, les niveaux d'ABA corrèlent positivement avec l'activité de *FUS3* (Gazzarrini *et al.*, 2004) et l'activation de gènes par *FUS3* et *ABI3* est fortement renforcée par l'ABA (Parcy & Valon, 1994; Kagaya *et al.*, 2005). Cependant, le remplissage du grain n'est pas déterminé uniquement par l'ABA, des contrôles hormonaux complexes interviennent. Par exemple, l'expression anormale de *LEC2* chez *A. thaliana* active des enzymes de biosynthèse de l'auxine et des gènes de réponse à l'auxine (Stone *et al.*, 2008), ce qui suggère que son action est médiée au moins en partie par la synthèse de l'auxine. De même, son induction provoque la régulation d'un promoteur sensible aux cytokines (Stone *et al.*, 2008). *LEC1* agit également de concert avec l'auxine pour promouvoir la différenciation embryonnaire (Casson & Lindsey, 2006). L'expression de *FUS3* est positivement régulée par l'auxine, il agit alors comme un répresseur de la biosynthèse des GA (Gazzarrini *et al.*, 2004). En outre, Takaiwa *et al.*, (2009) ont indiqué que la stabilité de l'ARNm des PR, la régulation de leur traduction, le repliement des protéines, les modifications post-traductionnelles et le renouvellement des protéines influent également le niveau d'accumulation de PR du grain.

Dans une étude sur la lysine, qui est l'acide aminé essentiel le moins abondant chez les céréales, il a été mis en évidence chez le riz que l'expression d'une enzyme de dégradation de la lysine (*lysine ketoglutarate reductase/saccharopine dehydrogenase*, *ProOsLKR/SDH*) est directement régulée par les régulateurs transcriptionnels majeurs de la synthèse des PR, RISBZ1 et RPBF (Kawakatsu & Takaiwa, 2010). Des cis-motifs (prolamin box, GCN4 motif) similaires à ceux qui peuvent être trouvés dans les régions promotrices des gènes codant pour les PR sont présents au niveau du promoteur de ProOsLKR/SDH. Il semblerait que la même combinaison de FT et cis-motifs soit impliquée dans la régulation des PR et de ProOsLKR/SDH mais que le mécanisme de régulation soit différent. L'étude des liens fonctionnels qui peuvent exister entre les acides aminés et le métabolisme des PR doit encore être approfondie.

Globalement, les mécanismes de régulation de la synthèse des PR sont étroitement liés aux mécanismes de développement du grain, pour les étudier, il est nécessaire d'analyse les réseaux de régulation de plusieurs fois métaboliques simultanément. Nos connaissances actuelles suggèrent que les métabolismes de l'N et du S sont impliqué dans ces mécanismes mais aussi les métabolismes des acides aminés, des sucres et des hormones. L'étude des régulateurs qui agissent en amont de ces voies métaboliques permettra d'étendre nos connaissances. Une étape importante dans la caractérisation de ces mécanismes est l'identification des FT impliqués dans les voies de synthèses des acides aminés et des PR. Afin

d'obtenir une vision plus globale de ces mécanismes en rapport avec la nutrition et le rendement, il est important de les étudier au niveau de la plante entière (Götz *et al.*, 2007). Ce sont ces observations, entre autres, qui ont motivé l'approche mise en place dans le cadre de cette thèse. Les ***publications 3 et 4*** s'inscrivent directement dans cette approche systémique des mécanismes de régulation de la synthèse des PR, en partant de la plante entière, en donnant une importance prépondérante aux FT et en observant conjointement le métabolisme des acides aminés en fonction des disponibilité en N et en S.

CONCLUSION DE LA SYNTHESE ET STRATEGIE D'ETUDE

L'étude bibliographique nous a permis de mettre en évidence que la synthèse des PR est soumise à des régulations transcriptionnelles spatiales et temporelles auxquelles peuvent s'ajouter des régulations post-transcriptionnelles. De nombreuses études ont démontré que la synthèse des PR est essentiellement régulée au niveau transcriptionnel. Ces mécanismes de régulation transcriptionnelle des PR sont conservés, au moins en partie, chez les dicotylédones et monocotylédones (Fauteux & Strömvik, 2009). La régulation transcriptionnelle met en jeu des séquences généralement courtes et bien conservées au niveau des promoteurs des gènes (cis-motifs) sur lesquelles se fixent de manière spécifique des protéines régulatrices (éléments trans-), les FT. Différents FT concourent généralement à la régulation d'un même gène, permettant ainsi d'intégrer des signaux de régulation multiples.

Bien que la teneur et la composition en PR reposent sur un déterminisme génétique, il existe des facteurs environnementaux capables d'altérer largement ces variables déterminantes pour la valeur d'utilisation du grain. L'accumulation des PR est relativement peu affectée par la température et la disponibilité en eau (Panozzo *et al.*, 2001) au cours du développement du grain. Ces facteurs de l'environnement modifient par contre significativement l'agrégation des PR au cours de la phase de dessiccation du grain. L'accumulation des PR est en fait essentiellement déterminée par la disponibilité en N et S au cours de la phase de remplissage du grain. Triboi *et al*. (2003) ont mis en évidence des relations allométriques entre la quantité de chaque classe de PR et la quantité d'N (protéines) totale par grain. Ces relations suggèrent que la synthèse des PR est finement régulée par des réseaux de régulations métaboliques et géniques (Triboï *et al.*, 2003a; Aguirrezábal *et al.*, 2009; Ravel *et al.*, 2009; Plessis *et al.*, 2013). De plus, un certain nombre d'études mettent en évidence la disponibilité en S comme un facteur déterminant de la composition en PR et capable dans certaines conditions d'agir en synergie avec l'N (Zhao *et al.*, 1993; Wieser *et al.*, 2004; Higashi *et al.*, 2006; Zörb *et al.*, 2010).

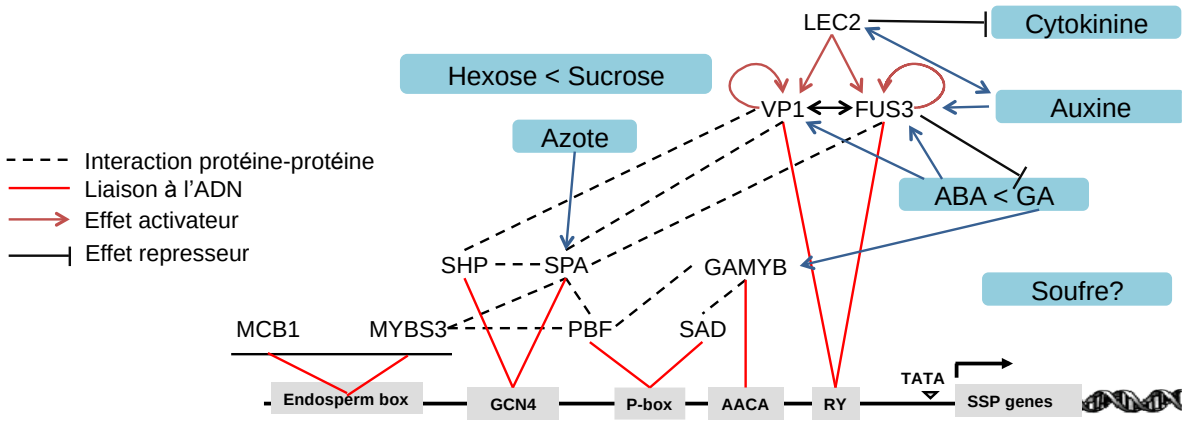


Figure 17 – Réseau hypothétique impliqué dans la régulation de la synthèse des protéines de réserve du grain chez les céréales.

L'étude bibliographique met donc en exergue le caractère systémique de la régulation de la synthèse des PR chez les céréales. Nous avons aussi constaté que les mécanismes moléculaires sous-jacents à ce système de régulation ont rarement été étudiés. Les études les plus détaillées concernent le réseau de régulation mis en évidence chez l'orge (Rubio-Somoza *et al.*, 2006a, b). Une review proposée par Verdier et Thompson (2008) permet d'exposer le caractère systémique de cette régulation chez les dicotylédones, avec notamment l'implication de différentes hormones (auxine, cytokinine, GA, ABA) ou métabolites (rapport hexose:sucrose). Une partie de nos connaissances sur le réseau de régulation de la synthèse des PR peut être synthétisée dans le réseau proposé en Figure 17. C'est ce réseau ainsi que les voies métaboliques qui l'entourent que nous avons exploré dans ce travail de thèse.

La plupart des études mettent en avant le caractère presque intégralement transcriptionnel dans la régulation de la synthèse des PR. C'est pour cette raison que des expériences transcriptomiques ont été mises en place, afin de suivre l'expression des gènes au cours du développement du grain et dans différentes conditions de disponibilité en N et en S. l'analyse des résultats de ces expériences de transcriptomique nécessite que l'on connaisse la fonction, au moins hypothétique d'un maximum de gènes. Or, le génome du blé n'est pas encore séquencé et la fonction d'une très grande majorité de ses gènes n'est pas encore identifiée. Afin de tirer le maximum d'information des données transcriptomiques, nous avons mis en place un outil d'annotation des séquences utilisées pour concevoir une puce à ADN développé précement laboratoire (Rustenholz *et al.*, 2011). L'outil, présenté dans la ***publication 1***,

possède de nombreuses fonctionnalités et nous a facilité l'annotation des gènes qui sont apparus dans les réseaux que nous avons inférés.

Une approche multi-échelles semble mieux adaptée afin d'inférer au mieux les réseaux impliqué dans la régulation de la synthèse des PR. Le caractère systémique de cette régulation nécessite une approche prenant en compte les niveaux de régulation transcriptomiques, protéomiques et métabolomiques. Ainsi, toujours dans un contexte méthodologique, nous nous sommes proposé d'étendre une approche d'inférence de réseaux (Agier *et al.*, 2007) ayant déjà été appliquée à des données biologiques. Pour répondre aux besoins de la biologie, nous nous sommes assurés que les réseaux d'interactions inférés possèdent bien une topologie *scale-free* représentative des réseaux biologiques complexes. Cette extension de l'approche permet d'avoir un contrôle maximum sur les sémantiques de recherche de liens. La plateforme et l'approche ainsi étendue sont présentées en détails dans la ***publication 2*****.**

Dans un second temps, nous avons mis en application les outils méthodologiques développés dans le cadre d'une étude transcriptomique du développement du grain de blé. Les résultats de cette étude sont présentés dans la ***publication 3***. Pour ce faire, nous avons tout d'abord mis à profit une expérience transcriptomiques de différents organes de la plante (Rustenholz *et al*., 2011) afin de dégager les spécificités du transcriptome du grain et identifier les FT spécifiquement exprimés dans le grain. La mise à profit d'une seconde expérience de transcriptomique (Romeuf, 2010) portant sur une cinétique de développement du grain nous a permis de mettre en évidence que la reprogrammation transcriptomique qui intervient entre les grandes phases de développement du grain est à l'origine portée par une reprogrammation de l'expression des FT. L'utilisation d'une troisième expérience transcriptomique portant sur le développement du grain dans différents niveaux de disponibilité en N et en S nous a permis d'observer l'influence du N et du S sur le transcriptome du grain.

La dernière publication présentée dans cette thèse (***publication 4***) est une étude systémique du développement du grain de blé dans différentes conditions d'approvisionnement en N et S. Des données transcriptomiques mais aussi protéomiques et métabolomiques ont été utilisées pour réaliser cette étude nous a permis d'étudier les régulations à l'origine des relations d'allométrie observées entre le niveau d'N total du grain et la composition en PR. Elle nous à également permis d'observer le comportement du métabolisme des acides aminés qui semble impliqué dans la régulation de la synthèse des PR. Cette étude enfin permis

d'observer l'effet sur le transcriptome, le protéome et le métabolome, d'un réapprovisionnement en N ou S au cours de la phase de remplissage du grain.

Dans sa globalité, cette thèse est surtout une étude exploratoire permettant d'évaluer l'efficacité d'une méthode d'inférence de réseaux pour étudier le caractère systémique de la régulation de la synthèse des PR. L'ordre des publications matérialise cette approche intégrative, avec tout d'abord une partie méthodologique sur l'annotation fonctionnelle (***publications 1***) et l'inférence de réseaux (***publications 2***). Une partie d'analyse transcriptomique tirant notamment partie des outils méthodologiques développés (***publication 3***). Enfin une analyse globale de la régulation de la synthèse des PR à plusieurs niveaux de régulation (***publication 4***). La ***publication 5*** qui est présentée en annexe est une analyse fonctionnelle des séquences promotrices des HMW-GS. A terme nous pourrions imaginer de combiner nos connaissances des FT et des séquences promotrices de ces PR afin d'étayer les interactions physiques qui peuvent intervenir entre les FT identifiés et les PR.

RÉFÉRENCES

Aceituno FF, Moseyko N, Rhee SY, Gutiérrez R a. **2008**. The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in Arabidopsis thaliana. *BMC genomics* **9**: 438.

Agier M. **2007**. Unifying Framework for Rule Semantics : Application to Gene Expression Data. **78**: 1–17.

Agier M, Petit J-M, Suzuki E. **2007**. Unifying Framework for Rule Semantics: Application to Gene Expression Data. *Fundam. Inf.* **78**: 543–559.

Aguirrezábal L, Martre P, Pereyra-Irujo G, Izquierdo N, Allard V. **2009**. Management and Breeding Strategies for the Improvement of Grain and Oil Quality. In: Sadras V, Calderini DBT-CP, eds. San Diego: Academic Press, 387–421.

Albani D. **1997**. The Wheat Transcriptional Activator SPA: A Seed-Specific bZIP Protein That Recognizes the GCN4-like Motif in the Bifactorial Endosperm Box of Prolamin Genes. *the Plant Cell Online* **9**: 171–184.

Altenbach SB, Kothari KM. **2007**. Omega gliadin genes expressed in Triticum aestivum cv. Butte 86: Effects of post-anthesis fertilizer on transcript accumulation during grain development. *Journal of Cereal Science* **46**: 169–177.

Aoki N, Whitfeld P, Hoeren F, Scofield G, Newell K, Patrick J, Offler C, Clarke B, Rahman S, Furbank R. **2002**. Three sucrose transporter genes are expressed in the developing grain of hexaploid wheat. *Plant Molecular Biology* **50**: 453–462.

Balasubramanian V, Alves B, Aulakh M, Bekunda M, Cai Z, Drinkwater L, Mugendi D, Van Kessel C, Oenema O. **2004**. Crop, Environmental and Management Factors Affecting Nitrogen Use Efficiency. In: Mosier A, Syers JK, Freney JR, eds. Agriculture and the nitrogen cycle: assessing the impacts of fertilizer use on food production and the environment. Washington, {DC}: Island Press, .

Balfourier F, Roussel V, Strelchenko P, Exbrayat-Vinson F, Sourdille P, Boutet G, Koenig J, Ravel C, Mitrofanova O, Beckert M, *et al.* 2007. A worldwide bread wheat core collection arrayed in a 384-well plate. *Theoretical and Applied Genetics* **114**: 1265–1275.

Banerjee MR, Chapman SJ. **1996**. The significance of microbial biomass sulphur in soil. *Biology and Fertility of Soils* **22**: 116–125.

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**: 101–13.

Baretta-Bekker HG, Duursma E, Kuipers B. **1992**. N. In: Baretta-Bekker HG, Duursma E, Kuipers B, eds. Encyclopedia of Marine Sciences SE - 14. Springer Berlin Heidelberg, 175–183.

Barton D, Gammack SM, Billett MF, Cresser MS. **1999**. Sulphate adsorption and acidification of Calluna heathland and Scots pine forest podzol soils in north-east Scotland. *Forest Ecology and Management* **114**: 151–164.

Bassel GW, Lan H, Glaab E, Gibbs DJ, Gerjets T, Krasnogor N, Bonner AJ, Holdsworth MJ, Provart NJ. **2011**. Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 9709–14.

Beaton JD. **1966**. SULFUR REQUIREMENTS OF CEREALS, TREE FRUITS, VEGETABLES, AND OTHER CROPS. *Soil Science* **101**.

Békés F, Kemény S, Morell M. **2006**. An integrated approach to predicting end-product quality of wheat. *European Journal of Agronomy* **25**: 155–162.

Bertalanffy L Von. **1969**. *General System Theory: Foundations, Development, Applications*. George Braziller, Inc.

Bertheloot J, Martre P, Andrieu B. **2008**. Dynamics of light and nitrogen distribution during grain filling within wheat canopy. *Plant physiology* **148**: 1707–20.

Bogard M, Jourdan M, Allard V, Martre P, Perretant MR, Ravel C, Heumez E, Orford S, Snape J, Griffiths S, *et al.* 2011. Anthesis date mainly explained correlations between post-anthesis leaf senescence, grain yield, and grain protein concentration in a winter wheat population segregating for flowering time QTLs. *Journal of experimental botany* **62**: 3621–36.

Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. **2006**. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology* **7**: R36.

Brancourt-Hulmel M, Doussinault G, Lecomte C, Bérard P, Le Buanec B, Trottet M. **2003**. Genetic Improvement of Agronomic Traits of Winter Wheat Cultivars Released in France from 1946 to 1992. : 37–45.

Branlard G, Dardevet M, Saccomano R, Lagoutte F, Gourdon J. **2001**. Genetic diversity of wheat storage proteins and bread wheat quality. : 59–67.

Brazma A, Vilo J, Ukkonen E, Valtonen K. **1997**. Data mining for regulatory elements in yeast genome. *Ismb*.

Brisson N, Gate P, Gouache D, Charmet G, Oury F-X, Huard F. **2010**. Why are wheat yields stagnating in Europe? A comprehensive data analysis for France. *Field Crops Research* **119**: 201–212.

Brunold C, Suter M. **1984**. Regulation of sulfate assimilation by nitrogen nutrition in the duckweed Lemna minor L. *Plant Physiology*: 579–583.

Buchanan B, Gruissem W, Jones R. **2000**. *Biochemistry & molecular biology of plants*.

Buchner P, Parmar S, Kriegel A, Carpentier M, Hawkesford MJ. **2010**. The Sulfate Transporter Family in Wheat: Tissue-Specific Gene Expression in Relation to Nutrition. *Molecular Plant* **3** : 374–389.

Campbell WH. **1999**. NITRATE REDUCTASE STRUCTURE, FUNCTION AND REGULATION: Bridging the Gap between Biochemistry and Physiology. *Annual review of plant physiology and plant molecular biology* **50**: 277–303.

Campbell CA, Selles F, Zentner RP, McConkey BG. **1993**. Available Water and Nitrogen Effects on Yield Components and Grain Nitrogen of Zero-Till Spring Wheat. : 114–120.

Cassman KG, Dobermann A, Walters DT, Yang H. **2003**. MEETING CEREAL DEMAND WHILE PROTECTING NATURAL RESOURCES AND IMPROVING ENVIRONMENTAL QUALITY. *Annual Review of Environment and Resources* **28**: 315–358.

Casson S, Lindsey K. **2006**. The turnip mutant of Arabidopsis reveals that LEAFY COTYLEDON1 expression mediates the effects of auxin and sugars to promote embryonic cell identity. *Plant physiology* **142**: 526–41.

Chardin B, Coquery E, Gouriou B. **2013**. Query Rewriting for Rule Mining in Databases. ….

Chope GA, Wan Y, Penson SP, Bhandari DG, Powers SJ, Shewry PR, Hawkesford MJ. **2014**. Effects of Genotype, Season, and Nitrogen Nutrition on Gene Expression and Protein Accumulation in Wheat Grain. *Journal of Agricultural and Food Chemistry* **62**: 4399–4407.

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier M-C, Magdelenat G, Gonthier C, *et al.* 2010. Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces. *The Plant Cell Online* **22** : 1686–1701.

Ciaffi M, Paolacci AR, Celletti S, Catarcione G, Kopriva S, Astolfi S. **2013**. Transcriptional and physiological changes in the S assimilation pathway due to single or combined S and Fe deprivation in durum wheat (Triticum durum L.) seedlings. *Journal of Experimental Botany* **64** : 1663–1675.

Cliquet J, Deléens E, Mariotti A. **1990**. C and N Mobilization from Stalk and Leaves during Kernel Filling by 13C and 15N Tracing in Zea mays L. *Plant Physiology* **94**: 1547–53.

Colot V, BEVAN M, THOMPSON R. **1987**. IDENTIFICATION OF DNA-SEQUENCES REQUIRED FOR SEED-SPECIFIC ACTIVITY OF A WHEAT STORAGE PROTEIN GENE PROMOTER IN TOBACCO. JOURNAL OF CELLULAR BIOCHEMISTRY.51.

Daniel C, Triboï E. **2001**. Effects of temperature and nitrogen nutrition on the accumulation of gliadins analysed by RP-HPLC. *Functional Plant Biology* **28**: 1197–1205.

Daniel-Vedele F, Filleur S, Caboche M. **1998**. Nitrate transport: a key step in nitrate assimilation. *Current Opinion in Plant Biology* **1**: 235–239.

Dewettinck K, Van Bockstaele F, Kühne B, Van de Walle D, Courtens TM, Gellynck X. **2008**. Nutritional value of bread: Influence of processing, food interaction and consumer perception. *Journal of Cereal Science* **48**: 243–257.

Diamond J. **2002**. Evolution, consequences and future of plant and animal domestication. *Nature* **418**: 700–7.

Dubcovsky J, Dvorak J. **2007a**. Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. *Science* **316** : 1862–1866.

Dubcovsky J, Dvorak J. **2007b**. Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. *Science* **316** : 1862–1866.

DuPont FM, Hurkman WJ, Vensel WH, Chan R, Lopez R, Tanaka CK, Altenbach SB. **2006**. Differential accumulation of sulfur-rich and sulfur-poor wheat flour proteins is affected by temperature and mineral nutrition during grain development. *Journal of Cereal Science* **44**: 101–112.

Durzinsky M, Wagler A, Marwan W. **2011**. Reconstruction of extended Petri nets from time series data and its application to signal transduction and to gene regulatory networks. *BMC systems biology* **5**: 113.

Eagles HA, Hollamby GJ, Gororo NN, Eastwood RF. **2002**. Estimation and utilisation of glutenin gene effects from the analysis of unbalanced data from wheat breeding programs. *Australian Journal of Agricultural Research* **53**: 367–377.

Eisenberg E, Levanon EY. **2003**. Preferential Attachment in the Protein Network Evolution. *Phys. Rev. Lett.* **91**: 138701.

Erdos P, Renyi A. **1960**. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **7**: 17.

Eriksen J. **2009**. Chapter 2 Soil Sulfur Cycling in Temperate Agricultural Systems. In: Agronomy DLSBT-A in, ed. Academic Press, 55–89.

Evans IM, Gatehouse J, Croy RD, Boulter D. **1984**. Regulation of the transcription of storage-protein mRNA in nuclei isolated from developing pea (Pisum sativum L.) cotyledons. *Planta* **160**: 559–568.

Fadda A, Fierro AC, Lemmens K, Monsieurs P, Engelen K, Marchal K. **2009**. Inferring the transcriptional network of Bacillus subtilis. *MOLECULAR BIOSYSTEMS* **5**: 1840–1852.

Fauteux F, Strömvik M V. **2009**. Seed storage protein gene promoters contain conserved DNA motifs in Brassicaceae, Fabaceae and Poaceae. *BMC plant biology* **9**: 126.

FELDMAN M. **2000**. Origin of cultivated wheat. *The World Wheat Book, a History of Wheat Breeding*.

Finkelstein RR, Gibson SI. **2002**. ABA and sugar interactions regulating development: cross-talk or voices in a crowd? *Current Opinion in Plant Biology* **5**: 26–32.

Friedman N, Linial M, Nachman I, Pe'er D. **2000**. Using Bayesian networks to analyze expression data. *Journal of computational biology : a journal of computational molecular cell biology* **7**: 601–20.

Fuller DQ. **2007**. Contrasting Patterns in Crop Domestication and Domestication Rates: Recent Archaeobotanical Insights from the Old World. *Annals of Botany* **100** : 903–924.

Gallais A, Coque M, Quilléré I, Prioul J-L, Hirel B. **2006**. Modelling postsilking nitrogen fluxes in maize (Zea mays) using 15N-labelling field experiments. *New Phytologist* **172**: 696–707.

Gazzarrini S, Tsuchiya Y, Lumba S, Okamoto M, McCourt P. **2004**. The Transcription Factor FUSCA3 Controls Developmental Timing in Arabidopsis through the Hormones Gibberellin and Abscisic Acid. *Developmental Cell* **7**: 373–385.

Gibson SI. **2005**. Control of plant development and gene expression by sugar signaling. *Current Opinion in Plant Biology* **8**: 93–102.

Giles J. **2005**. Nitrogen study fertilizes fears of pollution. *Nature* **433**: 791.

Goesaert H, Brijs K, Veraverbeke WS, Courtin CM, Gebruers K, Delcour JA. **2005**. Wheat flour constituents: how they impact bread quality, and how to impact their functionality. *Trends in Food Science & Technology* **16**: 12–30.

Götz K-P, Staroske N, Radchuk R, Emery RJN, Wutzke K-D, Herzog H, Weber H. **2007**. Uptake and allocation of carbon and nitrogen in Vicia narbonensis plants with increased seed sink strength achieved by seed-specific expression of an amino acid permease. *Journal of Experimental Botany* **58** : 3183–3195.

Graybosch RA, Peterson CJ. **2010**. Genetic Improvement in Winter Wheat Yields in the Great Plains of North America , 1959 – 2008.

Graybosch R a., Peterson CJ, Shelton DR, Baenziger PS. **1996**. Genotypic and Environmental Modification of Wheat Flour Protein Composition in Relation to End-Use Quality. *Crop Science* **36**: 296.

Gubler F, Jacobsen J V. **1992**. Gibberellin-Responsive Elements in the Promoter of a Barley. *The Plant Cell* **4**: 1435–1441.

Gupta VVSR, Germida JJ. **1989**. Influence of bacterial-amoebal interactions on sulfur transformations in soil. *Soil Biology and Biochemistry* **21**: 921–930.

Gupta RB, Masci S, Lafiandra D, Bariana HS, MacRitchie F. **1996**. Accumulation of protein subunits and their polymers in developing grains of hexaploid wheats. *Journal of Experimental Botany* **47** : 1377–1385.

Hamed RI, Ahson SI, Parveen R. **2010**. A new approach for modelling gene regulatory networks using fuzzy petri nets. *Journal of integrative bioinformatics* **7**: 1–16.

Hawkesford MJ. **2000**. Plant responses to sulphur deficiency and the genetic manipulation of sulphate transporters to improve S-utilization efficiency. *Journal of experimental botany* **51**: 131–8.

Heiner M, Lehrack S, Gilbert D, Marwan W. **2009**. Extended Stochastic Petri Nets for Model-Based Design of Wetlab Experiments. In: Priami C, Back R-J, Petre I, eds. Lecture Notes in Computer Science. Transactions on Computational Systems Biology XI SE - 7. Springer Berlin Heidelberg, 138–163.

Hesse H, Nikiforova V, Gakière B, Hoefgen R. **2004**. Molecular analysis and control of cysteine biosynthesis: integration of nitrogen and sulphur metabolism. *Journal of experimental botany* **55**: 1283–92.

Heun M, Scha R, Klawan D, Castagna R, Accerbi M, Borghi B, Salamini F. **1997**. Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. **278**: 1312–1314.

Higashi Y, Hirai MY, Fujiwara T, Naito S, Noji M, Saito K. **2006**. Proteomic and transcriptomic analysis of Arabidopsis seeds: molecular evidence for successive processing of seed proteins and its implication in the stress response to sulfur nutrition. *The Plant journal : for cell and molecular biology* **48**: 557–71.

Hirai MY, Fujiwara T, Chino M, Naito S. **1995**. Effects of Sulfate Concentrations on the Expression of a Soybean Seed Storage Protein Gene and its Reversibility in Transgenic Arabidopsis thaliana. *Plant and Cell Physiology* **36** : 1331–1339.

Hirel B, Le Gouis J, Ney B, Gallais A. **2007**. The challenge of improving nitrogen use efficiency in crop plants: towards a more central role for genetic variability and quantitative genetics within integrated approaches. *Journal of experimental botany* **58**: 2369–87.

Hirel B, Tétu T, Lea PJ, Dubois F. **2011**. Improving Nitrogen Use Efficiency in Crops for Sustainable Agriculture. *Sustainability* **3**: 1452–1485.

Hocking PJ, Steer BT. **1983**. Distribution of Nitrogen during Growth of Sunflower (Helianthus annuus L.). *Annals of Botany* **51**: 787–799.

Howarth JR, Parmar S, Barraclough PB, Hawkesford MJ. **2009**. A sulphur deficiency-induced gene, sdi1, involved in the utilization of stored sulphate pools under sulphur-limiting conditions has potential as a diagnostic indicator of sulphur nutritional status. *Plant Biotechnology Journal* **7**: 200–209.

Hu P, Bader G, Wigle DA, Emili A. **2007**. Computational prediction of cancer-gene function. *Nat Rev Cancer* **7**: 23–34.

Initiative TIB. **2010**. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**: 763–768.

Irving LJ, Robinson D. **2006**. A dynamic model of Rubisco turnover in cereal leaves. *New Phytologist* **169**: 493–504.

Isabel-LaMoneda I, Diaz I, Martinez M, Mena M, Carbonero P. **2003**. SAD: a new DOF protein from barley that activates transcription of a cathepsin B-like thiol protease gene in the aleurone of germinating seeds. *The Plant journal : for cell and molecular biology* **33**: 329–40.

Jamai a., Tommasini R, Martinoia E, Delrot S. **1996**. Characterization of glutathione uptake in broad bean leaf protoplasts. *Plant physiology* **111**: 1145–1152.

Johnston AE, McEwen J, Lane PW, Hewitt M V, Poulton PR, Yeoman DP. **1994**. Effects of one to six year old ryegrass-clover leys on soil nitrogen and on the subsequent yields and fertilizer nitrogen requirements of the arable sequence winter wheat, potatoes, winter wheat, winter beans (Vicia faba) grown on a sandy loam soil. *The Journal of Agricultural Science* **122**: 73–89.

Kaderali L, Dazert E, Zeuge U, Frese M, Bartenschlager R. **2009**. Reconstructing signaling pathways from RNAi data using probabilistic Boolean threshold networks. *Bioinformatics* **25** : 2229–2235.

Kagaya Y, Okuda R, Ban A, Toyoshima R, Tsutsumida K, Usui H, Yamamoto A, Hattori T. **2005**. Indirect ABA-dependent Regulation of Seed Storage Protein Genes by FUSCA3 Transcription Factor in Arabidopsis. *Plant and Cell Physiology* **46** : 300–311.

Kareiva P, Watts S, McDonald R, Boucher T. **2007**. Domesticated Nature: Shaping Landscapes and Ecosystems for Human Welfare. *Science* **316** : 1866–1869.

Kataoka T, Watanabe-Takahashi A, Hayashi N, Ohnishi M, Mimura T, Buchner P, Hawkesford MJ, Yamaya T TH. **2004**. Vacuolar sulfate transporters are essential determinants controlling internal distribution of sulfate in Arabidopsis. *The Plant Cell …* **16**: 2693–2704.

Kawakatsu T, Takaiwa F. **2010**. Differences in transcriptional regulatory mechanisms functioning for free lysine content and seed storage protein accumulation in rice grain. *Plant & cell physiology* **51**: 1964–74.

Klie S, Mutwil M, Persson S, Nikoloski Z. **2012**. Inferring gene functions through dissection of relevance networks: interleaving the intra-and inter-species views. *Molecular BioSystems* **8**: 2233–2241.

Kohanski MA, Dwyer DJ, Wierzbowski J, Cottarel G, Collins JJ. **2008**. Mistranslation of Membrane Proteins and Two-Component System Activation Trigger Antibiotic-Mediated Cell Death. *Cell* **135**: 679–690.

Kornberg RD. **1999**. Eukaryotic transcriptional control. *Trends in Biochemical Sciences* **24**: M46–M49.

Lafiandra D, Masci S, Blumenthal C, Wrigley CW. **1999**. The formation of glutenin polymer in practice. *Cereal foods world* **44**: 572–578.

Lakkineni KC, Abrol YP. **1992**. Sulphur Requirement of Rapeseed-Mustard, Groundnut and Wheat: A Comparative Assessment. *Journal of Agronomy and Crop Science* **169**: 281–285.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, *et al.* 2011. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research*: 1–9.

Large EC. **1954**. GROWTH STAGES IN CEREALS ILLUSTRATION OF THE FEEKES SCALE. *Plant Pathology* **3**: 128–129.

Lemmens K, De Bie T, Dhollander T, De Keersmaecker SC, Thijs IM, Schoofs G, De Weerdt A, De Moor B, Vanderleyden J, Collado-Vides J, *et al.* 2009. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. *Genome biology* **10**: R27.

Li G, Chandrasekharan M, Wolffe A, Hall T. **2001**. Chromatin structure and phaseolin gene regulation. *Plant molecular biology*.

Liang S, Fuhrman S, Somogyi R. **1998**. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific symposium on …* **29**: 18–29.

Ling H-Q, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, *et al.* 2013. Draft genome of the wheat A-genome progenitor Triticum urartu . *Nature* **496**: 87–90.

Lohmer S, Maddaloni M, Motto M. **1991**. The maize regulatory locus Opaque-2 encodes a DNA-binding protein which activates the transcription of the b-32 gene. *The EMBO …* **10**: 617–24.

Malagoli P, Laine P, Rossato L, Ourry A. **2005**. Dynamics of Nitrogen Uptake and Mobilization in Field-grown Winter Oilseed Rape (Brassica napus) from Stem Extension to Harvest: I. Global N Flows between Vegetative and Reproductive Tissues in Relation to Leaf Fall and their Residual N. *Annals of Botany* **95**: 853–861.

Martinez R, Pasquier N, Pasquier C. **2008**. GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. *Bioinformatics (Oxford, England)* **24**: 2643–4.

Martre P, Jamieson PD, Semenov MA, Zyskowski RF, Porter JR, Triboi E. **2006**. Modelling protein content and composition in relation to crop nitrogen dynamics for wheat. *European Journal of Agronomy* **25**: 138–154.

Martre P, Porter JR, Jamieson PD. **2003**. Modeling Grain Nitrogen Accumulation and Protein Composition to Understand the Sink / Source Regulations of Nitrogen Remobilization for Wheat. **133**: 1959–1967.

Masclaux-Daubresse C, Daniel-Vedele F, Dechorgnat J, Chardon F, Gaufichon L, Suzuki A. **2010**. Nitrogen uptake, assimilation and remobilization in plants: challenges for sustainable and productive agriculture. *Annals of botany* **105**: 1141–57.

McGrath SP, Zhao FJ. **1996**. Sulphur uptake, yield responses and the interactions between nitrogen and sulphur in winter oilseed rape (Brassica napus). *The Journal of Agricultural Science* **126**: 53–62.

Mena M, Vicente-Carbajosa J, Schmidt RJ, Carbonero P. **1998**. An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. *The Plant journal : for cell and molecular biology* **16**: 53–62.

Mendel RR. **1997**. Molybdenum cofactor of higher plants: biosynthesis and molecular biology. *Planta* **203**: 399–405.

Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juárez K, Contreras-Moreira B, *et al.* 2009. Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in <italic>E. coli</italic>. *PLoS ONE* **4**: e7526.

Michael Beman J, Arrigo KR, Matson PA. **2005**. Agricultural runoff fuels large phytoplankton blooms in vulnerable areas of the ocean. *Nature* **434**: 211–214.

Michoel T, De Smet R, Joshi A, Van de Peer Y, Marchal K. **2009**. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC systems biology* **3**: 49.

Miflin BJ, Habash DZ. **2002**. The role of glutamine synthetase and glutamate dehydrogenase in nitrogen assimilation and possibilities for improvement in the nitrogen utilization of crops. *Journal of Experimental Botany* **53** : 979–987.

Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K. **2009**. TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant physiology* **150**: 1135–46.

Moll RH, Kamprath EJ, Jackson WA. **1982**. Analysis and Interpretation of Factors Which Contribute to Efficiency of Nitrogen Utilization1. : 562–564.

Monaghan JM, Scrimgeour CM, Stein WM, Zhao FJ, Evans EJ. **1999**. Sulphur accumulation and redistribution in wheat (Triticum aestivum): a study using stable sulphur isotope ratios as a tracer system. *Plant, Cell & Environment* **22**: 831–839.

Moore KJ, Moser LE. **1995**. Quantifying Developmental Morphology of Perennial Grasses. *Crop Science* **35**: 37.

Moreno-Risueno MA, González N, Díaz I, Parcy F, Carbonero P, Vicente-Carbajosa J. **2008**. FUSCA3 from barley unveils a common transcriptional regulation of seed-specific genes between cereals and Arabidopsis. *The Plant journal : for cell and molecular biology* **53**: 882–94.

Morgan XC, Ni S, Miranker DP, Iyer VR. **2007**. Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. *BMC bioinformatics* **8**: 445.

Müller M, Knudsen S. **1993**. The nitrogen response of a barley C-hordein promoter is controlled by positive and negative regulation of the GCN4 and endosperm box. *The Plant Journal* **4**: 343–355.

Nesbitt M, Samuel D. **1998**. Wheat domestication: archaeobotanical evidence. *Science* **279**: 1431.

Nielsen N, Dickinson C. **1989**. Characterization of the glycinin gene family in soybean. *The Plant Cell …* **1**: 313–28.

Oliver S. **2000**. Proteomics: Guilt-by-association goes global. *Nature* **403**: 601–603.

Oury F-X, Godin C. **2007**. Yield and grain protein concentration in bread wheat: how to use the negative relationship between the two characters to identify favourable genotypes? *Euphytica* **157**: 45–57.

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, *et al.* 2007. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic acids research* **35**: D883–7.

Pan Y, Bradley G, Pyke K, Ball G, Lu C, Fray R, Marshall A, Jayasuta S, Baxter C, van Wijk R, *et al.* 2013. Network Inference Analysis Identifies an APRR2-Like Gene Linked to Pigment Accumulation in Tomato and Pepper Fruits. *Plant physiology* **161**: 1476–85.

Panozzo JF, Eagles HA, Wootton M. **2001**. Changes in protein composition during grain development in wheat. *Australian Journal of Agricultural Research* **52**: 485–493.

Paponov I, Aufhammer W, Kaul H-P, Ehmele F-P. **1996**. Nitrogen efficiency components of winter cereals. *European Journal of Agronomy* **5**: 115–124.

Parcy F, Valon C. **1994**. Regulation of gene expression programs during Arabidopsis seed development: roles of the ABI3 locus and of endogenous abscisic acid. *The Plant Cell …* **6**: 1567–82.

Pedersen CA, Knudsen L, Schnug E. **1998**. Sulphur Fertilisation. In: Schnug E, ed. Nutrients in Ecosystems. Sulphur in Agroecosystems SE - 4. Springer Netherlands, 115–134.

Petri CA. **1962**. Kommunikation mit Automaten.

Plessis A, Ravel C, Bordes J, Balfourier F, Martre P. **2013**. Association study of wheat grain protein composition reveals that gliadin and glutenin composition are trans-regulated by different chromosome regions. *Journal of experimental botany* **64**: 3627–44.

Prescott LM, Harley JP, Klein DA. **2002**. Microbiology.(5thedn).

Price DDS. **1976**. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27**: 292–306.

Qian J, Luscombe NM, Gerstein M. **2001**. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Molecular Biology* **313**: 673–681.

Randall PJ, Spencer K, Freney JR. **1981**. Sulfur and nitrogen fertilizer effects on wheat. I. Concentrations of sulfur and nitrogen and the nitrogen to sulfur ratio in grain, in relation to the yield response. *Australian Journal of Agricultural Research* **32**: 203–212.

Ravel C, Martre P, Romeuf I, Dardevet M, El-Malki R, Bordes J, Duchateau N, Brunel D, Balfourier F, Charmet G. **2009**. Nucleotide polymorphism in the wheat transcriptional activator Spa influences its pattern of expression and has pleiotropic effects on grain protein composition, dough viscoelasticity, and grain hardness. *Plant physiology* **151**: 2133–44.

Reneau RB, Brann DE, Donohue SJ. **1986**. Effect of sulfur on winter wheat grown in the coastal plain of Virginia. *Communications in Soil Science and Plant Analysis* **17**: 149–158.

Riechmann JL. **2000**. Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. *Science* **290**: 2105–2110.

Riechmann J. **2002**. Transcriptional regulation: a genomic overview. *The Arabidopsis Book*.

Riechmann JL, Ratcliffe OJ. **2000**. A genomic perspective on plant transcription factors. *Current Opinion in Plant Biology* **3**: 423–434.

Riley M. **1993**. Functions of the gene products of Escherichia coli. *Microbiological reviews* **57**: 862–952.

Rison SCG, Hodgman TC, Thornton JM. **2000**. Comparison of functional annotation schemes for genomes. *Functional & Integrative Genomics* **1**: 56–69.

Rolland F, Baena-Gonzalez E, Sheen J. **2006**. SUGAR SENSING AND SIGNALING IN PLANTS: Conserved and Novel Mechanisms. *Annual Review of Plant Biology* **57**: 675–709.

Romeuf I, Tessier D, Dardevet M, Branlard G, Charmet G, Ravel C. **2010**. wDBTF: an integrated database resource for studying wheat transcription factor families. *BMC genomics* **11**: 185.

Rosales EP, Iannone MF, Groppa MD, Benavides MP. **2011**. Nitric oxide inhibits nitrate reductase activity in wheat leaves. *Plant Physiology and Biochemistry* **49**: 124–130.

Rubio-Somoza I, Martinez M, Abraham Z, Diaz I, Carbonero P. **2006a**. Ternary complex formation between HvMYBS3 and other factors involved in transcriptional control in barley seeds. *The Plant Journal* **47**: 269–281.

Rubio-Somoza I, Martinez M, Diaz I, Carbonero P. **2006b**. HvMCB1, a R1MYB transcription factor from barley with antagonistic regulatory functions during seed development and germination. *The Plant Journal* **45**: 17–30.

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M, *et al.* 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic acids research* **32**: 5539–45.

Ruiz-apillas C, Moral A. **2002**. RELATION BETWEEN THE FREE AMINO ACIDS, ANSERINE AND THE TOTAL VOLATILE BASIC NITROGEN PRODUCED IN MUSCLE OF HAKE (MERLUCCIUS MERLUCCIUS, L.) DURING ICED STORAGE. *Journal of Food Biochemistry* **26**: 37–48.

Rustenholz C, Choulet F, Laugier C, Safár J, Simková H, Dolezel J, Magni F, Scalabrin S, Cattonaro F, Vautrin S, *et al.* 2011. A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant physiology* **157**: 1596–608.

Ruuska SA, Girke T, Benning C, Ohlrogge JB. **2002**. Contrapuntal Networks of Gene Expression during Arabidopsis Seed Filling. **14**: 1191–1206.

Rzhetsky A, Gomez SM. **2001**. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* **17** : 988–996.

Salon C, Munier-Jolain N, Duc G. **2001**. Grain legume seed filling in relation to nitrogen acquisition: a review and prospects with particular reference to pea. … **21**: 539–552.

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. **2008**. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *The Plant cell* **20**: 11–24.

Salvagiotti F, Castellarín JM, Miralles DJ, Pedrol HM. **2009**. Sulfur fertilization improves nitrogen use efficiency in wheat by increasing nitrogen uptake. *Field Crops Research* **113**: 170–177.

Salvagiotti F, Miralles DJ. **2008**. Radiation interception, biomass production and grain yield as affected by the interaction of nitrogen and sulfur fertilization in wheat. *European Journal of Agronomy* **28**: 282–290.

Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EKS, Sen TZ, Lawrence CJ. **2011**. MaizeGDB: curation and outreach go hand-in-hand. *Database* **2011** .

Scherer HW. **2001**. Sulphur in crop production — invited paper. *European Journal of Agronomy* **14**: 81–111.

Schiltz S, Munier-Jolain N, Jeudy C. **2005**. Dynamics of exogenous nitrogen partitioning and nitrogen remobilization from vegetative organs in pea revealed by 15N in vivo labeling throughout seed filling. *Plant Pysiology* **137**: 1463–1473.

Schneider A, Schatten T, Rennenberg H. **1994**. Exchange between phloem and xylem during long distance transport of glutathione in spruce trees [Picea abies[Karst.] L.). *Journal of Experimental Botany* **45** : 457–462.

Serres M, Riley M. **2000**. MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microbial & comparative genomics*.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. **2003**. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. : 2498–2504.

Shewry PR. **2009**. Wheat. *Journal of experimental botany* **60**: 1537–53.

Shewry PR, Halford NG. **2002**. Cereal seed storage proteins: structures, properties and role in grain utilization. *Journal of experimental botany* **53**: 947–58.

Shewry P, Halford N, Lafiandra D. **2003**. Genetics of wheat gluten proteins. *Advances in genetics*.

Shewry PR, Popineau Y, Lafiandra D, Belton P, Lellis SC De. **2001**. Wheat glutenin subunits and dough elasticity : findings of the EUROWHEAT project. **11**: 433–441.

Sieling K, Schröder H. **1998**. Yield, N uptake, and apparent N-use efficiency of winter wheat and winter barley grown in different cropping systems. *The Journal of …* **131**: 375–387.

Sinclair TR, Horie T. **1989**. Leaf Nitrogen, Photosynthesis, and Crop Radiation Use Efficiency: A Review. : 90–98.

Sinclair TR, de Wit CT. **1975**. Photosynthate and Nitrogen Requirements for Seed Production by Various Crops. *Science* **189** : 565–567.

De Smet R, Marchal K. **2010**. Advantages and limitations of current network inference methods. *Nature reviews. Microbiology* **8**: 717–29.

Smith IK. **1975**. Sulfate Transport in Cultured Tobacco Cells. *Plant Physiology* **55**: 303–307.

Smith CW. **1995**. *Crop production : evolution, history, and technology / C. Wayne Smith*. New York : J. Wiley.

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. **2011**. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)* **27**: 431–2.

Sreenivasulu N, Radchuk V, Strickert M, Miersch O, Weschke W, Wobus U. **2006**. Gene expression patterns reveal tissue-specific signaling networks controlling programmed cell death and ABA-regulated maturation in developing barley seeds. *The Plant journal : for cell and molecular biology* **47**: 310–27.

Stone SL, Braybrook SA, Paula SL, Kwong LW, Meuser J, Pelletier J, Hsieh T-F, Fischer RL, Goldberg RB, Harada JJ. **2008**. Arabidopsis LEAFY COTYLEDON2 induces maturation traits and auxin activity: Implications for somatic embryogenesis. *Proceedings of the National Academy of Sciences* **105** : 3151–3156.

Struhl K. **1999**. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* **98**: 1–4.

Takaiwa F, Hirose S, Takagi H, Yang L, Wakasa Y. **2009**. Deposition of a recombinant peptide in ER-derived protein bodies by retention with cysteine-rich prolamins in transgenic rice seed. *Planta* **229**: 1147–1158.

Thieffry D, Huerta AM, Pérez-Rueda E, Collado-Vides J. **1998**. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *BioEssays* **20**: 433–440.

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller L a., Rhee SY, Stitt M. **2004**. Mapman: a User-Driven Tool To Display Genomics Data Sets Onto Diagrams of Metabolic Pathways and Other Biological Processes. *The Plant Journal* **37**: 914–939.

Triboï E, Martre P, Triboï-Blondel A-M. **2003a**. Environmentally-induced changes in protein composition in developing grains of wheat are related to changes in total protein content. *Journal of experimental botany* **54**: 1731–42.

Triboï E, Martre P, Triboï-Blondel A-M. **2003b**. Environmentally-induced changes in protein composition in developing grains of wheat are related to changes in total protein content. *Journal of experimental botany* **54**: 1731–42.

Verdier J, Lalanne D, Pelletier S, Torres-Jerez I, Righetti K, Bandyopadhyay K, Leprince O, Chatelain E, Vu BL, Gouzy J, *et al.* 2013. A regulatory network-based approach dissects late maturation processes related to the acquisition of desiccation tolerance and longevity of Medicago truncatula seeds. *Plant physiology* **163**: 757–74.

Verdier J, Thompson RD. **2008**. Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant & cell physiology* **49**: 1263–71.

Verlag F, Anderson JW, Fitzgerald MA. **2001**. Physiological and metabolic origin of sulphur for the synthesis of seed storage proteins. **456**.

Vicente-Carbajosa J, Oñate L, Lara P, Diaz I, Carbonero P. **1998**. Barley BLZ1: a bZIP transcriptional activator that interacts with endosperm-specific gene promoters. *The Plant journal : for cell and molecular biology* **13**: 629–40.

Voit EO. **2000**. *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*. Cambridge University Press.

Wagner A. **2003**. How the global structure of protein interaction networks evolves. *Proceedings of the Royal Society of …* **270**: 457–66.

Wagner A, Fell DA. **2001**. The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268** : 1803–1810.

Walling L. **1986**. Transcriptional and post-transcriptional regulation of soybean seed protein mRNA levels. *Proceedings of the …* **83**: 2123–7.

Wan Y, Gritsch CS, Hawkesford MJ, Shewry PR. **2014**. Effects of nitrogen nutrition on the synthesis and deposition of the ω-gliadins of wheat. *Annals of Botany* **113** : 607–615.

Weber H, Borisjuk L, Wobus U. **2005**. MOLECULAR PHYSIOLOGY OF LEGUME SEED DEVELOPMENT. *Annual Review of Plant Biology* **56**: 253–279.

Weichert N, Saalbach I, Weichert H, Kohl S, Erban A, Kopka J, Hause B, Varshney A, Sreenivasulu N, Strickert M, *et al.* 2010. Increasing sucrose uptake capacity of wheat grains stimulates storage protein synthesis. *Plant physiology* **152**: 698–710.

Whisstock JC, Lesk AM. **2003**. Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics* **36**: 307–340.

Wickert E, Marcondes J, Lemos MV, Lemos EGM. **2007**. Nitrogen assimilation in Citrus based on CitEST data mining. *Genetics and Molecular Biology* **30**: 810–818.

Wieser H. **2007**. Chemistry of gluten proteins. *Food Microbiology* **24**: 115–119.

Wieser H, Gutser R, von Tucher S. **2004**. Influence of sulphur fertilisation on quantities and proportions of gluten protein types in wheat flour. *Journal of Cereal Science* **40**: 239–244.

Wilson K a, Rightmire BR, Chen JC, Tan-Wilson a L. **1986**. Differential Proteolysis of Glycinin and beta-Conglycinin Polypeptides during Soybean Germination and Seedling Growth. *Plant physiology* **82**: 71–6.

Witte C-P. **2011**. Urea metabolism in plants. *Plant Science* **180**: 431–438.

Wu J, O'Donnell AG, Syers JK. **1993**. Microbial growth and sulphur immobilization following the incorporation of plant residues into soil. *Soil Biology and Biochemistry* **25**: 1567–1573.

Xiong Y, Liu T, Tian C, Sun S, Li J, Chen M. **2005**. Transcription factors in rice: a genome-wide comparative analysis between monocots and eudicots. *Plant molecular biology*.

Zadoks JC, Chang TT, Konzak CF. **1974**. A decimal code for the growth stages of cereals. *Weed Research* **14**: 415–421.

Zaroug MG, Munns DN. **1980**. Effects of phosphorus and sulfur nutrition on soluble sugars and growth inClitoria ternatea L. *Plant and Soil* **55**: 243–250.

Zhang H, Sreenivasulu N, Weschke W, Stein N, Rudd S, Radchuk V, Potokina E, Scholz U, Schweizer P, Zierold U, *et al.* 2004. Large-scale analysis of the barley transcriptome based on expressed sequence tags. *The Plant journal : for cell and molecular biology* **40**: 276–90.

Zhao F, Evans EJ, Bilsborrow PE, Syers JK. **1993**. Influence of sulphur and nitrogen on seed yield and quality of low glucosinolate oilseed rape (Brassica napus L). *Journal of the Science of Food and Agriculture* **63**: 29–37.

Zheng Z, Kawagoe Y, Xiao S, Li Z, Okita T, Hau TL, Lin A, Murai N. **1993**. 5' distal and proximal cis-acting regulator elements are required for developmental control of a rice seed storage protein glutelin gene. *The Plant Journal* **4**: 357–366.

Zhou T. **2010**. Optimal control for probabilistic Boolean networks. *IET Systems Biology* **4**: 99–107(8).

Zohary D, Hopf M. **1988a**. Domestication of plants in the old world. Clarendon.

Zohary D, Hopf M. **1988b**. Domestication of plants in the old world. Clarendon.

Zohary D, Hopf M. **1993**. Domestication of plants in the world. The origin and spread of cultivated plants in west Asia, Europe and the Nile valley.

Zörb C, Grover C, Steinfurth D, Hermann Mühling K. **2010**. Quantitative proteome analysis of wheat gluten as influenced by N and S nutrition. *Plant and Soil* **327**: 225–234.

Zörb C, Steinfurth D, Seling S, Langenkämper G, Koehler P, Wieser H, Lindhauer MG, Mühling KH. **2009**. Quantitative protein composition and baking quality of winter wheat as affected by late sulfur fertilization. *Journal of agricultural and food chemistry* **57**: 3877–85.

PUBLICATION 1

dbWFA: a web-based database for functional annotation of Triticum aestivum transcripts

Jonathan Vincent[1,2,3], Zhanwu Dai[1,2,†], Catherine Ravel[1,2], Frédéric Choulet[1,2], Said Mouzeyar[1,2], M. Fouad Bouzidi[1,2], Marie Agier[3] and Pierre Martre[1,2]

[1]INRA, UMR1095 Génétique, Diversité et Ecophysiologie des Céréales, 5 chemin de Beaulieu, F-63 039 Clermont-Ferrand, France; [2]Université Blaise Pascal, UMR1095 Diversité et Ecophysiologie des Céréales, F-63 177 Aubière, France; [3]Université Blaise Pascal, UMR6158 CNRS LIMOS Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, F-63 173 Aubière, France; [†]Current address: INRA, UMR1287 Ecophysiologie et Génomique Fonctionnelle de la Vigne, Institut des Sciences de la Vigne et du Vin, F-33 882 Villenave d'Ornon, France.

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

Database tool

dbWFA: a web-based database for functional annotation of *Triticum aestivum* transcripts

Jonathan Vincent[1,2,3], Zhanwu Dai[1,2], Catherine Ravel[1,2], Frédéric Choulet[1,2], Said Mouzeyar[1,2], M. Fouad Bouzidi[1,2], Marie Agier[3] and Pierre Martre[1,2,*]

[1]INRA, UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 Chemin de Beaulieu, Clermont-Ferrand, F-63 039 Cedex 2, France, [2]Blaise Pascal University, UMR1095 Genetics, Diversity and Ecophysiology of Cereals, Aubière F-63 177, France and [3]Blaise Pascal University, UMR6158 CNRS LIMOS Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, Aubière F-63 173, France

Present address: Zhanwu Dai, INRA, ISVV, UMR1287 Écophysiologie et Génomique Fonctionnelle de la Vigne (EGFV), F-33 882 Villenave d'Ornon, France

***Corresponding author:** Tel: +33 473 624 351; Fax: +33 473 624 457; Email: pierre.martre@clermont.inra.fr

Submitted 11 July 2012; Revised 19 February 2013; Accepted 20 February 2013

Citation details: Vincent,J., Dai,Z.W., Ravel,C. *et al*. dbWFA: a web-based database for functional annotation of *Triticum aestivum* transcripts. *Database* (2013) Vol. 2013: article ID bat014; doi:10.1093/database/bat014

The functional annotation of genes based on sequence homology with genes from model species genomes is time-consuming because it is necessary to mine several unrelated databases. The aim of the present work was to develop a functional annotation database for common wheat *Triticum aestivum* (L.). The database, named dbWFA, is based on the reference NCBI UniGene set, an expressed gene catalogue built by expressed sequence tag clustering, and on full-length coding sequences retrieved from the TriFLDB database. Information from good-quality heterogeneous sources, including annotations for model plant species *Arabidopsis thaliana* (L.) Heynh. and *Oryza sativa* L., was gathered and linked to *T. aestivum* sequences through BLAST-based homology searches. Even though the complexity of the transcriptome cannot yet be fully appreciated, we developed a tool to easily and promptly obtain information from multiple functional annotation systems (Gene Ontology, MapMan bin codes, MIPS Functional Categories, PlantCyc pathway reactions and TAIR gene families). The use of dbWFA is illustrated here with several query examples. We were able to assign a putative function to 45% of the UniGenes and 81% of the full-length coding sequences from TriFLDB. Moreover, comparison of the annotation of the whole *T. aestivum* UniGene set along with curated annotations of the two model species assessed the accuracy of the annotation provided by dbWFA. To further illustrate the use of dbWFA, genes specifically expressed during the early cell division or late storage polymer accumulation phases of *T. aestivum* grain development were identified using a clustering analysis and then annotated using dbWFA. The annotation of these two sets of genes was consistent with previous analyses of *T. aestivum* grain transcriptomes and proteomes.

Database URL: urgi.versailles.inra.fr/dbWFA/

Introduction

Triticum aestivum (L.), common wheat or bread wheat, is one of the most important staple crops in the world. It is cultivated worldwide and provides >20% of the calories and proteins in the human diet (http://faostat.fao.org). Although ongoing sequencing efforts have already produced important genomic resources (1–4), the complete sequencing and annotation of the hexaploid (2n = 6× = 42, AABBDD) *T. aestivum* genome has yet to be achieved. A first version of the genome of the bread wheat cv. Chinese Spring has recently been published (4), providing the scientific community with highly valuable genomic and evolutionary information, which will facilitate

Database tool

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

genome-wide analysis of bread wheat. However, because of the low-coverage (5-fold) shotgun sequencing method used in this project, this resource does not represent a high-quality draft of the wheat genome in terms of sequence completion, quality and annotation. Genome-wide analysis of gene expression by means of expression microarrays or transcriptome sequencing (RNA-Seq) is now being adopted in *T. aestivum* (5, 6), but analysing such large datasets requires extensive annotation efforts. Data fragmentation and technical and semantic heterogeneity can severely limit the efficient extraction and interpretation of biological data (7, 8).

More and more genomic information is becoming available for *T. aestivum* research. Various resources and associated tools grant the user a structural overview of expressed sequence tag (EST) (ITEC, http://avena.pw.usda.gov/genome/) (9, 10) or bacterial artificial chromosome clone libraries (11, 12) for instance (13–15). Important initiatives are underway to facilitate the breeding of improved Triticeae varieties. The TriticeaeGenome project (www.triticeaegenome.eu) grants access to comprehensive information extracted from experimental data to provide a better understanding of Triticeae genomes (16). The global database GrainGenes (http://wheat.pw.usda.gov/) provides a variety of services and bioinformatics tools for the Triticeae and *Avena sativa* research communities. The HarvEST database (http://harvest.ucr.edu/) (17), dedicated to several crop species, including *T. aestivum* and *Hordeum vulgare*, provides access to curated EST assemblies, comparative analysis tools and links to orthologues in related model plant species. Together, these resources compile and cross-reference a great deal of information on physical and genetic mapping, markers, sequence variations and quantitative trait loci. To some extent, they also provide information leading indirectly to predicted gene product functions, but none of them is focused on functional gene annotation, and it is necessary to navigate through numerous unlinked resources to extract functional information.

Recently, pipelines for the automated annotation of genomic sequences of *T. aestivum* and related species have been developed (3, 18). These pipelines are based on the prediction of gene models within genome sequences, so they are not able to functionally annotate sequences originating from transcriptome sequencing like ESTs. Because no reference genome sequence is yet available for *T. aestivum*, a massive sequencing effort has produced more than a million ESTs (http://wheat.pw.usda.gov/genome/). To deal with the high level of redundancy of this resource, these sequences were clustered (i.e. overlapping and partial polyA-tailed expressed sequences are grouped) to provide a reference set of unique expressed genes, NCBI UniGenes (http://www.ncbi.nlm.nih.gov/UniGene). The assembly conditions used to build NCBI UniGenes make it the most comprehensive coding DNA (cDNA) assembly available to date, and UniGene assemblies have been used as a reference set of sequences for many species. An additional effort was made to construct full-length cDNA sequences that were included in the TriFLDB database (19). Full-length cDNA sequences are most commonly used in genome annotation as a resource for cross-species comparative analyses. Currently, TriFLDB is the most reliable source of full-length cDNA sequences in *T. aestivum*. TriFLDB includes annotations based on homologies found by searching protein databases, extensive Gene Ontology (GO) annotations and InterProScan results. Recently, a new collection of nearly 1 million ESTs, assembled into contigs and singlets, was annotated with GO terms (20), but meaningful prediction of gene function requires more than one system of annotation.

After the sequencing of the first plant genome, *Arabidopsis thaliana* in 2000 (21), several plant sequencing projects have been successful. The sequenced genomes most closely related to the *T. aestivum* genome are those of *Oryza sativa* ssp. *indica* (22), *Oryza sativa* ssp. *japonica* (23), *Zea mays* (24), *Glycine max* (25), *Sorghum bicolor* (26), *Brachypodium distachyon* (27) and *Hordeum vulgare* (28). Both structural and functional annotation resources for these species are developing steadily. One of the most effective methods to annotate a transcript is to find its orthologous counterparts in well-annotated closely related genomes (29, 30). Although *H. vulgare* (L.) would be expected to be the most useful reference because it is more closely related to *T. aestivum*, comprehensive and high-quality annotations of gene function are only available for *O. sativa* and *A. thaliana* (2), essentially owing to the lengthy and accurate annotation efforts undertaken.

To annotate ESTs or transcripts using sequence homology, it is necessary to navigate through unrelated databases. Some tools using this homology approach have been developed. For instance, Blast2GO (31) can be queried using *T. aestivum* sequences to give GO results. ONDEX (7), developed with the challenges of functional annotation of *T. aestivum* genome in mind, combines data integration from various sources and various mining methods, including graph-based analyses, to annotate wheat gene functions according to a wisely chosen set of annotation standards. However, ONDEX does not provide easy access to workable static results that are often required in research. To fill this gap, we developed dbWFA, an open-access database relating the *T. aestivum* UniGene set and the full-length cDNA sequences from TriFLDB to *A. thaliana* (TAIR10) (32) and *O. sativa* (pseudomolecules version 7.0) (33) annotation through BLAST (34) results. dbWFA also includes the inventory database of *T. aestivum* transcription factors (wDBTF) (35) and hand-curated gene families (36). As an all-in-one interface for the annotation of *T. aestivum* sequences, dbWFA will be

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

useful to the researchers working on *T. aestivum* and more generally on cereals, particularly for comparative cereal genomics and functional genomics. The web implementation of dbWFA provides an easy-to-use interface to annotate transcript sequences from *T. aestivum*, with functional information from multiple pervasive annotation systems. Here, the use of dbWFA is illustrated with several query examples, and the quality of the annotation method is assessed by comparing the MapMan bin annotation of all the transcripts of the *T. aestivum* NimbleGen 40 k microarray (37) with that of *A. thaliana* and *O. sativa*. The use of dbWFA is further illustrated by analysing the annotation of 433 genes specifically expressed during either the early cell division or the late storage polymer accumulation (SPA) phases of grain development.

Data Content, Database Architecture and Web Interface

Five functional classification/annotation systems were integrated (Figure 1) in dbWFA to offer a fast and efficient functional annotation tool for *T. aestivum* UniGenes:

- GO (http://www.geneontology.org) (38), a non-redundant structured hierarchy of ontologies, which is the most widely used functional annotation system in bioinformatics. The GO project provides an efficient annotation standard that can be applied to numerous species. It is built on a controlled vocabulary of terms for describing gene function. dbWFA includes GO annotation data (OBO version 1.2) for *A. thaliana* and *O. sativa*.

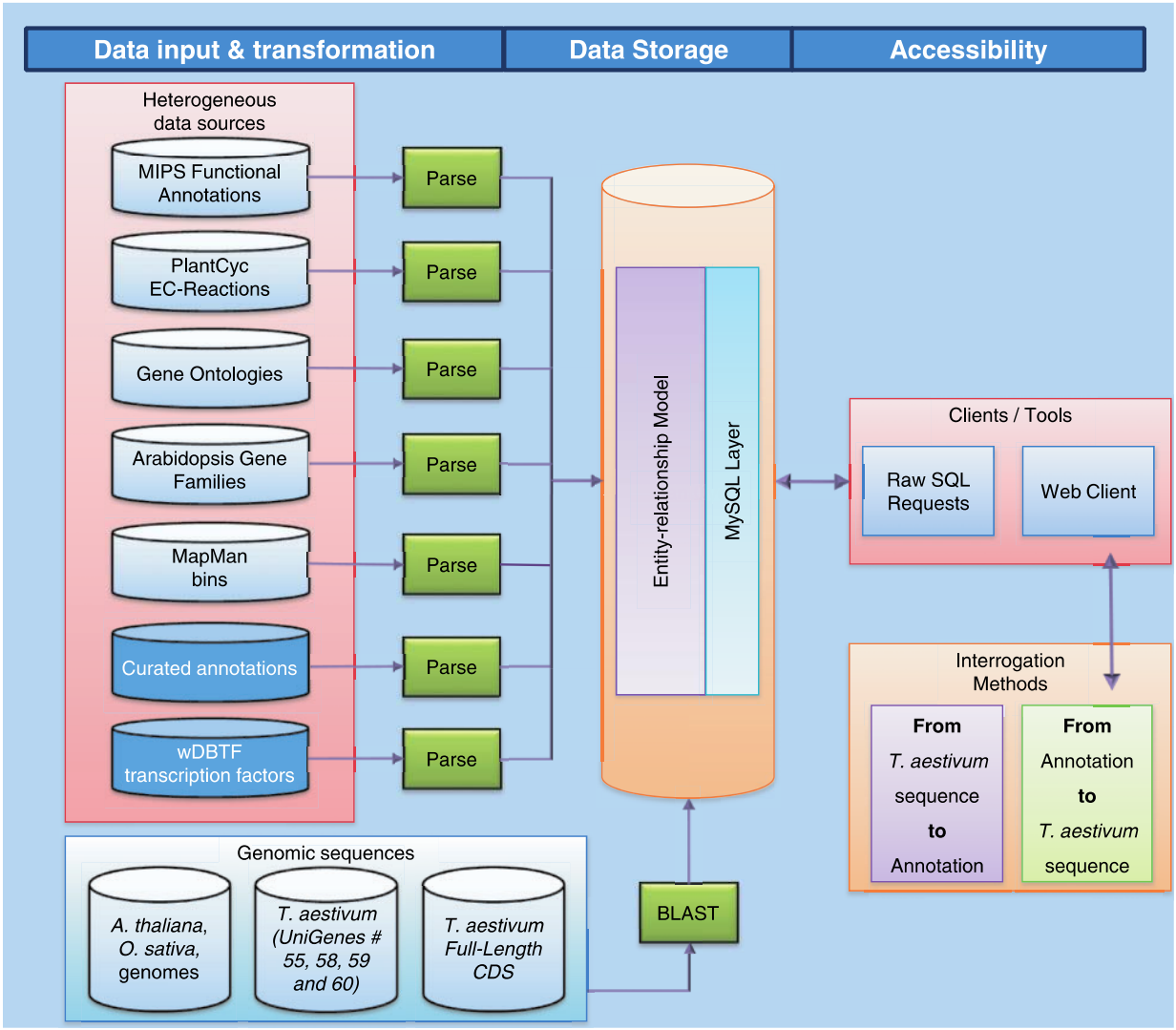


Figure 1. Simplified diagram of the data integration process.

Database tool

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

- Plant Metabolic Network (PMN; http://www.plantcyc.org) (39), which provides a broad network of curated databases on primary and secondary plant metabolism, including pathways, enzymes, genes, compounds and reactions from several plant species. dbWFA contains data from AraCyc (version 9.0) for *A. thaliana* and RiceCyc (version 3.2) for *O. sativa*.
- MapMan (http://mapman.gabipd.org) (40), which is a user-driven tool for large datasets (e.g. gene expression data from microarrays) visualized in the context of diagrams of metabolic pathways or other processes. MapMan annotation data (bin tree version 1.1) for both *A. thaliana* and *O. sativa* are stored in dbWFA. The dbWFA database also provides a function to automatically generate MapMan *T. aestivum* mapping files.
- Munich Information Center for Protein Sequences Functional Catalogue (MIPS FunCat; http://www.helmholtz-muenchen.de/en/mips/projects/funcat) (41), which provides a hierarchical scheme for the functional description of proteins of prokaryotic and eukaryotic origin. MIPS FunCat annotations for *A. thaliana* (MAtDB version 2.1) are stored in dbWFA.
- *A. thaliana* Gene Family Information (TAIR version 10; http://www.arabidopsis.org/browse/genefamily) (42), which provides gene family information for the plant model species *A. thaliana*.

The 17 541 full-length cDNA sequences from the TriFLDB and other public databases and all the transcript sequences from the *T. aestivum* UniGene set (builds #55, #58, #59 and #60) were processed using the BLASTx algorithm against *A. thaliana* and *O. sativa* predicted cDNA sequences (Figure 1). Build #55 (the one used to develop the *T. aestivum* NimbleGen 40 k microarray) (37) and the following major releases were retained, as users may have developed resources based on different builds of the UniGene even though NCBI only stores the most recent build. BLAST results with an e-value $>10^{-3}$ were not stored in the database, as we considered this would be too poor a match for most research. No other filter was applied to the BLAST results before their insertion into the database. All the parameters from the BLAST tabular results were kept, and $>30 \times 10^6$ BLAST results for the UniGenes and 95×10^6 BLAST results for the ESTs shaping the UniGene clusters were stored, so they could be rapidly screened when querying the database.

The database also contains curated information on *T. aestivum* transcription factors (2891 transcripts), E3 ubiquitin ligases of the ubiquitin-proteasome system (876 transcripts), hormone-responsive genes (467 transcripts) and seed storage proteins (55 transcripts; Figure 1). Transcription factor UniGenes were retrieved from the wDBTF database (34). E3 ligase and hormone-responsive UniGenes were recovered from the NCBI and TAIR databases using all *A. thaliana* and *O. sativa* E3 ligase and hormone-responsive sequences as the query in homology searches using the BLASTn, BLASTx and tBLASTx programs (36). The BLAST hits were filtered using an e-value threshold of 10^{-5} and an alignment length exceeding 80 bp. All sequences were checked for consistency and for the presence of specific protein signatures using the InterProScan program (http://www.ebi.ac.uk/Tools/pfa/iprscan/). For seed storage protein UniGenes, homology searches were performed on the whole UniGene build #55, using BLASTx and *T. aestivum* seed storage protein sequences as reference. No preliminary filter was applied to BLASTx results. Instead, all the alignments were carefully examined, and similarity in known conserved critical regions of seed storage proteins was given priority over e-value and BLAST score alone. In dbWFA, curated UniGene annotations are assigned to *T. aestivum* transcripts without any intermediate BLAST result.

Following the recommendations of the International Wheat Genome Sequencing Consortium (IWGSC) for annotating *T. aestivum* genomic sequences (3), the percentages of coverage (with respect to the length of the orthologous proteins) and identity are used to assign functional annotations to a transcript. In dbWFA, users can define the value of these two parameters, but we strongly recommend using the cutoff values suggested by the IWGSC, where BLAST results with an identity >45% and coverage >50% are assigned a 'putative function' and BLAST results with identity and coverage >90% are assigned a 'known function'.

All data are stored in a MySQL database. The integration of the database allows one to assign the functional annotation from any of the systems described above to the transcripts of interest and vice versa. The dbWFA database thus provides a very powerful resource for the annotation of *T. aestivum* UniGenes. To find the most commonly sought types of information from dbWFA, simple yet pertinent queries with their parameters can be sent through a web-based interface (Figure 2). The results are delivered as html pages, and an export procedure is available to retrieve data in spreadsheet. The html result pages provide links redirecting the user to websites of the different annotation systems, allowing a global analysis of the annotation results. The web interface can also be used to automatically create MapMan mapping files for the search results. Although the dbWFA web interface only allows data mining of common queries, specific queries can be performed using the SQL database, which can be downloaded from the dbWFA website. The modularity of the database will facilitate the integration of new *T. aestivum* data as transcripts are sequenced and annotated through different pipelines.

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014 **Database tool**

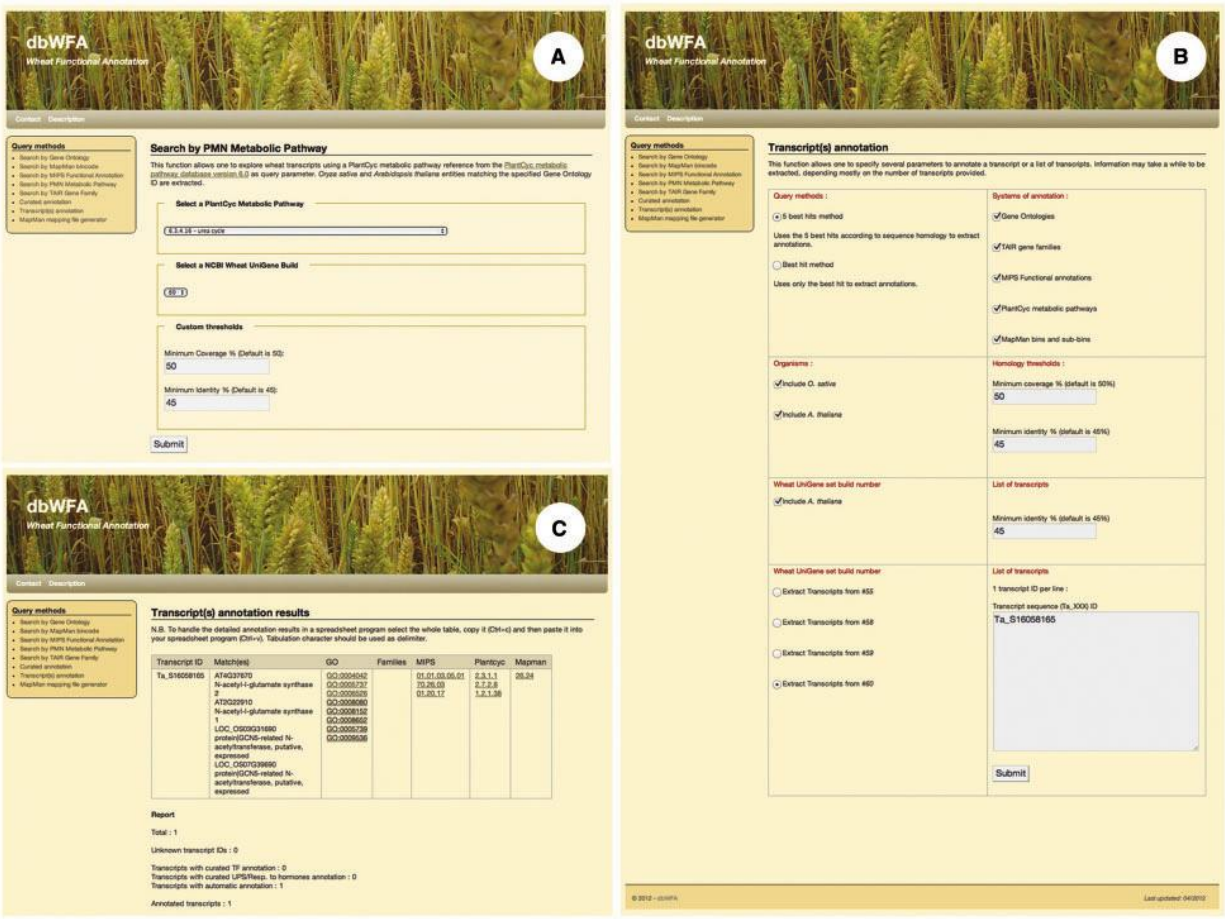


Figure 2. Screen capture of the web interface of the dbWFA database. **(A)** Page for querying PMN pathways. Similar pages can be used to query the MIPS Functional Category, TAIR gene families, GO and MapMan bins. A list of GO can be queried simultaneously. **(B)** Page for querying UniGene or Full-length cDNA sequences annotations. **(C)** Result page for annotated UniGenes.

Using dbWFA: Percentage of Annotated UniGenes, Comparison of *T. aestivum* UniGene and *A. thaliana* and *O. sativa* Whole-Genome Annotation and Query Examples

Thirty-four percent (13 713 transcript sequences), 40% (14 843), 35% (20 016) and 35% (20 034) of the transcript sequences of the UniGene builds #55, #58, #59 and #60, respectively, have a putative functional annotation in at least one of the annotation resources. Eighty-one percent of the 17 541 full-length cDNA sequences from TriFLDB have a putative functional annotation in at least one of the annotation resources. The number of transcripts and full-length cDNA sequences annotated in the different resources are given in Table 1. BLASTn analysis revealed that 12 478 full-length cDNA sequences matched a sequence in the UniGene set (build #60) with a coverage and identity threshold value >50 and 90%, respectively. Among these 12 478 correspondences, 10 996 and 5 932 full-length cDNA sequences and UniGene sequences, respectively, have a putative functional annotation in at least one of the annotation resources. This result highlights the additional information brought by the full-length cDNA sequences.

The quality of the annotation method is illustrated by comparing the MapMan bin annotation of all the transcripts of the *T. aestivum* NimbleGen 40 k microarray (developed with UniGene build #55) and the full-length cDNA sequences from TriFLDB with the annotation of *A. thaliana* and *O. sativa* imported from MapMan and recorded in the database. The MapMan bins were used here because this annotation system is available for the three species. Overall, there was no clear bias between the three species (Figure 3), and the percentages of genes

Database tool

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

Table 1. Number of *T. aestivum* transcripts from the NCBI UniGene set (build #60) and full-length cDNA (FL cDNA) sequences retrieved from the TriFLDB database, annotated with a putative function (coverage >50%, identity >45%) in at least one annotation system

Functional annotation systems	Number of annotated transcripts					
	O. sativa		*A. thaliana*		Total[a]	
	NCBI UniGene	FL cDNA	NCBI UniGene	FL cDNA	NCBI UniGene	FL cDNA
MIPS functional classification			12 943	10 864	12 943	10 864
PlantCyc pathway reactions	2193	2106	2093	2208	3067	2911
GOs	13 142	8014	10 444	10 850	16 079	12 279
TAIR *A. thaliana* gene families			4498	3797	4498	3797
MapMan bins	19 248	14 032	13 202	10 897	20 033	14 224
Curated pathways or functions						
Hormone-responsive genes					467	
Ubiquitin-proteasome system					876	
Transcription factors					2891	

[a]Number of transcripts and full-length cDNA sequences annotated with a putative function in at least one model species.

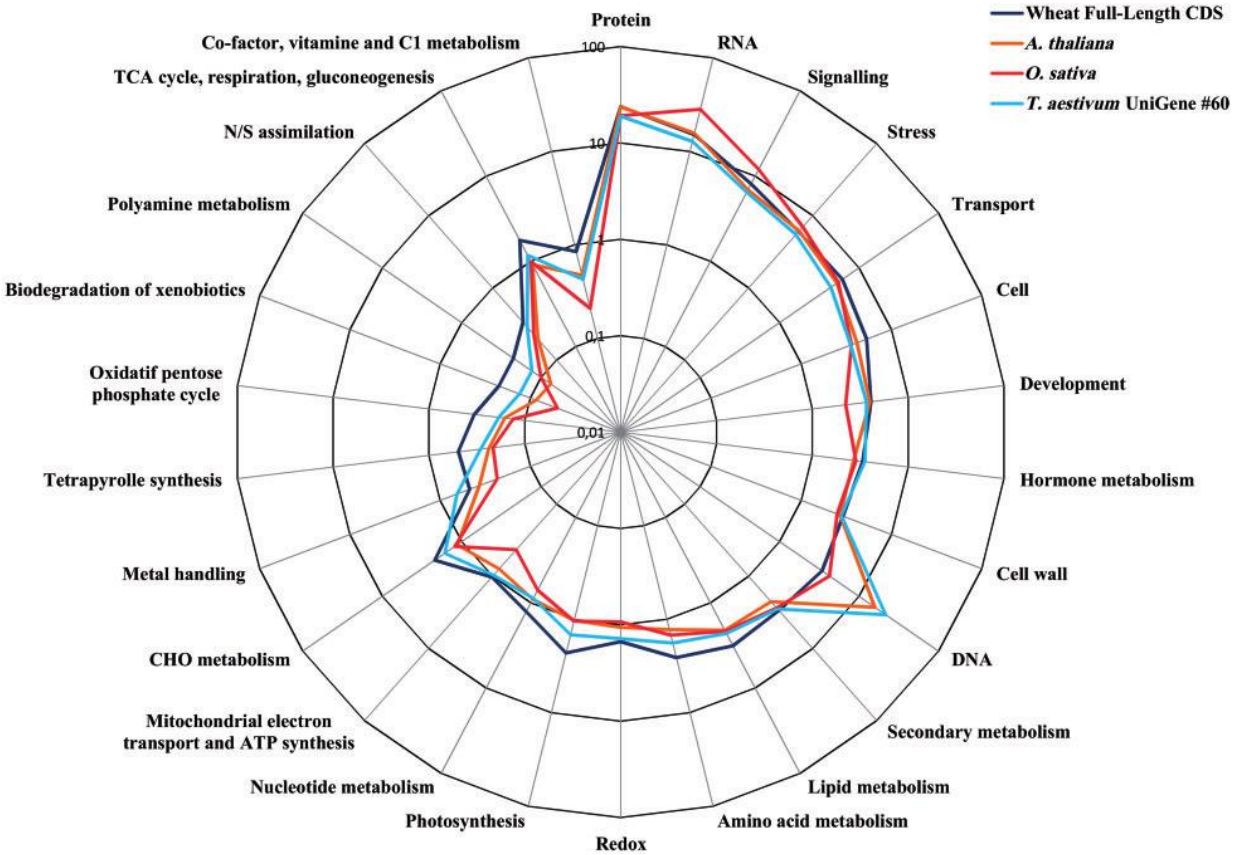


Figure 3. Radar plot (log scale) of the MapMan bin annotations for *A. thaliana*, *O. sativa* and *T. aestivum* UniGene (build #60) and full-length coding sequences. Data are percent of the total number of MapMan bin annotations (Table 1). Similar results were obtained with builds #55, #58 and #59 (data not shown). Some bins have been merged to make the figure clearer.

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

in the 26 categories for the three species were well correlated (*T. aestivum* versus *A. thaliana*: $r = 0.96$, $P < 0.001$; *T. aestivum* versus *O. sativa*: $r = 0.69$, $P < 0.001$), with no significant bias ($P < 0.001$). The higher correlation found with *A. thaliana* compared with *O. sativa* is mainly because there are fewer annotated transcripts in the DNA bin for *O. sativa* than for *A. thaliana* and *T. aestivum* ($r = 0.90$ for *T. aestivum* versus *O. sativa* when this bin is not considered). For the full-length coding sequences retrieved from TriFLDB and other public databases, the correlations between *T. aestivum* and *A. thaliana* and between *T. aestivum* and *O. sativa* were the same ($r = 0.90$, $P < 0.001$). The pairwise correlations between the four MapMan bin annotations presented were remarkably high, all >0.9 when the DNA bin was omitted. Similar results were obtained for the PlantCyc pathway reactions and GO (data not shown).

Unlike many annotation tools, dbWFA makes it possible to query multiple annotation systems simultaneously. To demonstrate various features of the dbWFA database, some query examples are presented in Box 1, using either the website or the database installed on a local machine.

Identification and Annotation of UniGenes Specifically Expressed During Either the Early or Late Stage of Grain Development

A total of 39 029 transcripts from the UniGene set (build #55) and 1613 transcription factors from the wDBTF database not present in the UniGene set are spotted on the custom *T. aestivum* NimbleGen 40 k microarray (36). Previous studies have shown that 18 140 (44.6%) of these transcripts are expressed during *T. aestivum* grain development (47). In dbWFA, 34–40% (depending on the build) of these transcripts have a putative functional annotation.

Box 1. Query Examples

To demonstrate the usefulness of dbWFA, several biologically relevant queries that can be performed using the current system are presented. In these examples, the UniGene build #55 was used, with coverage and identity thresholds of 50 and 45%, respectively, as recommended by the IWGSC to assign a putative function to a transcript.

Query 1. *Find all* T. aestivum *transcripts likely to have a phytoene synthase activity*

UniGene			Matching sequences		Alignment parameters	
Id number	Representing sequence	Description	Id number	Description	Coverage (%)	Identity (%)
Ta.41960	Ta_S16057905	*T. aestivum* clone wr1.pk0139.g3:fis, full insert mRNA sequence	LOC_OS06G51290	Phytoene synthase, chloroplast precursor, putative, expressed	59.7	81.4
			AT5G17230	Phytoene synthase	58.0	79.6
Ta.66029	Ta_S26027774	FGAS000498 *T. aestivum* FGAS: Library 2 Gate 3? *T. aestivum* cDNA, mRNA sequence	LOC_OS06G51290	phytoene synthase, chloroplast precursor, putative, expressed	55.3	48.9
			AT5G17230	Phytoene synthase	59.7	47.08

The first committed step in the biosynthesis of carotenoids is the condensation of two geranylgeranyl disphosphate molecules by phytoene synthase to produce phytoene, which catalyses a rate-controlling step in the plastid-localized carotenoid pathway (43). We could query the database for the PlantCyc pathway reaction 2.5.1.32 using its web interface. The result of this query is shown in the above table. Two *T. aestivum* transcripts were annotated with a putative phytoene synthase activity. In good agreement with this result, previous studies showed that Poaceae species possess a duplicated phytoene synthase gene (44). A thorough analysis of the two annotated UniGene sequences confirmed that they correspond to the duplicated phytoene synthase gene found in Poaceae. A third phytoene synthase has been isolated in *Z. mays* and *T. aestivum* (45, 46). Although the three *O. sativa* phytoene synthase genes are present in the database, the *T. aestivum* UniGene of this phytoene synthase gene was not found in dbWFA.

The phytoene synthase activity also corresponds to GO:0016767 MapMan bin 16.1.4.1. Searching dbWFA for this GO or MapMan bin yields the same results as above. It is possible to combine several overlying systems (e.g. PlantCyc pathway reaction and GO) in a single MySQL query when the database is installed on a local machine. It is also possible to compare and make the union or intersection of queries using MySQL, depending on the intended outcome.

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

Box 1: Continued

Query 2. *Find as much information as possible about a list of transcripts*

UniGene		GO	TAIR	MIPS	PlantCyc	MapMan
Id number	Match					
Ta.41960	**AT5G17230** Phytoene synthase	GO:0009507 GO:0016117 GO:0016767 GO:0046905		01.06.06.13 70.26.03	2.5.1.32 2.5.1.32	16.1.4.1

The efficiency of the database stems from its multiple systems of annotation. The cross-system annotation feature of dbWFA is integrated in the web interface in the 'Transcript(s) annotation' search method. This type of query could be used to obtain information for a list of UniGenes of interest in the different annotation systems integrated in dbWFA. Querying the UniGene set for the first phytoene synthase transcript retrieved in Query 1 yields the annotation shown in the above table. On the web interface, the user can choose to display only the best hit (as in the above table) or the five best hits with percentages of coverage and identity greater than the thresholds set by the user. The user can also choose the systems of annotation to include in the query and the model species. The results redirect the user to the web pages of the different annotation systems, which allows more detailed information to be obtained on the annotation of the list of transcripts of interest.

Query 3. *Find all the transcripts putatively involved in the glycolytic pathway for a transcriptome analysis in MapMan*

Bin code	Name	Identifier	Description	Type
4.1	Glycolysis.cytosolic branch	Ta_S16058223	Similar to UTP–glucose-1-phosphate uridylyltransferase, putative, expressed Coverage: 99.5745%, identity: 92.75%	T
4.1.10	Glycolysis.cytosolic branch.non-phosphorylating glyceraldehyde 3-phosphate dehydrogenase (NPGAP-DH)	Ta_S13048872	Similar to aldehyde dehydrogenase Coverage: 100%, identity: 87.1%	T
4.1.10	Glycolysis.cytosolic branch.non-phosphorylating glyceraldehyde 3-phosphate dehydrogenase (NPGAP-DH)	Ta_S13048873	Similar to aldehyde dehydrogenase Coverage: 100%, identity: 79.23%	T
4.1.11	Glycolysis.cytosolic branch.aldolase	Ta_S15902802	Similar to aldolase superfamily protein Coverage: 50.1873%, identity: 85.07%	T
4.1.11	Glycolysis.cytosolic branch.aldolase	Ta_S17888674	Similar to aldolase superfamily protein Coverage: 88.5475%, identity: 48.91%	T

In the search method 'MapMan mapping file generator', the user can select a metabolic pathway and automatically create a mapping file to visualise the results of transcriptomic experiences performed with the *T. aestivum* custom NimbleGen 40 k microarray using the -omic data viewing and analysing tool MapMan. The glycolytic pathway corresponds to the bin code 4. The first five lines of the table generated by dbWFA for this query are shown above. When the database is installed on a local machine, several pathways could be queried simultaneously to create a custom *T. aestivum* mapping file for MapMan.

T. aestivum grain development comprises several distinct phases, starting with a syncytial then a cellularization phase (ca. 0–100°Cdays after anthesis), followed by a first differentiation phase of active endosperm cell division (ECD), expansion and differentiation (ca. 100–250°Cdays after anthesis), a second differentiation phase when storage polymers rapidly accumulate (ca. 250–750°Cdays after anthesis) and a maturation phase when grain rapidly desiccates (ca. 750–900°Cdays after anthesis) (48, 49). The transitions between these phases are associated with major changes in the grain transcriptome (5, 36, 50, 51) and proteome (52, 53).

To validate the principle underpinning the database and provide another example of the usefulness of dbWFA, we analysed the functional annotation of the transcripts specifically expressed during either the ECD or SPA phase of grain development. We used transcriptome data obtained with the custom *T. aestivum* NimbleGen 40 k microarray for

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

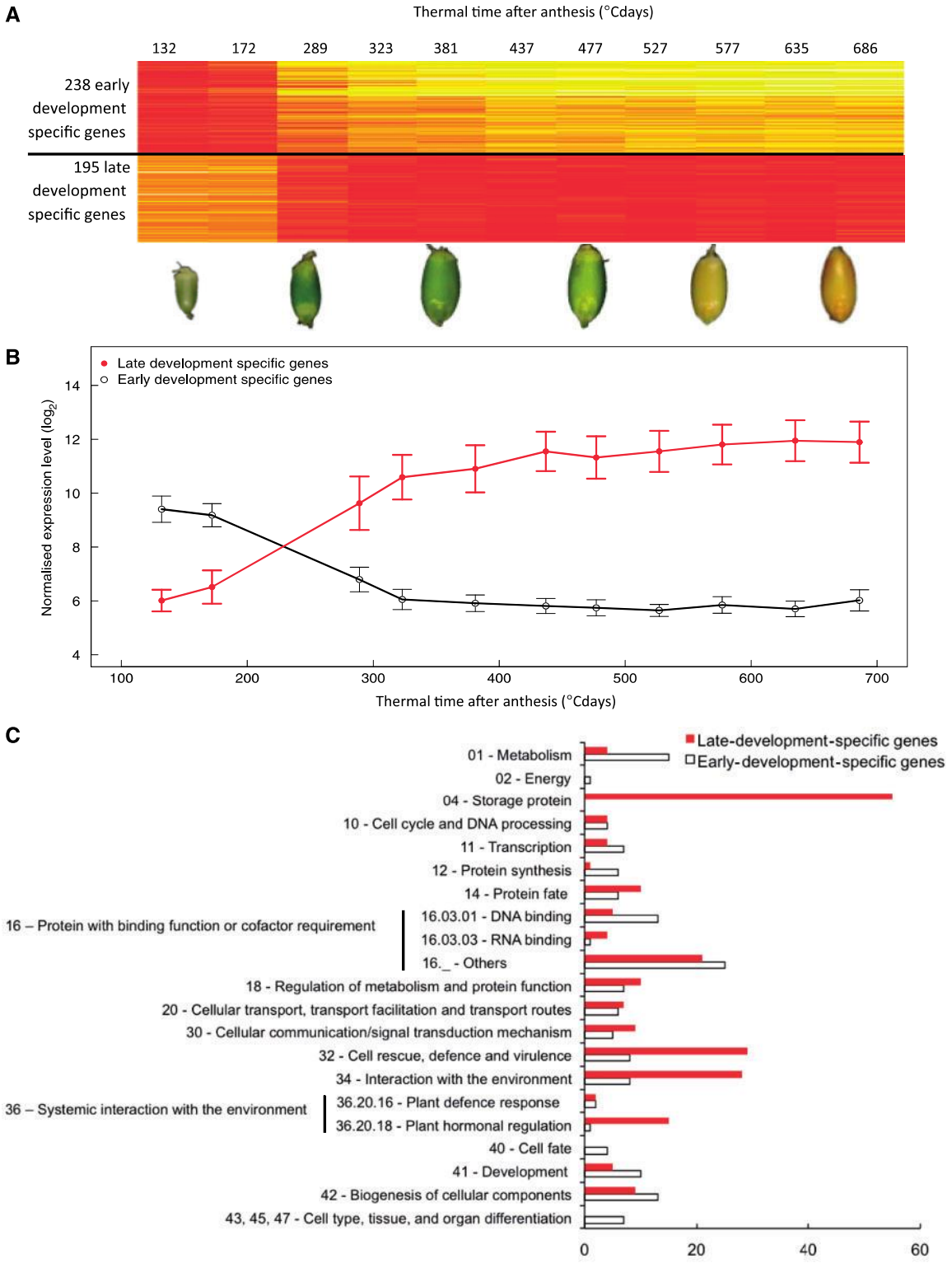


Figure 4. Functional annotation of genes specifically expressed during either the early cell division or late SPA phases of *T. aestivum* grain development. **(A)** Heat map of expression for early- and late-development-specific genes. **(B)** Normalized expression of the early and late development specific gene clusters. Transcripts with normalized expression <7 were not considered to be expressed (i.e. not different from the background noise). Data are medians ± 1 SD. **(C)** MIPS Functional Categories of genes from both UniGene clusters.

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

the *T. aestivum* cultivar Recital grown under standard conditions in a greenhouse and sampled every 34–117°Cdays between 132 and 686°Cdays after anthesis (35). Transcripts with different patterns of expression were classified with J-Express 2012 software package (54) using Euclidean distance–based k-means clustering. For this analysis, the number of clusters was empirically set at 25 because it allowed us to clearly discriminate between gene expression clusters specific to the ECD and SPA phases of grain development. One cluster of 238 genes contained genes expressed exclusively during ECD stages (Figure 4A and B). Two other clusters contained genes expressed exclusively during SPA stages. The latter two clusters were merged to form a single SPA cluster of 195 genes. dbWFA was then used to retrieve the functional classification of the transcripts from both clusters. The MIPS Functional Classification was used, as it was the most informative and straightforward for comparisons with previous studies.

Using IWGSC-recommended coverage (50%) and identity (45%) percentages, 68 (29%) ECD-specific transcripts and 129 (66%) SPA-specific transcripts were assigned to an MIPS functional category, respectively (Figure 4C). The annotation results were consistent with previous transcriptome (5, 51) and proteome (52, 53) studies of developing *T. aestivum* and *H. vulgare* (55) grain. The functional classifications of the gene clusters were different. Not surprisingly, 12 transcripts involved in cell fate and cell type, tissue differentiation and organ differentiation were specifically expressed during the ECD phase, whereas no SPA phase–specific transcripts were annotated as being in these MIPS functional categories. Also in good agreement with our knowledge of grain development, 55 seed storage protein transcripts were specifically expressed during the SPA phase, while none was found among the annotated ECD-specific genes.

Quantitative differences in the annotation of these two clusters of transcripts were also observed. Several transcripts in the SPA-specific cluster were involved in cell rescue, defence and virulence and in the interaction with the environment. In particular, transcripts involved in plant hormonal regulation were overrepresented in the SPA-specific gene cluster. Transcripts coding for proteins involved in protein synthesis and proteins with metabolic functions were overrepresented in the ECD cluster. These results coincide with previous transcriptome (5, 36, 56) and proteome analyses (53). Finally, we note a substantial difference in the MIPS functional category 'protein with binding function or co-factor requirement', with more transcripts involved in DNA binding in the ECD cluster and more transcripts involved in RNA binding in the SPA cluster.

All these data show great similarity to published results for *T. aestivum* and *H. vulgare*, reflecting the accuracy of the automatic annotation provided by dbWFA. Unlike several other *T. aestivum* transcriptome analyses where a complex process had to be carried out to assign a functional annotation to selected transcripts and/or proteins (5, 57), here only a single request was made to the dbWFA database to retrieve the functional classification of 45% of the transcripts of interest, and 40% of the 40 642 transcripts of the *T. aestivum* NimbleGen 40 k microarray. This percentage of annotated transcripts is similar to that previously reported (38%) for the *T. aestivum* Affymetrix GeneChip® microarray (4).

Outlook

The dbWFA database was created by integrating numerous data sources. As a result, it is a practical source of heterogeneous data for functional annotation of *T. aestivum* transcripts. The website grants access to the most common queries that can be applied to the database, and the freely available MySQL database is a powerful tool for more specific requests. Although further analyses are required to confirm the dbWFA annotation results, the database provides an efficient and fast solution for acquiring a wide range of functional information. cDNA resources are useful to predict exonic regions from genomic sequences; thus, efforts to annotate the UniGene resources will significantly contribute to the analysis of sequence data produced by ongoing *T. aestivum* initiatives and other genome sequencing projects.

The version of dbWFA presented here is operational, but the aim is not to restrict the database to storing *O. sativa* and *A. thaliana* annotations but to expand it to include data from other plant species genomes as their functional annotation becomes more consistent. Integration of InterProScan (58) in the workflow could be a valuable way of augmenting the process. Also, the integration of AFAWE (59) would provide an annotation workflow with different function prediction tools. However, the current version of AFAWE cannot be used independently from its web interface and would thus have to be implemented using the tools called in its workflow, which are available as web services. Finally, the upcoming integration of a BLAST program in the workflow will allow users to annotate their own sequences and will also make dbWFA applicable to other species.

Acknowledgements

The authors thank Dr Etienne Paux and Dr Catherine Feuillet (INRA, UMR1095 GDEC, Clermont-Ferrand, France) for useful discussions and advice, and Mr Sébastien Reboux, Ms Claire Viseux and Mr Michael Alaux (INRA, URGI, Versailles, France) for installing and maintaining the database on the URGI server.

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

Funding

This work was supported by a PhD grant from the French Ministry for Higher Education and Research to J.V.

Conflict of interest. None declared.

References

1. Feuillet,C. and Eversole,K. (2007) Physical mapping of the wheat genome: a coordinated effort to lay the foundation for genome sequencing and develop tools for breeders. *Isr. J. Plant Sci.*, **55**, 307–313.
2. Feuillet,C., Leach,J.E., Rogers,J. *et al.* (2011) Crop genome sequencing: lessons and rationales. *Trends Plant Sci.*, **16**, 77–88.
3. Leroy,P., Guilhot,N., Sakai,H. *et al.* (2012) TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. *Front. Plant Sci.*, **3**, 1–14.
4. Brenchley,R., Spannagl,M., Pfeifer,M. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.
5. Wan,Y., Poole,R.L., Huttly,A.K. *et al.* (2008) Transcriptome analysis of grain development in hexaploid wheat. *BMC Genomics*, **9**, 121.
6. Pellny,T.K., Lovegrove,A., Freeman,J. *et al.* (2012) Cell walls of developing wheat starchy endosperm: comparison of composition and RNA-Seq transcriptome. *Plant Physiol.*, **158**, 612–627.
7. Köhler,J., Baumbach,J., Taubert,J. *et al.* (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, **22**, 1383–1390.
8. Lysenko,A., Hindle,M.M., Taubert,J. *et al.* (2009) Data integration for plant genomics–exemplars from the integration of *Arabidopsis thaliana* databases. *Briefings Bioinformatics*, **10**, 676–693.
9. Lazo,G.R., Chao,S., Hummel,D.D. *et al.* (2004) Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. *Genetics*, **168**, 585–593.
10. Zhang,H., Sreenivasulu,N., Weschke,W. *et al.* (2004) Large-scale analysis of the barley transcriptome based on expressed sequence tags. *Plant J.*, **40**, 276–290.
11. Allouis.,S., Moore.,G., Bellec.,A. *et al.* (2003) Construction and characterisation of a hexaploid wheat (*Triticum aestivum* L.) BAC library from the reference germplasm 'Chinese Spring'. *Cereal Res. Commun.*, **31**, 331–338.
12. Safár,J., Bartos,J., Janda,J. *et al.* (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.*, **39**, 960–968.
13. Wilkinson,P.A., Winfield,M.O., Barker,G.L.A. *et al.* (2012) CerealsDB 2.0: an integrated resource for plant breeders and scientists. *BMC Bioinformatics*, **13**, 219.
14. Lai,K., Berkman,P.J., Lorenc,M. *et al.* (2012) WheatGenome.info: an integrated database and portal for wheat genome information. *Plant Cell Physiol.*, **53**, e2.
15. Dong,Q., Schlueter,S.D. and Brendel,V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.*, **32**(Database issue), D354–D359.
16. Feuillet,C., Stein,N., Rossini,L. *et al.* (2012) Integrating cereal genomics to support innovation in the Triticeae. *Funct. Integr. Genomics*, **12**, 573–583.
17. Close,T.J., Wanamaker,S., Roose,M.L. *et al.* (2007) HarvEST. *Methods Mol Biol.*, **406**, 161–177.
18. Estill,J.C. and Bennetzen,J.L. (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods*, **5**, 8.
19. Mochida,K., Yoshida,T., Sakurai,T. *et al.* (2009) TriFLDB: a database of clustered full-length coding sequences from triticeae with applications to comparative grass genomics. *Plant Physiol.*, **150**, 1135–1146.
20. Manickavelu,A., Kawaura,K., Oishi,K. *et al.* (2012) Comprehensive functional analyses of expressed sequence tags in common wheat (*Triticum aestivum*). *DNA Res.*, **19**, 165–177.
21. The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
22. Yu,J., Hu,S., Wang,J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
23. The International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
24. Schnable,P.S., Ware,D., Fulton,R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
25. Schmutz,J., Cannon,S.B., Schlueter,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
26. Paterson,A.H., Bowers,J.E., Bruggmann,R. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
27. The International Brachypodium Initiative. (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
28. Mayer,K.F.X., Waugh,R., Langridge,P. *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.
29. Van Bel,M., Proost,S., Wischnitzki,E. *et al.* (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.*, **158**, 590–600.
30. Dassanayake,M., Oh,D.H., Haas,J.S. *et al.* (2011) The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.*, **43**, 913–918.
31. Conesa,A., Götz,S., Garcia-Gomez,J.M. *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
32. Lamesch,P., Berardini,T.Z., Li,D. *et al.* (2011) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
33. Ouyang,S., Zhu,W., Hamilton,J. *et al.* (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
34. Altschul,S.F., Madden,T.L., Schäffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
35. Romeuf,I., Tessier,D., Dardevet,M. *et al.* (2010) wDBTF: an integrated database resource for studying wheat transcription factor families. *BMC Genomics*, **11**, 185.
36. Capron,D., Mouzeyar,S., Boulaflous,A. *et al.* (2012) Transcriptional profile analysis of E3 ligase and hormone-related genes expressed during wheat grain development. *BMC Plant Biol.*, **12**, 35.
37. Rustenholz,C., Choulet,F., Laugier,C. *et al.* (2011) A 3,000-loci transcription map of chromosome 3B unravels the structural and

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

functional features of gene islands in hexaploid wheat. *Plant Physiol.*, **157**, 1596–1608.

38. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2011) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
39. Zhang,P., Dreher,K., Karthikeyan,A. *et al.* (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.*, **153**, 1479–1491.
40. Thimm,O., Bläsing,O., Gibon,Y. *et al.* (2004) Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.
41. Ruepp,A., Zollner,A., Maier,D. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
42. Rhee,S.Y., Beavis,W., Berardini,T.Z. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
43. Cunningham,F.X. and Gantt,E. (1998) Genes and enzymes of carotenoids biosynthesis in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **49**, 557–583.
44. Gallagher,C.E., Matthews,P.D., Li,F. *et al.* (2004) Gene duplication in the carotenoid biosynthetic pathway preceded evolution of the grasses. *Plant Physiol.*, **135**, 1776–1783.
45. Li,F., Vallabhaneni,R. and Wurtzel,E.T. (2008) PSY3, a new member of the phytoene synthase gene family conserved in the poaceae and regulator of abiotic stress-induced root carotenogenesis. *Plant Physiol.*, **146**, 1333–1345.
46. Dibari,B., Murat,F., Chosson,A. *et al.* (2012) Deciphering the genomic structure, function and evolution of carotenogenesis related phytoene synthases in grasses. *BMC Genomics*, **13**, 221.
47. Romeuf,I. (2010) Identification *in silico* des facteurs de transcription du blé tendre (*Triticum aestivum*) et mise en évidence des facteurs de transcription impliqués dans la synthèse des protéines de réserve, *Ph.D. Thesis*. Université Clermont-Ferrand II, Blaise Pascal, Clermont-Ferrand, France, pp. 223.
48. Bennett,M.D., Rao,M.K., Smith,J.B. *et al.* (1975) Cell development in the anther, the ovule, and the young seed of *Triticum aestivum* L. Var. chinese spring. *Philos. T. R. Soc. B*, **266**, 6–81.
49. Evers,T. and Millar,S. (2002) Cereal grain structure and development: some implications for quality. *J. Cereal Sci.*, **36**, 261–284.
50. Drea,S., Leader,D.J., Arnold,B.C. *et al.* (2005) Systematic spatial analysis of gene expression during wheat Caryopsis. *Plant Cell.*, **17**, 2172–2185.
51. Laudencia-Chingcuanco,D.L., Stamova,B.S., You,F.M. *et al.* (2007) Transcriptional profiling of wheat caryopsis development using cDNA microarrays. *Plant Mol. Biol.*, **63**, 651–668.
52. Nadaud,I., Girousse,C., Debiton,C. *et al.* (2010) Proteomic and morphological analysis of early stages of wheat grain development. *Proteomics*, **10**, 2901–2910.
53. Tasleem-Tahir,A., Nadaud,I., Chambon,C. *et al.* (2012) Expression profiling of starchy endosperm metabolic proteins at 21 stages of wheat grain development. *J. Proteome Res.*, **11**, 2754–2773.
54. Dysvik,B. and Jonassen,I. (2001) J-Express: exploring gene expression data using Java. *Bioinformatics*, **17**, 369–370.
55. Sreenivasulu,N., Radchuk,V., Strickert,M. *et al.* (2006) Gene expression patterns reveal tissue-specific signaling networks controlling programmed cell death and ABA- regulated maturation in developing barley seeds. *Plant J.*, **47**, 310–327.
56. Clarke,B.C., Hobbs,M., Skylas,D. *et al.* (2000) Genes active in developing wheat endosperm. *Funct. Integr. Genomics*, **1**, 44–55.
57. Szucs,A., Jäger,K., Jurca,M.E. *et al.* (2010) Histological and microarray analysis of the direct effect of water shortage alone or combined with heat on early grain development in wheat (*Triticum aestivum*). *Physiol Plant.*, **140**, 174–188.
58. Goujon,M., McWilliam,H., Li,W. *et al.* (2010) A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Res.*, **38**, W695–W699.
59. Jöcker,A., Hoffmann,F., Groscurth,A. *et al.* (2008) Protein function prediction and annotation in an integrated environment powered by web services (AFAWE). *Bioinformatics*, **24**, 2393–2394.

CONCLUSION PUBLICATION 1

La base de données qui a été développée intègre plusieurs sources de données : les données d'annotation disponibles pour *A. thaliana* et le riz comme présentées dans l'article, mais aussi les données d'annotation fonctionnelle du maïs. Les homologies entre les séquences de ces organismes et les différentes séquences de blé tendre (UniGene et TriFLdb) sont également stockées dans la base de données, permettant de lier rapidement les séquences de blé aux annotations fonctionnelles disponibles dans les différents systèmes d'annotation utilisés. Par conséquent, cette base de données représente une ressource adaptée à l'annotation fonctionnelle de ces séquences par homologie. La base de données fournit une solution efficace et rapide pour obtenir un large éventail d'informations fonctionnelles et peut contribuer de manière significative à l'analyse des données de séquences produites dans les projets de transcriptomique appliqués au blé tendre.

L'intégration des quatre systèmes d'annotation fonctionnelle présentés et des FT (Figure I-1) permet d'interroger la base de données de deux façons distinctes. D'une part elle permet d'annoter facilement une liste de séquences de blé, d'autre part elle permet, à l'inverse, de trouver une liste de séquences de blé possédant une fonction commune. Des requêtes prédéfinies sont disponibles sur une page web dédiée à l'interrogation de la base de données (Figure I-2) tandis que les requêtes plus complexes peuvent être effectuées directement par le biais de MySQL.

La qualité de l'annotation fournie par cette méthode a été évaluée par plusieurs analyses. Tout d'abord par la comparaison de l'annotation MapMan de toutes les séquences de blé présentes dans la base de données, à savoir l'UniGene version 60 et TriFLdb comme présenté dans la Figure I-3 mais aussi avec les version 58, 59 et 60 de l'UniGene (données non présentées) et avec l'annotation des séquences d'*A. thaliana*, de riz (Figure I-3) et de maïs (données non présentées). Les résultats suggèrent qu'il n'existe pas de biais fonctionnel significatif entre les différents organismes et les pourcentages de gènes dans les différentes catégories fonctionnelles pour les trois espèces sont bien corrélés. Des résultats similaires ont été obtenus avec l'utilisation des autres systèmes d'annotation (données non présentées).

Ensuite, afin de valider l'objectif affiché de la base de données, nous avons utilisé les données transcriptomiques obtenues sur une cinétique de développement du grain (132 à 686°Cj après floraison). Nous avons analysé l'annotation fonctionnelle des transcrits sur-exprimés soit pendant la phase précoce de développement du grain de blé, soit pendant la phase d'accumulation des PR. Un groupe de 238 gènes contenait les gènes sur-exprimés pendant la

phase précoce de développement du grain tandis qu'un autre groupe de 195 contenait les gènes sur-exprimés pendant la phase d'accumulation des réserves (Figures I-4A et I-4B). La base de données a ensuite été utilisée pour obtenir l'annotation fonctionnelle des gènes de ces deux groupes.

L'utilisation des seuils d'homologie recommandés par l'IWGSC (*International Wheat Genome Sequencing Consortium*) a permis d'annoter 68 (29%) transcrits spécifiques de la phase précoce de développement du grain et 129 (66%) transcrits spécifiques de la phase d'accumulation des réserves (figure I-4C). Les résultats obtenus, qui coïncident avec des résultats obtenus chez d'autres espèces proches permettent à la fois de valider l'annotation mais aussi et surtout de préparer les analyses pour la suite du travail de thèse.

L'objectif de la base de données n'est pas de se limiter aux organismes déjà intégrés mais de s''étendre pour inclure des données provenant d'autres génomes d'espèces végétales dont les données d'annotation fonctionnelle s'accumulent afin de fournir des résultats plus consistants. Enfin, l'intégration d'un outil de recherche d'homologie (*BLAST*) directement dans le workflow de la base de données permettra aux utilisateurs d'annoter leurs propres séquences et rendra également dbWFA utilisable pour d'autres espèces.

L'inférence de réseaux dans cette thèse devant être réalisée à l'échelle du transcriptome, ou dans certains contextes, à l'échelle de quelques centaines ou milliers de gènes, la base de données permettra d'identifier rapidement les groupes de gènes jugés intéressants dans la suite du travail.

PUBLICATION 2

RulNet: A web-oriented platform for regulatory network inference, application to wheat –omics data

Jonathan Vincent[1,2,3], Pierre Martre[1,2], Benjamin Gouriou[3], Catherine Ravel[1,2], Zhanwu Dai[1,2,†], Jean-Marc Petit[4] and Marie Pailloux[3]

[1]INRA, UMR1095 Génétique, Diversité et Ecophysiologie des Céréales, 5 chemin de Beaulieu, F-63 039 Clermont-Ferrand, France; [2]Université Blaise Pascal, UMR1095 Diversité et Ecophysiologie des Céréales, F-63 177 Aubière, France; [3]Université Blaise Pascal, UMR6158 CNRS LIMOS Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, F-63 173 Aubière, France; [4]INSA Lyon, UMR5205 CNRS LIRIS Laboratoire d'Informatique en Images et Systèmes d'Information, Villeurbanne, F-69 621, France; [†]Current address: INRA, UMR1287 Ecophysiologie et Génomique Fonctionnelle de la Vigne, Institut des Sciences de la Vigne et du Vin, F-33 882 Villenave d'Ornon, France.

ABSTRACT

Concurrently with the increasing amount of –omics data available, a particular effort to provide suitable analysis tools has to be made. Unraveling molecular regulations from massive heterogeneous data is therefore a major challenge. Here we describe RulNet, a web-oriented platform dedicated to the inference and analysis of regulatory networks from qualitative and quantitative –omics data by means of rule discovery. Queries for rule discovery can be written in an extended form of the RQL language, a textual query syntax similar to SQL. RulNet also offers users interactive possibilities to progressively adjust and refine the inferred networks. Here, we carried out a functional characterization of RulNet and inferred networks were compared with correlation-based approaches. Then the performance of RulNet was evaluated using the three benchmark datasets used for the transcriptional network inference challenge DREAM5. Overall, RulNet ranked among the best methods that participated to this challenge and unlikely other methods it ranked equally well for the three datasets. Finally, the elucidation of regulatory networks involved in the response to nitrogen and sulfur supply in wheat (Triticum aestivum L.) grains allowed assessing the suitability of RulNet to analyze experimental –omics data. Results highlight putative actors governing the response to nitrogen and sulfur supply in wheat grains. We introduce the main characteristics and features.

INTRODUCTION

Regulation of gene expression is defined as the spatiotemporal control of the amount of gene products and allows cell differentiation and efficient adaptation of living organisms to their environment. Out of the diverse levels of regulation, transcriptional control is onsidered crucial and is conserved in eukaryotes and prokaryotes. Regulatory sequences upstream of genes as well as proteins able to recognize and bind specifically to these sequences, called transcription factors (TFs), were substantiated in the late 1960's (Miller et al., 1968). TFs are themselves encoded by genes and the complex interactions between genes and TFs put together are called Gene Regulatory Networks (GRNs). GRNs have been inferred efficiently in the past decade, leading to significant breakthroughs (Shen-Orr et al., 2002; Yoon et al., 2009; Belcastro et al., 2011; Pan et al., 2013).However, GRNs might be considered too restrictive to elucidate the complex physiology of a living system. Cell physiology is now addressed more broadly and the overall regulatory relationships between genes, their products and the various molecules present in a cell are called Regulatory Networks (RNs).

Inference of RNs is accordingly a convenient approach to elucidate the complexity of molecular interactions. Existing RN inference approaches can be categorized according to the strategy developed (De Smet & Marchal, 2010) as supervised (Mordelet & Vert, 2008) or unsupervised (Margolin *et al.*, 2006), i.e. whether learning samples should be provided or not, and global or query driven, i.e. whether all or only a portion of the hypothetic relationships are inferred. Query driven approaches allow prioritizing predictions likely to be valuable in the context of a research. These approaches can be further divided according to the data mining or statistical methods they rely on to infer interactions. The most widely used approaches are Boolean networks (Kauffman *et al.*, 2003; Liang & Han, 2012), Bayesian networks (Friedman *et al.*, 2000), Petri nets (Küffner *et al.*, 2000; Durzinsky *et al.*, 2011) and association rule discovery(Georgii *et al.*, 2005; Agier *et al.*, 2007; Martinez *et al.*, 2008).

Several tools based on association rule discovery have been developed. For instance, GenMiner (Martinez *et al.*, 2008) allows the simultaneous analysis of biological data sets such as gene expression or annotation, but quantitative data are preprocessed and discretized in the workflow. Georgi *et al.* (Georgii *et al.*, 2005) also presented an approach for association rule discovery that uses quantitative data and allows users to specify parameters to discover rules matching their personal interest. These applications do not provide an integrated tool for the visualization and edition of inferred networks. Only a few web-oriented platforms are available to infer RNs. GENIES (Kotera *et al.*, 2012) for instance provides a supervised method

to construct networks using partially known network information. It allows loading heterogeneous data but lacks visualization features for resulting networks. Predictive Networks (Haibe-Kains *et al.*, 2012) is an all-in-one web-oriented platform with visualization and analysis tools. Its workflow consists of a data- and text-mining pipeline combined with a seeded Bayesian network inference method. Literature mining can provide valuable information for model species but are less efficient when applied to less studied species.

RulNet, presented herein relies on an integrative, query driven approach using a data mining method extended from Agier *et al.* (Agier *et al.*, 2007). It is implemented as a web-oriented platform dedicated to the inference of RNs suitable for qualitative and quantitative –omics data. It provides users with the possibility to specify entities (genes, metabolites,…) of interest to discover rules involving these entities. A specific query language was integrated to RulNet allowing users to elaborate and perform custom queries. Each type of rules has its own biological meaning, thus providing users with different information. The approach can also be used to infer very distinct networks, e.g. between traits and genetic markers in association studies. This all-in-one platform allows uploading experimental data or interrogating data directly in a database management system (DBMS) and discovering different types of rules in a unifying framework. The platform has several functionalities such as validation or invalidation of interactions, network generation from one or two lists of components, parameterization of the number of intermediate components. The ultimate aim of RulNet is to provide biological researchers with more control over their search for interactions.

In order to assess the efficiency and suitability of RulNet to infer RNs, we first benchmarked it using gold standard datasets used for the transcriptional network inference challenge DREAM5 (Marbach *et al.*, 2012). Overall, RulNet ranked among the best methods that participated to this challenge and unlikely other methods it ranked equally well for the three datasets. Then we used RulNet to analyze changes in gene expression, grain storage protein (GSP) composition, free amino acids and key metabolites concentrations in response to nitrogen (N) and sulfur (S) supply for developing wheat (Triticum aestivum L.) grain. Wheat GSPs accumulate during the effective filling phase of grain development and account for 60% to 80% of the total protein content of mature grains (Gupta et al., 1996). Increased N supply via the use of fertilizers increases grain yield and protein concentration (Triboi et al., 2000), but also modifies the balance between GSP fractions, i.e. gliadins and glutenins, the most important single trait determining wheat end-use value (Gupta et al., 1989; Martre et al., 2003). Concurrently, S is essential for an efficient use of N and to maintain balanced GSP

composition (Wrigley et al., 1984; Shewry et al., 2000). The RN inferred highlights genes responsive to N and S supply as well as their links with various GSP fractions and metabolites.

RulNet is publicly available at http://rulnet.isima.fr. Users can also download and run the application on their own servers or computers.

METHODS

Multipredicate rules and RQL language

The notion of rules or implications is widely used in the database, data mining and artificial intelligence communities. A rule reflects an observation on the data. Examples of rules are functional dependencies (Armstrong, 1974), implications (Ganter & Wille, 1997) or association rules (Agrawal *et al.*, 1993). A rule is written $X \rightarrow Y$ and is read *X* implies *Y*, where *X* and *Y* are two sets of attributes (i.e. entities). A rule means that when a property (called a predicate) is observed on *X* (the left hand side or antecedent) then a property is observed on *Y* (the right hand side or consequent) in the data. The two predicates can be of the same type but not necessarily.

The principle of the approach is to indicate which type of rules (semantics) to discover then to apply algorithms to infer all the rules satisfying this semantics in the data. The biologists then have to interpret the set of rules inferred and determine if they have a biological meaning. One may also be interested in approximate rules, i.e. rules that are not always satisfied in the data. This is especially true for biological data since they usually have some degree of uncertainty (noise or error).

A generic SQL-like language called RQL (Rule Query Language) has been recently developed to specify in a simple way, rule semantics (Chardin *et al.*, 2013). RQL allows defining association rules or functional dependencies on a relation but can also capture rules between attributes from different tables of a database, or conditional rules such as conditional functional dependencies (Bohannon *et al.*, 2007; Fan *et al.*, 2008). However, RQL is limited to rules with single predicate (i.e. the same condition on all the attributes of the left hand side implies the same condition on all the attributes on the right hand side). For instance, for exact association rules from a binary relation, the predicate is that the attribute has to be equal to 1 ($AB \rightarrow C$ if for all samples, when *A* = 1 and *B* = 1, then *C* = 1). For functional dependencies, the predicate is the equality between two values ($AB \rightarrow C$ if for all couples of samples with a same value for *A* and a same value for *B*, then *C* has a same value).

This restriction to a single predicate may not allow reflecting all the behaviors in the data, in particular for biological data. Moreover, if users want to treat different types of attributes (quantitative, categorical or binary) a single predicate may not be used for all attributes. Therefore, we extended RQL where multipredicate rules (MP-rules) can be defined and generated. Each predicate can have a different meaning corresponding to a specific objective for the application domain and the type of attributes. The advantage of MP-rules, is that we can capture different perspectives from the data. The advantage is also that we can treat quantitative, categorical or binary attributes in the same process.

To illustrate the use of MP-rules, let us consider a data set composed of three tables (Fig. 1). In this data set, each time *A*1 = 1 in table ann then *M*2 > 0.5 in table met. This observation can be represented by the MP-rule $A1_1 \rightarrow M\ 2$, which means that *A*1 = 1 guarantees *M*2 > 0.5 (i.e. *A*1 = 1 is a sufficient condition to observe *M*2 > 0.5). Another exact MP-rule in this data set is $G\ 3 \rightarrow A1_0$ (i.e. if *G*3 > 0.5 then *A*1 = 0). Here *A*1 = 0 is a necessary condition to observed *G*3 > 0.5.

In RulNet, a query for MP-rules has the following template:

```
FINDRULES
  SCOPE t1 IN (dataset1), t2 IN (dataset2), …
  WHERE condition(t1, t2, …)
  HAVING label1: predicate1(t1, t2, …) OVER attributes1
  AND label2: predicate2(t1, t2, …) OVER attributes2
  AND …;
```

where the "SCOPE" line defines the tuples (rows) of the data sets on which the "HAVING" clause is tested. Data sets can be data files, database tables or SQL queries. The "HAVING" and "AND" lines define the predicates and the associated lists of attributes, respectively. The "WHERE" line is optional, it allows setting conditions on the tuples. For instance, in the data set of Figure 1, the clause WHERE t2.time = t1.time + 1 means that all the couples of consecutive samples (tuples) will be tested. The theoretical results on rules expressed in RQL presented by Chardin et al. (Chardin *et al.*, 2013) are verified in RulNet, but these theoretical aspects are beyond the scope of this paper.

sample	*G1*	*G2*	*G3*	*G4*
S1	-0.2	0.4	0.3	0.7
S2	-0.9	0.7	0.9	0.8
S3	0.3	0.1	0.7	0.6
S4	0.5	0.3	-0.2	0.8
S5	0.8	0.9	-0.8	-0.7
S6	0.4	0.6	-0.1	-0.9

sample	M1	M2	M3
S1	-0.4	0.7	0.8
S2	0.1	-0.3	0.6
S3	-0.2	0.6	-0.7
S4	-0.1	0.8	0.3
S5	-0.2	-0.5	-0.2
S6	0.1	0.6	-0.1

sample	*A1*	*A2*	*A3*	*time*
S1	1	1	18	1
S2	0	1	17	2
S3	0	0	19	3
S4	1	0	17	4
S5	0	1	34	5
S6	1	1	32	6

Fig. 1. Example of a database composed of three tables: gen (top left), met (top right) and ann (bottom), containing heterogeneous data for six samples. *These tables contain quantitative attributes (G1, G2, G3, G4, M1, M2, M3, A3), binary attributes (A1, A2) and categorical attributes (time).*

Rule generation and interest measures

In RulNet, users have the possibility to generate both exact and approximate 1-1 MP-rules (e.g. $G1 \rightarrow G2$). Users can also choose to infer a subset of n-1 MP-rules (e.g. $G1G2 \rightarrow G3$). In this case, considering the very large number of possible rules, only the exact MP-rules with the smaller left hand sides (in other words, the direct interactions) and the approximate MP-rules with the larger left hand sides are generated. This corresponds respectively to the canonical and the Gottlob and Libkin covers (Gottlob & Libkin, 1990) generalized to MP-rules.

The first step of the processe consists in generating an intermediate set from which the rules will be deduced. This set is a generalization to MP-rules of a result given for functional dependencies (Beeri *et al.*, 1984) and for well-formed semantics (Agier *et al.*, 2007). In RulNet, its generation can be accomplished with a single SQL query and thus benefits of the optimization possibilities of the DBMS (Lopes *et al.*, 2002; Chardin *et al.*, 2013). The second step for the generation of n-1 MP-rules consists in computing the two sets of exact and approximate MP-rules (Eiter & Gottlob, 1995). It is the most expensive step in terms of computation time. RulNet, uses the algorithm and the code proposed by Murakami and Uno, which is recognized as the best solution for such problems (Murakami & Uno). Once generated, rules can be objectively evaluated by calculating quality measures. Support, confidence, lift and leverage are four commonly used measures of significance and interestingness for association rules (Tan *et al.*, 2004; Geng & Hamilton, 2006). In RulNet, these metrics were adapted for MP-rules and are defined as follows.

The support is a measure of significance (importance) of a rule and is given by:

$$Support\left(X \rightarrow Y\right) = \frac{\text{count of } \left(X \cup Y\right)}{N} \qquad (1)$$

where N is the number of possible tuples in the data set verifying the specified condition if it exists (the "WHERE" clause) if only one tuple variable (t1) is defined. If two or more tuple variables are specified, N is equal to the number of possible combinations of tuples in each data set verifying the specified condition if it exists. The support is often used to reduce the search space (i.e. to retain only the rules with a support higher than a user defined threshold value). The confidence is a measure of the strength of a rule. It is defined as the probability of seeing the rule's consequent under the condition that the antecedent is satisfied. The confidence is directed and gives different values for the rules $X \rightarrow Y$ and $Y \rightarrow X$. The confidence is given by:

$$Confidence\left(X \rightarrow Y\right) = \frac{\text{count of } \left(X \cup Y\right)}{\text{count of } X} \qquad (2)$$

The lift (Brin *et al.*, 1997) and the leverage (Piatetsky-Shapiro, 1991) compare the observed support of the rule and the expected support if X and Y are statistically independent. The lift compares how many times X and Y occur together with the number of times they would occur together it they were statistically independent:

$$Lift\left(X \rightarrow Y\right) = \frac{\text{count of } \left(X \cup Y\right) \times N}{\text{count of } X \times \text{count of } Y} \qquad (3)$$

The leverage measures the difference of X and Y appearing together in the data set and what would be expected if X and Y were statistically dependent:

$$Leverage\left(X \rightarrow Y\right) = \frac{\text{count of } \left(X \cup Y\right)}{N} - \frac{\text{count of } X}{N} \times \frac{\text{count of } Y}{N} \qquad (4)$$

A lift equal to 1 and a leverage equal to 0 means that X and Y are statistically independent.

Workflow

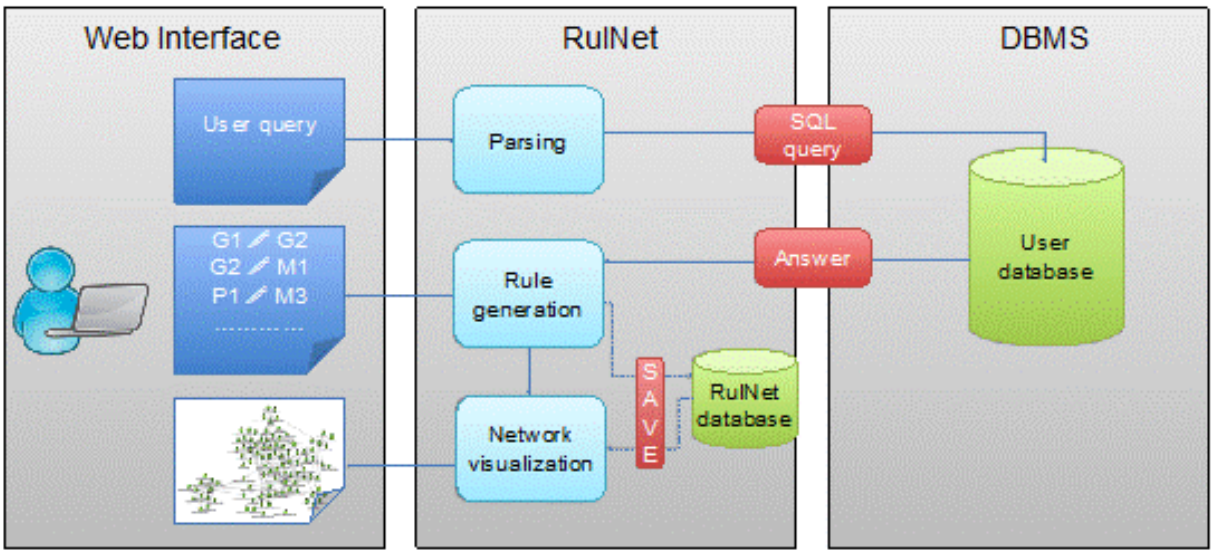


Fig. 2. ***Scheme of the workflow of the RulNet platform and leading to the discovery and visualization of rules.*** *The overall workflow from user's point of view consists of three steps that are data upload, query design and visualization and edition of inferred networks. These steps can be saved and reloaded afterwards when using a registered account.*

The steps leading to the establishment of a rule network are described in Figure 2. Users can upload their data as tabulation delimited text files that contain samples (rows) and attributes (columns) data via the web interface. Each file is then parsed and tuned into a table of a database. RQL is coupled with the standard SQL language; therefore, users can skip the file upload step by connecting directly to a database by specifying the corresponding parameters. The advantage is that the data can be queried directly from where they are stored without any fastidious preprocessing task. The second benefit is that it takes advantage of DBMS technologies for the optimization of computation time.

A syntactic parsing step is implemented to verify that queries respect the defined Backus-Naur form standards (Backus *et al.*, 1960) (Supplementary Text S1) and a query builder is proposed to facilitate the understanding of the language. Two commonly used queries are predefined in the query builder. The first one (called Q1) is similar to association rules and discovers rules between entities if they show similar high or low values in the same samples. The second one (called Q2) generates rules between attributes showing similar profiles in the data. For each query, users can specify the parameters and the thresholds to use for the quality measures.

Networks of generated rules can be visualized with different layouts and the platform offers different features to manipulate these networks. For instance, rules obtained using different queries can be visualized on the same network. Registered users are given the possibility to save and reload the rules and networks inferred at different steps of the analysis. Generated rules can be exported as Cytoscape 3.0 compliant tabulation delimited file for further analysis or to enhance the display of the networks in Cytoscape, a popular software for visualization and analysis of biological networks (Shannon *et al.*, 2003; Smoot *et al.*, 2011).

Characteristics and features

RulNet is available as a web-oriented platform (http://rulnet.isima.fr) installed on a Transtec 2300L Data Storage Server running Intel Xeon E5506 quad-core 2.13 GHz processors with 4 Mb of L2 cache and 8 Gb of RAM. The form parts are implemented using the framework Richfaces 3.3.2 from JBOSS that simplifies the using of AJAX with JSF 1.2 technology. The display layer is a Java applet, built in with the JDK6 version. Users can also download the application to install it on their own server or local machine.

Computational performances

In order to assess the computational performances of RulNet, the two predefined queries were run on the platform using data sets consisting of randomized numerical data. Computing times and the total number of rules discovered (exact and approximate rules) were recorded for data sets of 10 to 4,000 samples (with 100 attributes) and 10 to 4,000 attributes (with 100 samples; Fig. 3). The number of rules given is indicative as it could vary depending on the data used.

For Q1, computing time is low (< 3 s) and does not increase significantly with the number of samples (Fig. 3A). For Q2, computing time is substantially higher than for Q1 (Fig. 3C), because two tuple variables are defined in Q2, which increases with the number of samples. The number of rules discovered increases with the number of samples and is maximum for 500 and 100 samples for Q1 and Q2, respectively. For Q1 and Q2, both computation time and the number of rules increase with the number of attributes (Fig. 3B-D). The number of rules scales with the number of attributes, with a scaling exponent (γ) equal to two.

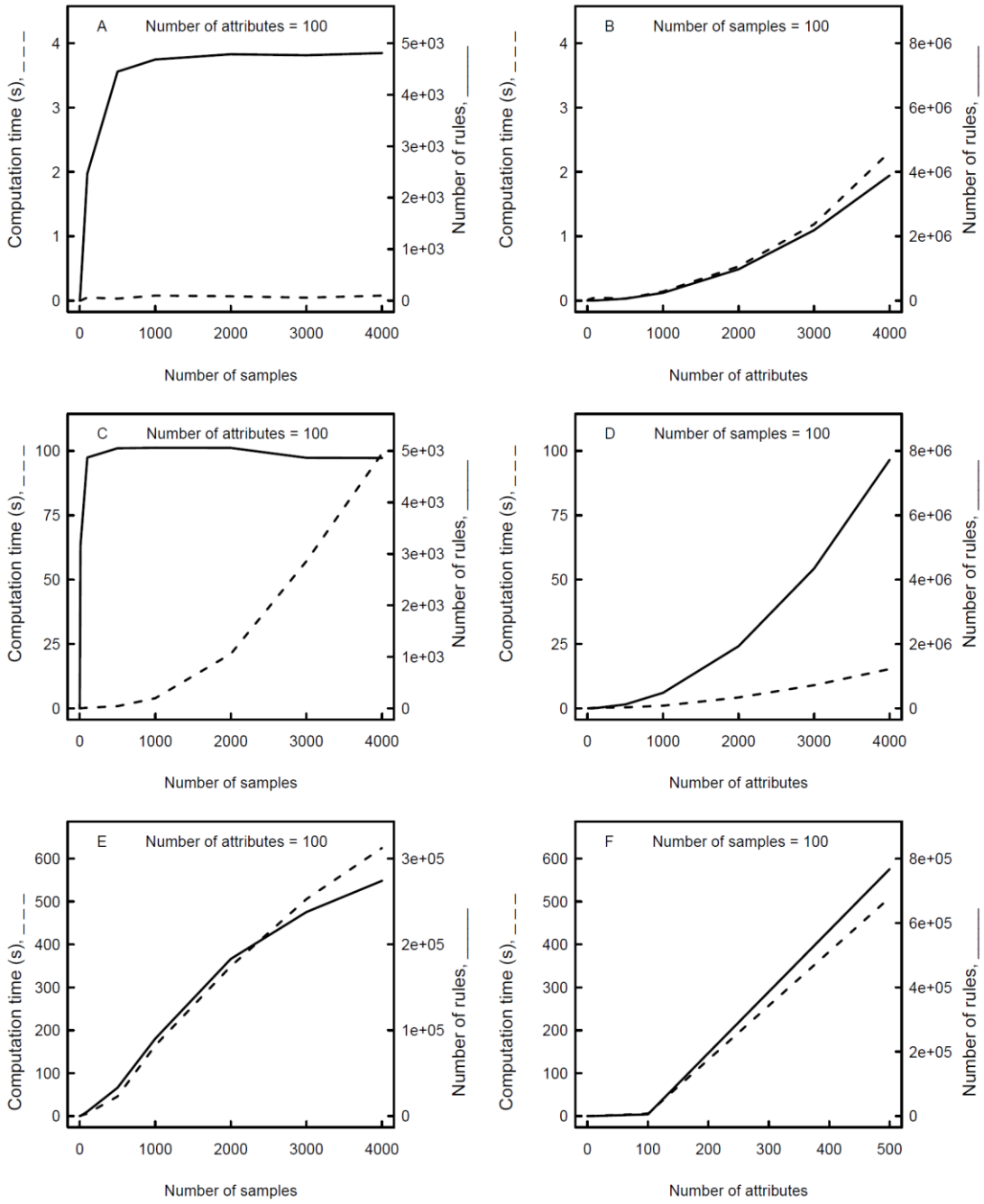


Fig. 3. ***Computing time*** *(dashed lines) and number of rules (solid lines) of networks inferred with the RulNet platform for query Q1 (A, B) and Q2 (C, D) for 1-1 rules and for query Q1 for n-1 rules with n ≤ 3 (E, F) applied to a randomized data set with 10 to 4,000 samples (A, C, E) and attributes (B, D, F).*

With query Q1 the number of n-1 MP-rules (n ≤ 3) and associated computing time increase steadily with the number of samples (Fig. 3E). Computing time increases steeply for 1,000 attributes and exceeds the session timeout of the web server and results are reported for up

to 500 attributes only (Fig. 3F). As for 1-1 rules, the number of rules scales with the number of attributes, but the scaling exponent is 50% higher than for 1-1 rules.

RESULTS

To illustrate the results that can be obtained with RulNet, the platform was used to mine the results of an experiment carried out to better understand wheat grain development adaptive responses to N and S supply (Z. Dai *et al*., in prep.).

Vernalized plants of the winter bread wheat cultivar Récital were grown in a growth chamber from growth stage 13 (Zadoks *et al.*, 1974) to ripeness maturity with a modified Hoagland's nutrient solution (Castle & Randall, 1987) with different rates of N and S supply, either 3 mM N and 2 mM S (control treatment), 3 mM N and 0.02 mM S (low S treatment), or 15 mM N and 0.02 mM S (low S and supra-optimal N treatment). Three additional treatments were applied to each of the three main treatments, complementing the nutrient solution with N and/or S from 25 days after flowering (i.e. midway through the effective grain filling phase) to grain ripeness maturity. Grains were sampled every 3 to 5 days between 10 (i.e. two thirds through the lag phase of grain development) and 34 (i.e. physiological maturity) days after flowering for each of the three main treatments. Grains from the six treatments were also sampled at 0, 6, 9, 12, 18, 24, 48 and 72 h after the shift in N and S supply.

Three independent replicates were used and the following measurements were done for each sampling date. Gene expressions were monitored using a custom 40k Nimblegen® wheat microarray (Rustenholz *et al.*, 2011). Transcriptomic data are available from ArrayExpress repository under the experiment names "N and S regulation of wheat storage protein accumulation" (accessions E-MTAB-1782) and "short-term effect of N and S supply shifts on gene expression in wheat grain" (accessions E-MTAB-1920). The main grain storage protein (GSP) fractions contributing to wheat bread making quality are glutenins and gliadins. High (HMW-GS) and low (LMW-GS) molecular weight-glutenin subunits and ω1,2-, ω5-, γ- and α/β-gliadin proteins were separated and quantified by RP-HPLC (Daniel & Triboï, 2001). The concentration of free amino acids was also assayed by RP-HPLC (Cohen & Michaud, 1993). Metabolites, including organic acids (malate and citrate), soluble sugars (glucose, fructose, and sucrose), oxidized and reduced glutathione and starch were assayed as described by Gibon *et al*. (Gibon *et al.*, 2009).

In the context of this experiment, we defined two distinct objectives. First we want to compare the topology of networks obtained using RulNet with those obtained using two

widely used RN inference methods that are a straightforward Pearson's correlation- based approach (hereafter Pearson) and the Weighted Gene Co-Expression Network Analysis (WGCNA) (Zhang & Horvath, 2005; Langfelder & Horvath, 2008). The second objective is to illustrate the use of RulNet as a biology-driven clustering method.

Topological comparison of correlation-based and RulNet regulatory networks

Using the Significant Analysis of Microarrays (SAM) dimension reduction technique (Tusher *et al.*, 2001), out of the 40,642 transcripts spots on the microarray we identified a total of 984 transcripts associated with N and/or S supply. Among these transcripts, 96 were associated with N supply, 237 with S supply, and 640 with both N and S supply. The expression data for these 973 transcripts were used to infer undirected RNs.

For the three RN inference methods performed here, data were only scaled and centered and used indiscriminately. Threshold parameters for the three methods were set using the scale-free criterion (Supplementary Fig. S1). A common property of large non-random networks is that the distribution of their local connectivity is free of scale, following $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that an attribute in the network interacts with k other attributes (Barabási, 1999). Pearson's RNs were generated using a distance matrix computed using Pearson's correlations with a cutoff threshold of 0.8. In RulNet, we used the Q1 query from the query builder with Confidence and Support thresholds of 0.90 and 0.10, respectively. WGCNA RNs were inferred using the R package WGCNA version 1.34 (Langfelder & Horvath, 2008) with power adjacency set to 9.

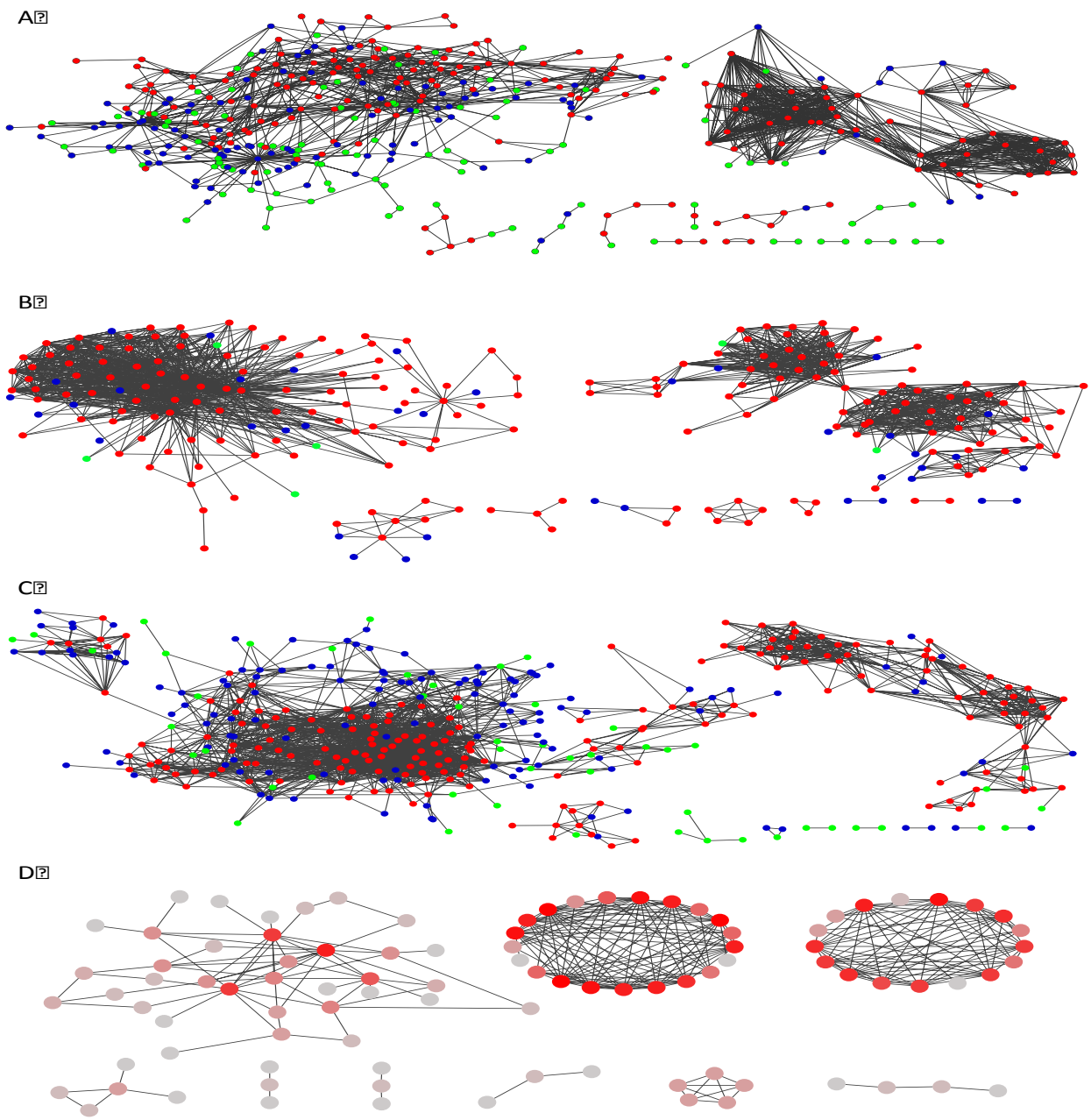


Fig. 4. ***Undirected scale-free regulatory networks inferred with the RulNet platform (A) and the WGCNA (B) and Pearson (C) methods and consensus network obtained using the intersection algorithm of Cytoscape (D)***. *In A, B, and C node color indicate the conservation between the three methods: green, nodes specific to the network; blue, nodes common with one other network; and red nodes found with all three methods. In D, node color is dependent on node degree, from gray for less connected nodes to red for nodes showing the highest connectivity.*

The three methods show common features as well as topological differences (Fig. 4, Table 1). For the three methods, the scaling exponent of the power law distribution is lower than 2 (Table 1), indicating a high importance of the hubs in the networks (Barabási & Oltvai, 2004). The number of nodes is similar for Pearson and RulNet but is 38% lower for WGCNA, while

the number of edges is similar for RulNet and WGCNA but is 60% higher for Pearson. The network connectivity assessed by the average number of neighbors (i.e. the edge-to–node ratio) is preserved between Pearson and WGCNA, but is lower for RulNet (Table 1). It results in a slightly higher number of connected components for the correlation-based methods than for RulNet. Both correlation-based methods have a higher tendency to create "community structures" than RulNet, resulting in higher network density and average clustering coefficient (which characterizes the overall tendency of nodes to form clusters). Although the network density is similar for both correlation-based methods, the network centralization is much greater for WGCNA than for the two other methods, giving a greater importance to hubs in the network. This topological property is apparent in the networks shown in Figure 4. Network heterogeneity represents the tendency of a network to contain hubs and is similar for all three methods. Finally, the characteristic path length, which represents the average over the shortest paths between all pairs of nodes and offers a measure of a network's overall navigability, is lower for WGCNA than for the two other methods.

Table 1. *Network topological properties of regulatory networks inferred using RulNet, WGCNA and Pearson methods*

Topological parameters	RulNet	WGCNA	Pearson
Scale-free r^2	0.82	0.99	0.81
Absolute scaling exponent γ	1.21	1.41	1.08
no. of nodes	397	238	375
no. of edges	1,562	1,564	2,502
Avg. no. of neighbours	6.36	13.14	13.34
Connected components	14	10	10
Network density	0.02	0.05	0.04
Avg. clustering coefficient	0.27	0.71	0.53
Network centralization	0.13	0.29	0.16
Network heterogeneity	1.17	1.14	1.10
Characteristic path length	3.52	2.71	3.71

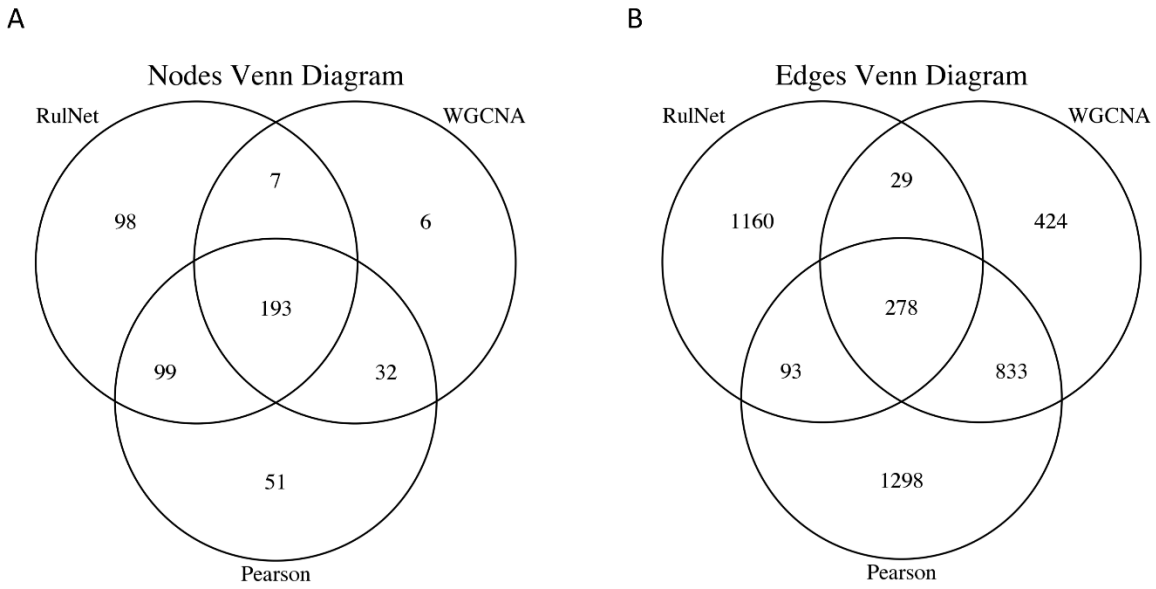


Fig. 5. ***Venn diagram showing node (A) and edge (B) homology between the RulNet platform, WGCNA and Pearson network inference methods.***

Out of the total number of 486 nodes and 4,115 edges found with the three network inference methods, 40% of the nodes and only 7% of the edges are conserved in the three methods (Fig. 4D and 5).

Only 6 nodes are specific to WGCNA while 51 and 98 are specific to Pearson and RulNet, respectively. The conservation of edges between WGCNA and Pearson increases almost linearly with the connectivity of the nodes, from 13% for nodes with a connectivity degree of 1 to 5, to 66% for nodes with a connectivity degree of 81 to 85 (data not shown). The conservation of edges between RulNet and WGCNA and between RulNet and Pearson is independent of the connectivity of the nodes and is close to 10%. The greater conservation of edges between WGCNA and Pearson than between RulNet and the two other methods is not surprising as both WGCNA and Pearson are based on correlation coefficients. These results illustrate the complementarity of RN inference methods reported in previous studies (Marbach *et al.*, 2012a).

Evaluation of RulNet using the DREAM5 network inference challenge datasets

In the context of the DREAM5 challenge, the community of network inference experts was invited to infer genome-scale transcriptional regulatory networks from gene-expression microarray datasets for a prokaryotic model organism (E. coli), a eukaryotic model organism

(S. cerevisiae) and an in silico generated network (Marbach et al., 2012). Each of these three datasets compiles a wide range of experimental conditions. Twenty nine teams participated to this challenge and thirty five methods were compared. For each method, participants were asked to provide a ranked list of 100 000 interactions per dataset which was used to assess the performance of the methods. Two standard quality metrics from machine learning were calculated, the area under the precision-recall (AUPR) and receiver operating characteristic (AUROC) curves (Prill et al., 2010). AUPR is a single measure that summarizes the tradeoff between the completeness (recall) and fidelity (precision) of the inferred network, while AUROC is a single measure that summarizes the tradeoff between the rate of true and false positive predicted edges in a gold standard network. Therefore these two metrics are complementary and provide a comprehensive characterization of the predicted networks. To summarize the performance of each individual method across the three networks overall scores were derived from the AUPR and the AUROC scores by calculating the geometric mean of the network specific scores (Marbach et al., 2012 ; Prill et al., 2010). Finally, an overall score was obtained as the mean of the overall AUPR and AUROC scores. These metrics were calculated using only the interactions for which experimentally supported interactions exists (gold standard interactions). Network predictions from individual teams were integrated to form community networks by rescoring interactions according to their average rank across all methods (Marbach et al., 2012). We benchmarked RulNet using the three datasets and methodology used in the DREAM5 challenge and compared our results with that of the 35 methods that participated to this challenge (Marbach et al., 2012). Two queries called QD1 and QD2 were written for that purpose (Supplementary Text S3). QD1 allows discovering rules between highly and weakly expressed genes, while QD2 allows discovering rules between knockout transcription factors and their putative target genes. For each of these two queries interactions were ranked by descending leverage and confidence. The best rank obtained with QD1 and QD2 were kept for each interaction, which defines a third method named QD1+2. The ranked interactions for each dataset and each method are given in Supplementary Data S1 to S9. Figure 6 shows the values of the AUPR, AUROC and overall scores over the three networks for the three QD queries. Results for the three individual networks and each individual methods of DREAM5 and RulNet are given in Supplementary (Fig. S2). QD1 gives higher AUROC scores than QD1, while QD2 gives higher AUPR scores than QD1. This results is to be related to the biological meaning of the queries and to the fact that the AUROC reflects the quality of the results obtained throughout the full list of interactions while the AUPR is strongly affected by the performance at the top of the prediction list (i.e.

by the most reliable interactions), giving a particular importance to the ranking criteria. The scores obtained with QD1+2 are much higher than those obtained with either QD1 or QD2. The overall AUPR obtained with QD1+2 is close to that of the best individual method evaluated in the context of DREAM5 (Fig. 6) and is second to only one of the 35 individual methods (Supplementary Fig. S2). More importantly, while some methods provide uneven results depending on the dataset, RulNet performs well for each one of them (Supplementary Fig. S2).

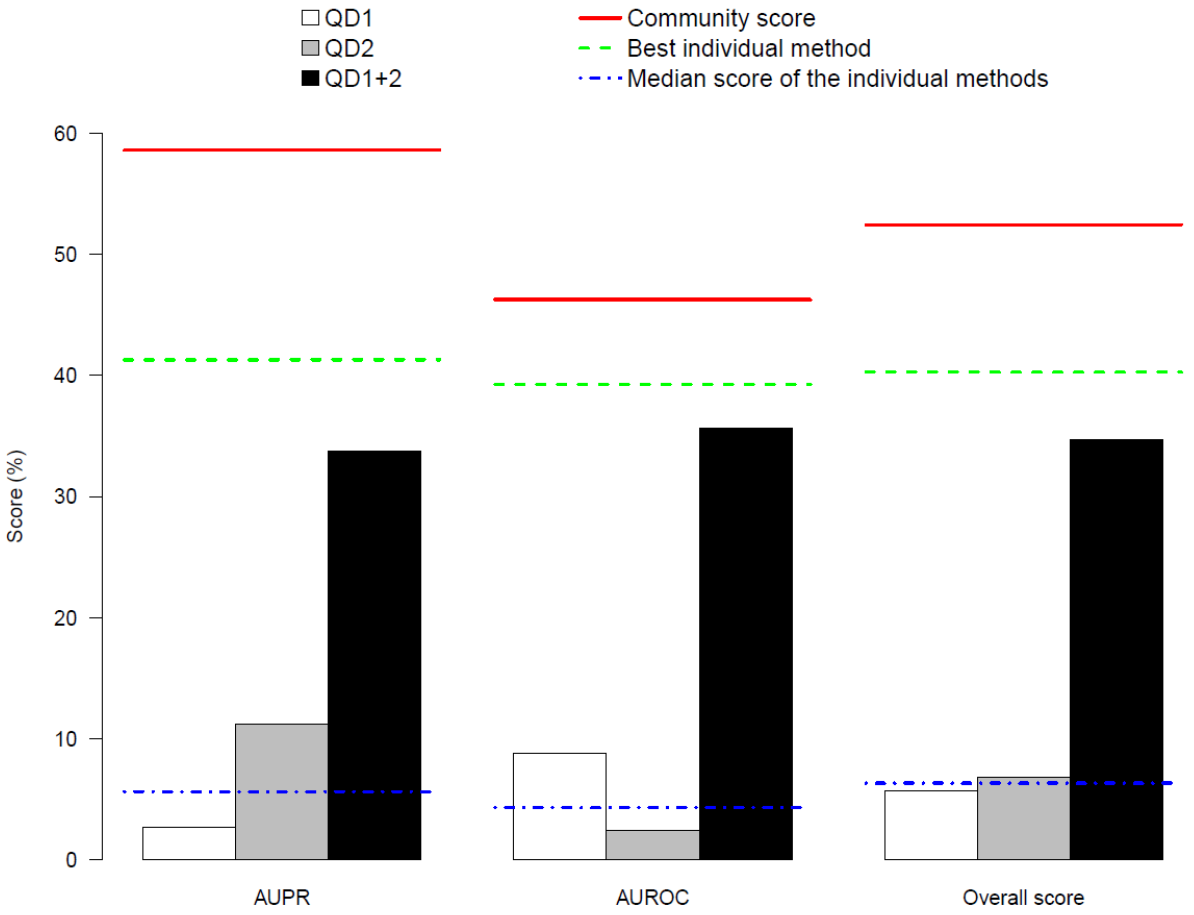


Fig. 6. ***Evaluation of RulNet performance using the DREAM5 network inference challenge datasets***. *The overall (geometric mean across the three inferred networks) area under the precision-recall (AUPR) and receiver operating characteristic (AUROC) and overall score (mean of the overall AUPR and AUROC scores) obtained using RulNet with the QD1, QD2 and QD1+2 queries (Supplementary Text S3) were compared to the 35 methods evaluated in the DREAM5 challenge. The vertical continuous red lines indicate the scores of the integrated community predictions from DREAM5, the vertical dashed green lines indicate the scores of the best individual model from DREAM5 based on the overall score, and the vertical dash-dot blue lines indicate the median score for 35 methods evaluated in the DREAM5 challenge.*

Identification of transcription factors involved in the transcriptional reprogramming occurring during grain development and in response to N and S supply

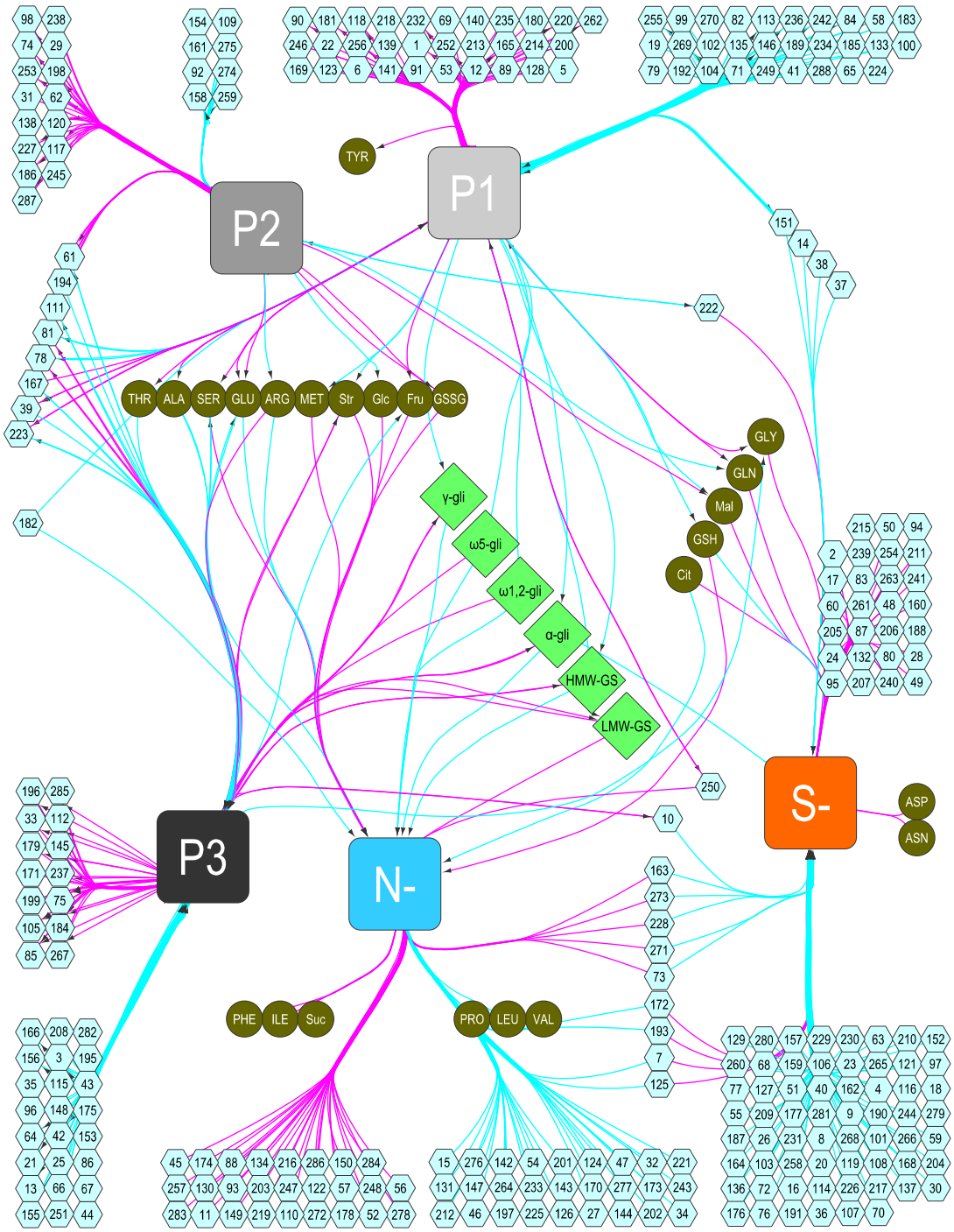


Fig. 7. ***Directed network inferred using the RulNet platform and illustrating the use of central attributes.*** *Linkages of transcription factors expression (hexagons), and the quantity per grain of storage proteins (squares) and metabolites (circles) with the phases of grain development and nitrogen and sulfur deficiencies defined as central attributes. Pink and light blue edges indicate rules discovered with the QNS1 and QNS2 queries, respectively. For metabolites and transcription factors, the correspondence between the node id and actual entity names is given in Supplementary Table S1 and S2, respectively.*

Here we demonstrate how RulNet can be used as a biology-driven clustering method. For this purpose, two features of RulNet have to be used. A unique property of RulNet is that central attributes can be defined for the rule discovery algorithm. When attributes are defined as central, only rules involving at least one of these attributes are discovered. Central attributes could be of any type. The second feature of RulNet that is leveraged in this biology-driven perspective is the possibility to use equivalently quantitative and qualitative data.

The second objective is to highlight TFs involved in the transcriptomic shifts observed between the three major phases of grain development and, independently or also, involved in the adaptive response of wheat grain to N and S deficiency. Moreover, we also try to shed light on TFs potentially involved in the regulation of GSP synthesis in wheat. Two queries called QNS1 and QNS2 were written (Supplementary Text S2) with Support and Confidence thresholds set at 0.15 and 0.90, respectively.

To illustrate this feature of RulNet, N and S deficiency (qualitative data) and the three phases of grain development (qualitative data) were defined as central attributes to discover TFs, GSPs and metabolites (quantitative data) associated with these attributes. Each sample was associated with a phase of development and a level of N and S (low or high) supply. Phase 1 of grain development was assigned to samples taken at 10 and 14 days after flowering, phase 2 to samples taken at 18, 24, and 27 days after flowering, and phase 3 to samples taken at 30 and 34 days after flowering. The expression of all 2,891 TFs transcripts spotted on the microarray was used, as well as the quantity per grain of GSPs and metabolites (Fig. 6).

This network can be considered as a biology-driven clustering visualization rather than as a standard interaction network. The edges between attributes do not testify of a likely interaction but rather of functional links between the central attributes and the other attributes. Overall the network highlights TFs transcripts, metabolites and proteins whose relative abundance is linked to the nutrition (N and/or S) or to a particular phase of the grain development or a combination of both. It is also possible to infer interaction between non-central attributes in the same network.

All five GSPs show expected linkages with the early (Phase 1) and late (Phase 3) phases of grain development, being found in low abundance in the early phase and in high abundance during the late phase (Fig. 6). Overall, the clustering network obtained through this approach allows retrieving some well-known effects of N and S supply on GSP synthesis (Shewry *et al.*, 2000). N deficiency implies a low abundance of the four classes of gliadin and HMW-GS while

it implies a high abundance of LMW-GS. A S deficiency implies a low abundance of the S-rich α-gliadins.

Regarding the metabolites, many respond to N deficiency but only a few are regulated by S deficiency. For instance, S deficiency implies a high abundance of asparagine (ASN) and aspartate (ASP). It was previously reported that ASN could be increased up to a 30-folds in conditions of S deficiency (Muttucumaru *et al.*, 2006; Granvogl *et al.*, 2007; Albrecht Weber *et al.*, 2008). Reduced glutathione (GSH) is also found in low abundance in conditions of S deficiency, while citrate (Cit) and malate (Mal) are found up-regulated, in good agreement with previous reports (Tea *et al.*, 2005; Reinbold *et al.*, 2008; Zörb *et al.*, 2009).

Besides well know responses of GSPs or metabolites to N and S deficiency, this network reveals putative candidate TFs involved in the molecular response of wheat grains to N and S deficiency. In particular it identifies TFs whose expression is independent of grain development but are down- or up-regulated under conditions of N and/or S deficiency. Sixty-two and 71 TFs are associated with N and S deficiency, respectively, but not with grain development. Among these TFs, nine are associated with both N and S deficiency. Interestingly these TFs show opposite response to N and S, suggesting that they orchestrate the response to the N-to-S balance For instance, R2R3-MYB_191 is an orthologous to Arabidopsis thaliana AtMYB7, a seed expressed TF involved in hormone mediated signaling pathways (ethylene, abcisic acid, jasmonic acid, salicylic acid) and overall in the response to abiotic stress (Heyndrickx & Vandepoele, 2012). C2H2_132 A. thaliana ortholog, namely STOP2 is notably involved nitrate transport and in the response to intracellular nitrate concentration (Heyndrickx & Vandepoele, 2012). WRKY_181 A. thaliana ortholog, WRKY30 is involved in the response to ozone and salicylic acid (Besseau *et al.*, 2012) while WRKY_195 A. thaliana ortholog, namely WRKY40 is involved in many pathways including response to hormones, abiotic stress and pathogen defense (Yanhui *et al.*, 2006; Heyndrickx & Vandepoele, 2012). All these observations make these TFs good candidates in the context of the study of N/S response in wheat grain. More complex queries could allow us to further elucidate complex regulatory circuitry involved in the adaptation of grain to N/S supply.

DISCUSSION

Scale-free topology: Most functional features of a cell are driven by groups of molecules, which are not secluded but rather linked to each other or even overlapping. In strong contrast with random networks, this characteristic is reflected by a scale-free topology of the networks at different levels of the cell functional organization. One can therefore assume that a consistent network inference method should tend towards scale-free topology resulting networks. The main feature of scale-free networks is the presence of nodes of widely different connectivity, including weakly connected nodes representing molecules with restricted field of action and hubs (highly connected nodes) representing important regulators. Here we showed that RulNet infers scale-free networks depending on the thresholds chosen for quality measures.

TFs interdependency: In this work, we illustrate the efficiency of a RN inference and visualization platform using –omics data including TFs expression. It is well known that TFs are interdependent, i.e. they often work in combinations with other TFs. In the work presented, different parameters would allow discovering rules involving multiple attributes in the left hand side (n-1 MP-rules), meeting the needs of a TFs study.

Programming knowledge: RN inference is the most efficient tool to predict and analyze RNs using high throughput data. Powerful algorithms and tools have been made available to infer RNs. Unlike some of these tools, preliminary programming skills might not be required to perform a simple yet complete analysis using the platform presented here. Whereas writing simple queries or designing simple queries is made easy thanks to the query builder provided on the platform, prior knowledge of the SQL query language could facilitate the assimilation of the RQL language available to perform custom queries and take the best out of RulNet.

Dimensionality issue: A first issue, of technical nature should be addressed. Even in the context of high throughput technologies, experimental limitations do not allow the acquisition of a number of observations commensurate with the number of variables. The consistency of the results obtained should increase proportionally to the number of observations. The complexity of the approach together with the relatively restricted number of independent data points compared to the size of the search space can limit the scope and significance of the results. However, RN inference using –omics data allows researchers to model complex systems of regulation although they do not reveal the full complexity of the system but only the observable components. Moreover, similarly to most approaches, the interaction inferred between two variables attests only that their relative quantity are

somehow linked, but does not elucidate the hypothetic complexity of the interaction. The latter issue applies to most RN inference approaches.

Taking advantage of the complementarity: The main asset of RulNet stems from its capability to perform different queries in the same framework It has been highlighted by comparative analyses that most approaches used to infer RNs from the same data sets show relatively distinct results (Marbach et al., 2012a). The discrepancy observed between the RNs inferred does not demonstrate any failure from a particular approach but rather highlights their complementarity. In the approach presented here, this issue is partly addressed by offering the possibility to use multiple queries to discover rules in a unifying framework. This is clearly seen in the evaluation of RulNet with the DREAM5 datasets, the evaluation of QD1+2 is better than that of either QD1 or QD2. The main conclusion of the DREAM consortium is that integrated community networks tend to get closer to the reality. To a reasonable extent, RulNet provides by itself the possibility to infer integrated networks composed by results from different queries. Considering the quality metrics of the DREAM consortium, we observe that different queries from RulNet allow obtaining different results highlighting the importance of query design prior to the inference process.

The web-oriented platform offers multiple features to visualize and edit a global RN gathering the interactions inferred using different queries. Moreover we conducted a comparative analysis of RulNet inference method with two other commonly used inference methods, namely correlation network inference and WGCNA. This analysis highlights RulNet specificities but also the consistency of the typological properties of the networks inferred with this three methods. Considering the complexity of eukaryotic organisms, it is essential to consider more than one type of interactions, which is not proposed by most of the available tools. However, it is ascertained that no approach reveals the absolute and complete truth of complex RNs; only exploiting the complementarity of the approaches can narrow the gap between the hypothetical network and the actual complex network of interactions (Marbach *et al.*, 2012a).

REFERENCES

Agier M, Petit J-M, Suzuki E. **2007**. Unifying Framework for Rule Semantics: Application to Gene Expression Data. *Fundam. Inf.* **78**: 543–559.

Agrawal R, Imieliński T, Swami A. **1993**. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*: 207–216.

Albrecht Weber E, Koller W-D, Graeff S, Hermann W, Merkt N, Claupein W. **2008**. Impact of different nitrogen fertilizers and an additional sulfur supply on grain yield, quality, and the potential of acrylamide formation in winter wheat. *Journal of Plant Nutrition and Soil Science* **171**: 643–655.

Armstrong WW. **1974**. Dependency Structures of Data Base Relationships. IFIP Congress.580–583.

Backus JW, Bauer FL, Green J, Katz C, McCarthy J, Perlis AJ, Rutishauser H, Samelson K, Vauquois B, Wegstein JH, *et al.* 1960. Report on the algorithmic language ALGOL 60 (P Naur, Ed.). *Commun. ACM* **3**: 299–314.

Barabási a. 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**: 101–13.

Beeri C, Dowd M, Fagin R, Statman R. **1984**. On the Structure of Armstrong Relations for Functional Dependencies. *J. ACM* **31**: 30–46.

Belcastro V, Siciliano V, Gregoretti F, Mithbaokar P, Dharmalingam G, Berlingieri S, Iorio F, Oliva G, Polishchuck R, Brunetti-Pierri N, *et al.* 2011. Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. *Nucleic acids research* **39**: 8677–88.

Besseau S, Li J, Palva ET. **2012**. WRKY54 and WRKY70 co-operate as negative regulators of leaf senescence in Arabidopsis thaliana. *Journal of experimental botany* **63**: 2667–79.

Bohannon P, Fan W, Geerts F, Jia X, Kementsietsidis A. **2007**. Conditional Functional Dependencies for Data Cleaning. *2007 IEEE 23rd International Conference on Data Engineering*: 746–755.

Brin S, Motwani R, Ullman J, Tsur S. **1997**. Dynamic Itemset Counting and Implication Rules for Market Basket Data. ACM SIGMOD Conf.\ Management of Data.

Castle SL, Randall PJ. **1987**. Effects of Sulfur Deficiency on the Synthesis and Accumulation of Proteins in the Developing Wheat Seed. *Functional Plant Biology* **14**: 503–516.

Chardin B, Coquery E, Gouriou B. **2013**. Query Rewriting for Rule Mining in Databases. ….

Cohen SA, Michaud DP. **1993**. Synthesis of a Fluorescent Derivatizing Reagent, 6-Aminoquinolyl-N-Hydroxysuccinimidyl Carbamate, and Its Application for the Analysis of Hydrolysate Amino Acids via High-Performance Liquid Chromatography. *Analytical Biochemistry* **211**: 279–287.

Daniel C, Triboï E. **2001**. Effects of temperature and nitrogen nutrition on the accumulation of gliadins analysed by RP-HPLC. *Functional Plant Biology* **28**: 1197–1205.

Durzinsky M, Wagler A, Marwan W. **2011**. Reconstruction of extended Petri nets from time series data and its application to signal transduction and to gene regulatory networks. *BMC systems biology* **5**: 113.

Eiter T, Gottlob G. **1995**. Identifying the Minimal Transversals of a Hypergraph and Related Problems. *SIAM J. Comput.* **24**: 1278–1304.

Fan W, Geerts F, Jia X, Kementsietsidis A. **2008**. Conditional functional dependencies for capturing data inconsistencies. *ACM Transactions on Database Systems* **33**: 1–48.

Friedman N, Linial M, Nachman I, Pe'er D. **2000**. Using Bayesian networks to analyze expression data. *Journal of computational biology : a journal of computational molecular cell biology* **7**: 601–20.

Ganter B, Wille R. **1997**. *Formal Concept Analysis: Mathematical Foundations*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Geng L, Hamilton HJ. **2006**. Interestingness measures for data mining. *ACM Computing Surveys* **38**: 9–es.

Georgii E, Richter L, Rückert U, Kramer S. **2005**. Analyzing microarray data using quantitative association rules. *Bioinformatics (Oxford, England)* **21 Suppl 2**: ii123–9.

Gibon Y, Pyl E-T, Sulpice R, Lunn JE, Höhne M, Günther M, Stitt M. **2009**. Adjustment of growth, starch turnover, protein content and central metabolism to a decrease of the carbon supply when Arabidopsis is grown in very short photoperiods. *Plant, Cell & Environment* **32**: 859–874.

Gottlob G, Libkin L. **1990**. Investigations on Armstrong Relations , Dependency Inference and Excluded Functional Dependencies Investigations on Armstrong Relations , Dependency Inference , and Excluded Functional Dependencies. **9**: 385–402.

Granvogl M, Wieser H, Koehler P, Tucher S Von, Schieberle P. **2007**. Influence of sulfur fertilization on the amounts of free amino acids in wheat. correlation with baking properties as well as with 3-aminopropionamide and acrylamide generation during baking. *Journal of agricultural and food chemistry* **55**: 4271–7.

Gupta RB, Masci S, Lafiandra D, Bariana HS, MacRitchie F. **1996**. Accumulation of protein subunits and their polymers in developing grains of hexaploid wheats. *Journal of Experimental Botany* **47** : 1377–1385.

Gupta R, Singh N, Shepherd K. **1989**. The cumulative effect of allelic variation in LMW and HMW glutenin subunits on dough properties in the progeny of two bread wheats. *Theoretical and Applied Genetics*: 57–64.

Haibe-Kains B, Olsen C, Djebbari A, Bontempi G, Correll M, Bouton C, Quackenbush J. **2012**. Predictive networks: a flexible, open source, web application for integration and analysis of human gene networks. *Nucleic acids research* **40**: D866–75.

Heyndrickx KS, Vandepoele K. **2012**. Systematic identification of functional plant modules through the integration of complementary data sources. *Plant physiology* **159**: 884–901.

Kauffman S, Peterson C, Samuelsson B, Troein C. **2003**. Random Boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 14796–9.

Kotera M, Yamanishi Y, Moriya Y, Kanehisa M, Goto S. **2012**. GENIES: gene network inference engine based on supervised analysis. *Nucleic acids research* **40**: W162–7.

Küffner R, Zimmer R, Lengauer T. **2000**. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics (Oxford, England)* **16**: 825–36.

Langfelder P, Horvath S. **2008**. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**: 559.

Liang J, Han J. **2012**. Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks. *BMC systems biology* **6**: 113.

Lopes S, Petit J-M, Lakhal L. **2002**. Functional and approximate dependency mining: database and FCA points of view. *Journal of Experimental & Theoretical Artificial Intelligence* **14**: 93–114.

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Consortium TD, Kellis M, Collins JJ, *et al.* 2012a. Wisdom of crowds for robust gene network inference.

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G. **2012b**. Wisdom of crowds for robust gene network inference. *Nat Meth* **9**: 796–804.

Margolin A a, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. **2006**. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7 Suppl 1**: S7.

Martinez R, Pasquier N, Pasquier C. **2008**. GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. *Bioinformatics (Oxford, England)* **24**: 2643–4.

Martre P, Porter JR, Jamieson PD. **2003**. Modeling Grain Nitrogen Accumulation and Protein Composition to Understand the Sink / Source Regulations of Nitrogen Remobilization for Wheat. **133**: 1959–1967.

Miller JH, Beckwith JON, Muller-Hill B. **1968**. Direction of Transcription of a Regulatory Gene in E. coli. *Nature* **220**: 1287–1290.

Mordelet F, Vert J-P. **2008**. SIRENE: supervised inference of regulatory networks. *Bioinformatics (Oxford, England)* **24**: i76–82.

Murakami K, Uno T. Efficient Algorithms for Dualizing Large-Scale Hypergraphs.

Muttucumaru N, Halford NG, Elmore JS, Dodson AT, Parry M, Shewry PR, Mottram DS. **2006**. Formation of high levels of acrylamide during the processing of flour derived from sulfate-deprived wheat. *Journal of agricultural and food chemistry* **54**: 8951–5.

Pan Y, Bradley G, Pyke K, Ball G, Lu C, Fray R, Marshall A, Jayasuta S, Baxter C, van Wijk R, *et al.* 2013. Network Inference Analysis Identifies an APRR2-Like Gene Linked to Pigment Accumulation in Tomato and Pepper Fruits. *Plant physiology* **161**: 1476–85.

Piatetsky-Shapiro G. **1991**. Discovery, analysis and presentation of strong rules. In: Piatetsky-Shapiro G, Frawley WJ, eds. Knowledge Discovery in Databases. AAAI Press, 229–248.

Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, et al. 2010. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PLoS One **5**:e9202.

Reinbold J, Rychlik M, Asam S, Wieser H, Koehler P. **2008**. Concentrations of total glutathione and cysteine in wheat flour as affected by sulfur deficiency and correlation to quality parameters. *Journal of agricultural and food chemistry* **56**: 6844–50.

Rustenholz C, Choulet F, Laugier C, Safár J, Simková H, Dolezel J, Magni F, Scalabrin S, Cattonaro F, Vautrin S, *et al.* 2011. A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant physiology* **157**: 1596–608.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. **2003**. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. : 2498–2504.

Shen-Orr SS, Milo R, Mangan S, Alon U. **2002**. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics* **31**: 64–8.

Shewry PR, Tatham AS, Halford NG. **2000**. Nutritional control of storage protein synthesis in developing grain of wheat and barley.

De Smet R, Marchal K. **2010**. Advantages and limitations of current network inference methods. *Nature reviews. Microbiology* **8**: 717–29.

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. **2011**. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)* **27**: 431–2.

Tan P-N, Kumar V, Srivastava J. **2004**. Selecting the right objective measure for association analysis. *Information Systems* **29**: 293–313.

Tea I, Genter T, Violleau F, Kleiber D. **2005**. Changes in the glutathione thiol-disulfide status in wheat grain by foliar sulphur fertilization: consequences for the rheological properties of dough. *Journal of Cereal Science* **41**: 305–315.

Triboi E, Abad A, Michelena A, Lloveras J, Ollier JL, Daniel C. **2000**. Environmental effects on the quality of two wheat genotypes: 1. quantitative and qualitative variation of storage proteins. *European Journal of Agronomy* **13**: 47–64.

Tusher VG, Tibshirani R, Chu G. **2001**. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 5116–21.

Wrigley CW, Cros DL Du, Fullington JG, Kasarda DD. **1984**. Changes in polypeptide composition and grain quality due to sulfur deficiency in wheat. *Journal of Cereal Science* **2**: 15–24.

Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, Zhiqiang L, Yunfei Z, Xiaoxiao W, Xiaoming Q, *et al.* 2006. The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. *Plant molecular biology* **60**: 107–24.

Yoon H, McDermott JE, Porwollik S, McClelland M, Heffron F. 2009. Coordinated regulation of virulence during systemic infection of Salmonella enterica serovar Typhimurium. *PLoS pathogens* **5**: e1000306.

Zadoks JC, Chang TT, Konzak CF. 1974. A decimal code for the growth stages of cereals. *Weed Research* **14**: 415–421.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4**: Article17.

Zörb C, Steinfurth D, Seling S, Langenkämper G, Koehler P, Wieser H, Lindhauer MG, Mühling KH. 2009. Quantitative protein composition and baking quality of winter wheat as affected by late sulfur fertilization. *Journal of agricultural and food chemistry* **57**: 3877–85.

SUPPLEMENTARY INFORMATION

Text S1 Grammar specifications of the RQL language

The grammar specifications of the language to define different semantics can be defined using Backus-Naur Form (BNF) standards (Backus *et al.*, 1960) (Box S1). A BNF consists of a set of derivation rules of the form <*symbol*> ::= _expression_, where _expression_ consists of one or more sequences of symbols, separated by a vertical bar (|) indicating a choice. A symbol is terminal if it does not appear on the left hand side of any derivation rule. The non-terminal symbol <*query*> shows the global form of a query, from which we can see the SQL inspiration.

Box S1. Backus-Naur Form grammar of the extended form of the RQL language

```
<query> ::=
   "FINDRULES "
   "SCOPE " <psi>
   "HAVING " <pred_list>";"
<psi> ::=
   <tupvar_list> " [WHERE " <SQL_condition>]
 <tupvar_list> ::=
   <tupvar_def> | <tupvar_def> " , " <tupvar_list>
<tupvar_def> ::=
   <tupvar> " IN " <data>
<data> ::=
   <table_name> | "("<SQL_query>")"
<pred_list> :=
   <predicate> | <predicate> " AND " <pred_list>
<predicate> := [<label>" : "] <delta> " OVER " <att_list> | ALL | ALL MINUS <att_list>";"
<delta> ::= <delta_atomic>  |
   "(" <delta> ")"  |
   " NOT " <delta> |
   <delta> <logical_connector> <delta>
<delta_atomic> ::=
   <tupvar> ".ATT " <comparison_op> <tupvar> ".ATT" |
   <tupvar> ".ATT " <comparison_op> <constant>
<logical_connector> ::= " OR " | " AND "
<comparison_op> :: "=" | "<" | ">" | "<=" | ">=" | "!="
```

<*SQL_query*> represents a classical (SELECT FROM WHERE) query and <*SQL_condition*> is the clause used for instance in the "WHERE" statement of a SQL query. RQL also provides most of the functions provided by SQL, such as SQRT, ABS, ROWID, SUBSTR or LTRIM.

The "OVER" clause specifies the attributes on which the predicates have to be discovered. These attributes have to be present in the datasets otherwise the query cannot be semantically correct.

A predicate is defined as a label (<label>), a property (<delta>) and an attribute list (<att_list>) on which the predicate is tested. The clause <label> is optional but it is necessary if multiple predicates are defined for an attribute. The keyword "ATT" will then be replaced by each attribute of the <att_list> clause and tested to check if the rule is satisfied.

Text S2 Semantics used to infer regulatory networks from wheat –omics data

QNS1: Discover rules between highly abundant molecules and N/S supply or development phase.

```
FINDRULES
SCOPE t1 IN Omic_data
HAVING high : t1.ATT > 0.5 OVER ALL MINUS sample, Nitrogen, Sulfur, Phase_1, Phase_2,
Phase_3
AND NoNitrogen : t1.ATT = 0 OVER Nitrogen
AND Nitrogen : t1.ATT = 15 OVER Nitrogen
AND NoSulfur : t1.ATT = 0.2 OVER Sulfur
AND Sulfur: t1.ATT = 2 OVER Sulfur
AND Cellularization : t1.ATT = 1 OVER Phase_1
AND Storage : t1.ATT = 1 OVER Phase_2
AND Senescence : t1.ATT = 1 OVER Phase_3;
```

QNS2: Discover rules between not abundant molecules and N/S supply or development phase.

```
FINDRULES
SCOPE t1 IN Omic_data
HAVING high : t1.ATT < -0.5 OVER ALL MINUS sample, Nitrogen, Sulfur, Phase_1, Phase_2,
Phase_3
AND NoNitrogen : t1.ATT = 0 OVER Nitrogen
AND Nitrogen : t1.ATT=15 OVER Nitrogen
AND NoSulfur : t1.ATT = 0.2 OVER Sulfur
AND Sulfur: t1.ATT = 2 OVER Sulfur
AND Cellularization : t1.ATT = 1 OVER Phase_1
AND Storage : t1.ATT = 1 OVER Phase_2
AND Senescence : t1.ATT = 1 OVER Phase_3;
```

Text S3 Semantics used to infer regulatory networks from the DREAM5 datasets

QD1: Discover rules between highly and weakly expressed genes.

```
FINDRULES
SCOPE t1 IN netx
HAVING high : t1.ATT > 0.5 OVER ALL
AND low : t1.ATT < -0.5 OVER ALL
```

QD2: Discover rules between knockout transcription factors and their putative target genes.

```
FINDRULES
SCOPE t1 IN netx, t2 IN netx
WHERE t1.Experiment = t2.Experiment AND t1.Perturbations = t2.Perturbations AND
t1.PerturbationLevels = t2.PerturbationLevels AND t1.Treatment = t2.Treatment AND
t1.OverexpressedGenes = t2.OverexpressedGenes AND t1.Time = t2.Time AND
t1.DeletedGenes != t2.DeletedGenes AND t1.DeletedGenes = 'NA'
HAVING higher : t2.ATT-t1.ATT > 0.5 OVER ALL
AND lower : t1.ATT-t2.ATT > 0.5 OVER ALL
AND G1 : t2.ATT='G1' OVER DeletedGenes
AND …
AND G195 : t2.ATT='G195' OVER DeletedGenes
```

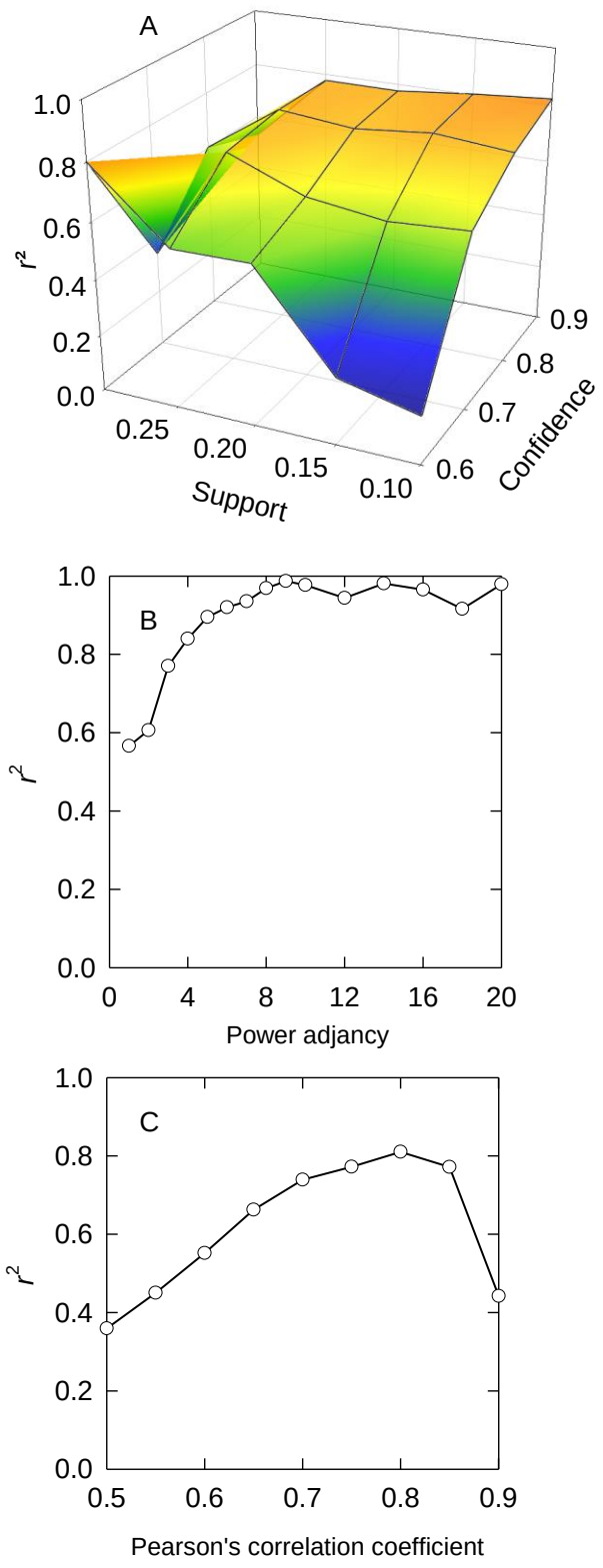


Fig. S1. Power law analysis of regulatory networks inferred with three different methods. *(A) RulNet platform, (B) GCNA, and (C) Pearson correlation methods. For each method, the coefficient of determination (r^2) of the power-law connectivity distribution is plotted against the threshold values.*

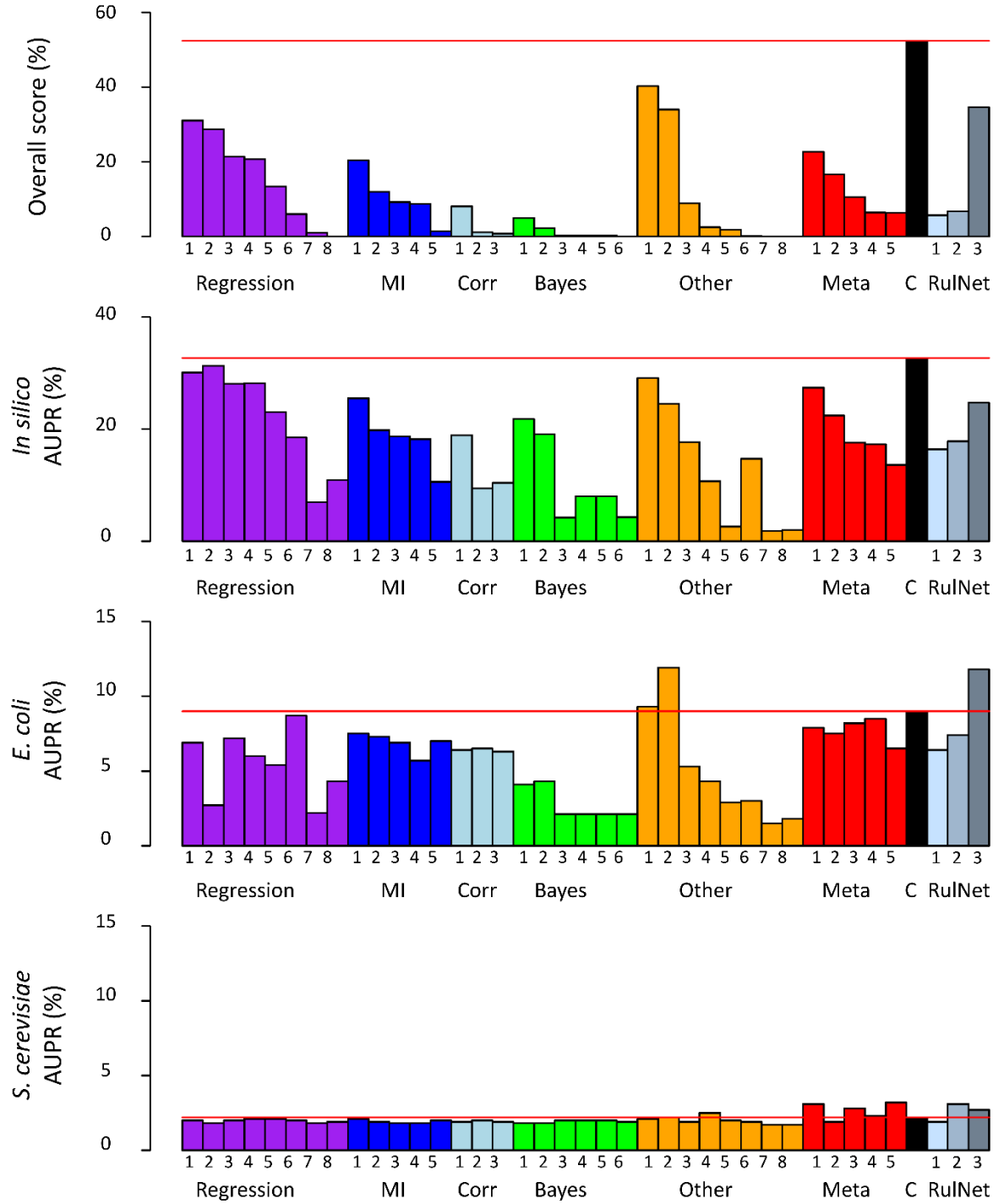


Fig. S2. Comparison of the performance of RulNet for the DREAM5 networks with that of the 35 network inference methods that participated to the DREAM5 challenge. *The plots show the area under precision-recall curve (AUPR) for the individual networks and the overall score summarizing the performance across networks. Inference methods are grouped by type of methods (Regression, mutual information [MI], correlation [Corr], bayesian network [Bayesian], other approaches [Other], meta predictor[Meta]) and within a group of methods are indexed as in [1]. For RulNet, the index 1, 2, and 3 are the queries QD1 and QD2, and the method QD1+2. C and vertical red lines are integrated community predictions. Figure Based on Marbach et al., 2012.*

Table S1. Abbreviations of metabolites used in Figure 6.

Metabolite	Abbreviation
Alanine	ALA
Arginine	ARG
Asparagine	ASN
Aspartic acid	ASP
Glutamine	GLN
Glutamic acid	GLU
Glycine	GLY
Reduced Glutathione	GSH
Glutathione disulfide	GSSG
Isoleucine	ILE
Leucine	LEU
Methionine	MET
Phenylalanine	PHE
Proline	PRO
Serine	SER
Threonine	THR
Tyrosine	TYR
Valine	VAL
Citrate	Cit
Fructose	Fru
Glucose	Glc
Malate	Mal
Starch	Str
Sucrose	Suc

Table S2. Correspondence of transcription factors identifiers and names used in Figure 6. The names are those used in the wheat transcription factor database wDBTF (http://wwwappli.nantes.inra.fr:8180/wDBFT/).

Id	Name	Id	Name	Id	Names
1	Alfin_8	40	bHLH_90	79	C3H_106
2	AP2_16	41	bZIP_117	80	C3H_32
3	ARF_48	42	bZIP_131	81	C3H_4
4	ARF_49	43	bZIP_133	82	C3H_49
5	ARF_77	44	bZIP_134	83	C3H_59
6	ARF_80	45	bZIP_139	84	C3H_93
7	ARID_3	46	bZIP_147	85	CCAAT-HAP2_7
8	AS2_16	47	bZIP_153	86	CCAAT-HAP3_1
9	AS2_20	48	bZIP_157	87	CCAAT-HAP3_2
10	AS2_21	49	bZIP_162	88	CCAAT-HAP5_18
11	AS2_5	50	bZIP_163	89	CPP_16
12	ASH1_3	51	bZIP_35	90	CPP_3
13	AUX/IAA_1	52	bZIP_53	91	E2F-DP_11
14	AUX/IAA_14	53	bZIP_80	92	E2F-DP_4
15	AUX/IAA_57	54	bZIP_91	93	EIL_11
16	AUX/IAA_75	55	bZIP_94	94	EIL_16
17	BBR/BPC_3	56	C2C2-CO-like_64	95	EIL_19
18	BES1_10	57	C2C2-CO-like_71	96	EREBP_102
19	BES1_9	58	C2C2-CO-like_8	97	EREBP_106
20	BHL_13	59	C2C2-CO-like_89	98	EREBP_108
21	BHL_16	60	C2C2-DOF_35	99	EREBP_149
22	BHL_20	61	C2C2-DOF_9	100	EREBP_153
23	BHL_6	62	C2C2-GATA_10	101	EREBP_157
24	bHLH_119	63	C2C2-GATA_13	102	EREBP_159
25	bHLH_120	64	C2C2-GATA_17	103	EREBP_16
26	bHLH_153	65	C2C2-GATA_20	104	EREBP_161
27	bHLH_161	66	C2C2-GATA_21	105	EREBP_162
28	bHLH_165	67	C2C2-YABBY_10	106	EREBP_163
29	bHLH_171	68	C2C2-YABBY_4	107	EREBP_176
30	bHLH_173	69	C2H2_102	108	EREBP_191
31	bHLH_184	70	C2H2_103	109	EREBP_197
32	bHLH_222	71	C2H2_118	110	EREBP_2
33	bHLH_23	72	C2H2_121	111	EREBP_206
34	bHLH_265	73	C2H2_132	112	EREBP_213
35	bHLH_278	74	C2H2_138	113	EREBP_225
36	bHLH_283	75	C2H2_170	114	EREBP_232
37	bHLH_302	76	C2H2_74	115	EREBP_246
38	bHLH_35	77	C2H2_80	116	EREBP_267
39	bHLH_69	78	C3H_10	117	EREBP_39

Table S2. Continued

118	EREBP_42	157	M-type_2	196	NAC_32
119	EREBP_56	158	MBF1_4	197	NAC_34
120	EREBP_58	159	MBF1_43	198	NAC_90
121	EREBP_68	160	MBF1_9	199	NAC_95
122	EREBP_86	161	MIKC-type_13	200	Nin-like_2
123	GARP-ARR-B_19	162	MIKC-type_32	201	PAT1_2
124	GARP-G2-like_22	163	MIKC-type_4	202	PAT1_4
125	GARP-G2-like_23	164	MIKC-type_63	203	PAT1_9
126	GARP-G2-like_4	165	MIKC-type_7	204	PHD-ZNFG_1
127	GARP-G2-like_79	166	MIKC-type_87	205	PHD-ZNFG_146
128	GARP-G2-like_85	167	MIKC-type_96	206	PHD-ZNFG_149
129	GARP-G2-like_89	168	MYB-related_11	207	PHD-ZNFG_186
130	GeBP_16	169	MYB-related_114	208	PHD-ZNFG_201
131	GeBP_7	170	MYB-related_125	209	PHD-ZNFG_211
132	GeBP_8	171	MYB-related_126	210	PHD-ZNFG_244
133	HAM_11	172	MYB-related_162	211	PHD-ZNFG_247
134	HAM_3	173	MYB-related_27	212	PHD-ZNFG_248
135	HD-ZIP_14	174	MYB-related_29	213	PHD-ZNFG_261
136	HD-ZIP_17	175	MYB-related_36	214	PHD-ZNFG_262
137	HD-ZIP_32	176	MYB-related_7	215	PHD-ZNFG_287
138	HD-ZIP_6	177	MYB-related_98	216	PHD-ZNFG_289
139	HD-ZIP_88	178	MYB-related_99	217	PHD-ZNFG_60
140	HMG_46	179	NAC_109	218	PHD-ZNFG_61
141	HMG_47	180	NAC_118	219	PHD-ZNFG_71
142	HMG_50	181	NAC_119	220	PHD-ZNFG_73
143	HMG_77	182	NAC_12	221	PHD-ZNFG_77
144	HMG_81	183	NAC_125	222	PHD-ZNFG_83
145	HSF_12	184	NAC_126	223	PLATZ_1
146	HSF_48	185	NAC_13	224	R2R3-MYB_105
147	HSF_54	186	NAC_14	225	R2R3-MYB_132
148	HSF_58	187	NAC_159	226	R2R3-MYB_166
149	JUMONJI_13	188	NAC_16	227	R2R3-MYB_181
150	KNOX_15	189	NAC_172	228	R2R3-MYB_191
151	KNOX_17	190	NAC_20	229	R2R3-MYB_193
152	LIM_10	191	NAC_203	230	R2R3-MYB_208
153	LIM_28	192	NAC_222	231	R2R3-MYB_210
154	LIM_36	193	NAC_244	232	R2R3-MYB_219
155	LUG_13	194	NAC_249	233	R2R3-MYB_232
156	LUG_14	195	NAC_250	234	R2R3-MYB_27

Table S2. *Continued*

235	R2R3-MYB_31	274	WRKY_199
236	R2R3-MYB_83	275	WRKY_201
237	REM_12	276	WRKY_210
238	REM_18	277	WRKY_215
239	REM_30	278	WRKY_222
240	REM_46	279	WRKY_31
241	REM_54	280	WRKY_46
242	REM_57	281	WRKY_81
243	REM_62	282	WRKY_82
244	REM_65	283	WRKY_87
245	REM_66	284	ZF-HD_5
246	REM_7	285	ZF-HD_9
247	REM_75	286	ZIM_20
248	SBP_22	287	ZIM_4
249	SBP_24	288	ZIM_75
250	SBP_5		
251	SUVAR_24		
252	SUVAR_40		
253	TAZ_13		
254	Trihelix_1		
255	Trihelix_2		
256	Trihelix_24		
257	Trihelix_33		
258	Trihelix_37		
259	TUB_28		
260	TUB_29		
261	TUB_41		
262	TUB_9		
263	Whirly_4		
264	WRKY_113		
265	WRKY_120		
266	WRKY_137		
267	WRKY_150		
268	WRKY_17		
269	WRKY_171		
270	WRKY_173		
271	WRKY_181		
272	WRKY_190		
273	WRKY_195		

CONCLUSION PUBLICATION 2

Il est aujourd'hui communément accepté que la topologe *scale-free* est caractéristique des grands réseaux biologiques. La plupart des propriétés fonctionnelles d'une cellule sont portées par des groupes de molécules qui ne sont pas isolés mais plutôt liés les uns aux autres ou même chevauchants. On peut donc supposer qu'une méthode d'inférence de réseau cohérente pour l'application à des données biologiques devrait tendre vers cette topologie *scale-free*. Une caractéristique principale des réseaux *scale-free* est la présence de nœuds de connectivités très différentes, à savoir des nœuds faiblement connectés représentant des molécules au champ d'action restreint et des hubs représentant des régulateurs importants. Ici nous avons montré que RulNet infère bien des réseaux à la topologie *scale-free* (Figure II-4 et II-S1), ou tout du moins qu'il existe des paramètres seuils permettant de s'en approcher. Dans les résultats présentés pour comparer les réseaux inférés par l'approche et des méthodes communément utilisées (Figures II-4,et II-5), une sémantique simple a été conçue mais toute la puissance de l'approche réside dans la possibilité de concevoir une infinité de sémantiques capables de répondre à des problématiques biologiques différentes.

Dans l'article, nous avons mis en exergue une particularité de l'approche, celle de pouvoir être utilisée comme une méthode de « clustering dirigé » (Figure II-6). Les sémantiques utilisées dans la partie de l'article concernée et l'utilisation des « attributs centraux » ont permis de mettre en relation les données omiques, que ce soit les FT, les métabolites ou les PR et de résumer, leur comportement dans les données utilisées et leurs éventuels liens avec les niveaux d'approvisionnement en N et en S ou des phases de développement du grain. Dans cette partie, plutôt que des liens d'interaction, ce sont des relations fonctionnelles qui sont inférées. Globalement, les résultats obtenus sont consistants avec nos connaissances de la régulation des PR et des métabolites.

Dans ce travail, nous illustrons également l'efficacité de l'approche et de la plate-forme dans le cadre d'une étude sur des données omiques. Il est aujourd'hui admis que les FT sont interdépendants, c'est à dire qu'ils agissent le plus souvent en combinaison avec d'autres FT. Dans l'approche RulNet, bien que cette caractéristique ne soit pas illustrée dans cet article, différents paramètres permettent de découvrir des règles impliquant plusieurs antécédents (n-1 MP-règles), idéalement adapté à l'étude de réseaux d'interactions impliquant des FT.

Bien que l'accumulation des données omiques ces dernières années soit très importante, il existe un problème de dimensionnalité des données. Les technologies utilisées permettent aujourd'hui, particulièrement au niveau transcriptomique, de générer des données à l'échelle

du génome. En revanche, les limites expérimentales permettent très difficilement d'obtenir un nombre d'observations compatible avec les méthodes d'inférence de réseaux. Ce nombre d'observations restreint peut limiter la portée des résultats d'inférence même lorsque les résultats semblent consistants avec des résultats obtenus par des approches fonctionnelles.

Aucune méthode d'inférence n'est actuellement capable d'annoncer une précision absolue dans l'application à des données biologiques, il existe de plus, de fortes divergences entre les résultats obtenus entre les différentes approches. Des études comparatives des méthodes d'inférence de réseaux permettent aujourd'hui d'affirmer que les meilleurs résultats peuvent être obtenus en tirant parti de la complémentarité des approches (Marbach *et al.*, 2012). L'approche développée ici s'inscrit dans cette recherche de nouvelles méthodes d'inférence capables de participer à cet effort.

Dans la suite de ce travail de thèse, nous avons choisi d'étudier d'abord tout particulièrement le transcriptome des FT dans la plante, et dans le grain dans diverses conditions d'approvisionnement en N et S. Le niveau transcriptionnel de la régulation ayant été mis en évidence comme niveau principale de la régulation des PR, cette étude nous permettra de dégager des FT candidats dans le développement du grain en fonction de la disponibilité en N et S. Ensuite, nous avons mis en place une approche systémique afin de dégager les différentes voies métaboliques en relation avec l'accumulation des PR. L'approche d'inférence de réseaux que nous avons développée dans la partie précédente est utilisée régulièrement dans les différents contextes de la suite de l'étude.

RÉFÉRENCES

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G. **2012b**. Wisdom of crowds for robust gene network inference. *Nat Meth* **9**: 796–804.

PUBLICATION 3

Transcriptome analysis of wheat transcription factors reveals key regulators of grain development and their response to nitrogen and sulfur

Jonathan Vincent[1,2,3], Pierre Martre[1,2], Zhanwu Dai[1,2,†], Isabelle Romeuf[1,2], Camille Rustenholz[1,2,††], Alicia Besson[1,2], Mireille Dardevet[1,2], Etienne Paux[1,2], Marie Pailloux[3] and Catherine Ravel[1,2]

[1]INRA, UMR1095 Génétique, Diversité et Ecophysiologie des Céréales, 5 chemin de Beaulieu, F-63 039 Clermont-Ferrand, France; [2]Université Blaise Pascal, UMR1095 Diversité et Ecophysiologie des Céréales, F-63 177 Aubière, France; [3]Université Blaise Pascal, UMR6158 CNRS LIMOS Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, F-63 173 Aubière, France; [†]Current address: INRA, UMR1287 Ecophysiologie et Génomique Fonctionnelle de la Vigne, Institut des Sciences de la Vigne et du Vin, F-33 882 Villenave d'Ornon, France; [††]Current address: Université de Strasbourg, UMR1131 Santé de la Vigne et Qualité du Vin, F-68 021 Colmar, France.

ABSTRACT

Bread wheat (*Triticum aestivum* L.) is a staple crop of major importance. Study of the global transcriptional time and/or organ specific programming as well as transcriptional regulation arising from nitrogen (N) and sulfur (S) supply should provide valuable information toward nutritional value improvement. Three distinct microarray experiments were carried out to measure the expression of 40 642 genes, including 2 891 transcription factors (TFs) in 68 conditions of development, organ and nutrition. A comprehensive TF transcriptome atlas based on these samples was developed, representing five organs at key developmental stages, grain development. The expression of TF in grains in response to N and S supply was also studied. Specificities of wheat grain transcriptome were highlighted as well as its underlying developmental reprogramming and the effects of N and S supply. The results showed that TFs carry the original information underlying transcriptional signature of organs and transcriptional reprogramming of wheat grain. Overall, TFs were not different to other genes in terms of level of expression. Seventy grain-specific TFs were discovered. The inference of regulatory gene network highlighted transcription factors implied in grain response to N and S supply. Even more than N and S independently, N:S balance had a major influence on wheat transcriptome.

Key words: gene regulatory networks, grain development, nitrogen, sulfur, transcription factor, transcriptome, bread wheat (*Triticum aestivum* L).

INTRODUCTION

In 2012 the annual global production of wheat (*Triticum aestivum* L.) was estimated at more than 700 million of tones, it is the main cereal crop regarding the area of growth, the most consumed and the main staple food in various regions around the world (FAO 2013). As a large part of the world's food calories and proteins come from cereal grains, extensive research efforts are carried out to improve cereal grain productivity, end-use quality and nutritional value. However, the grain is a complex organ, and improvement, either through genetics or agronomic solutions, requires an understanding of mechanisms governing its development. Knowledge of the molecular mechanisms underlying wheat grain development will be helpful not only for further improvement but also as a model for the exploration of developmental mechanisms of other angiosperms.

Balanced nutrition is key to cultivating high yield and high quality crops (Altenbach *et al.,* 2002; Triboi *et al.*, 2000). In wheat, grain storage proteins (GSPs) content and composition are essential as they are the main determinants of flour and dough end-use value. High N supply via the use of fertilizers increases not only grain dry mass and protein yields but also modifies the ratios between GSP fractions (Gupta *et al.*, 2003; Zhu *et al.*, 1999). Sulfur nutrition has also a major impact on crop growth and nitrogen economy but even more in terms of quality. High S supply favor the accumulation of S-rich protein fractions (DuPont *et al.*, 2006; Wieser *et al.*, 2004) and has profound effect on GSP composition and dough functional properties (Zhao *et al.*, 1999). Sulfur deficiency may also led to high concentration of free asparagine in the grains, the precursor of the suspect carcinogen acrylamide (Muttucumaru *et al*., 2006).

Although several reports show that GSP accumulation is mainly regulated at the transcriptional level in cereals (reviewed by Verdier et al., 2008; She *et al*., 2011), knowledge of the underlying molecular mechanisms remains elusive. Transcriptional activation or repression, as well as the processing of transcripts are critical for a finely tuned control of development. Transcription factors (TFs) recognize and undergo sequence-specific binding to the promoter regions of target genes and thus regulate promoter activity and gene transcription (Brivanlou & Darnell, 2002). Transcriptional regulation of a gene often involves interactions between multiple TFs and the complex regulation networks often lead to the control of gene expression via the concomitance of multiple signaling pathways. In terms of coordinated plant development, each process is the result of extensive regulatory interactions between TFs, including hierarchical, cross-regulatory and auto-regulatory

interactions (Benfey & Weigel, 2001). TFs have been identified by comprehensive analyses in various plant species (Gao *et al.*, 2006; Le *et al.*, 2010; Riechmann, 2000; Rushton *et al.*, 2008; Zhu *et al.*, 2007). Genomic studies of *Arabidopsis thaliana* revealed that about 5% of its genes encode TFs (Riechmann, 2000). However, less than 10% of these TFs have been genetically or functionally characterized to date. Hexaploid wheat genome is substantially larger than *A. thaliana* and composed of three homoeologous genomes (2n = 6× = 42, AABBDD), thus generally each gene is represented by three homoeologous copies. The complete genomic sequence od wheat is not yet available. However, extensive cDNA libraries have been constructed (Mochida *et al.*, 2009) as well as EST libraries (Boguski *et al.*, 1993; Manickavelu *et al.*, 2012), and a thorough analysis of expressed sequences available for wheat has led to the identification of 2,891 TFs (Romeuf *et al.*, 2010).

Profiling transcript abundance is currently the most efficient, scalable, and informative method for investigating complex biological systems at the molecular level. To date, many studies have been carried out using microarrays to identify genes that are expressed at different stages of development in different organs (Benedito *et al.*, 2008; Jiao *et al.*, 2009; Ma *et al.*, 2005; Schmid *et al.*, 2005; Schreiber *et al.*, 2009). These studies showed that organs can be differentiated based on their transcriptional signatures, with sets of genes expressed only in specific organs and at definite stages of development. Profiling of gene expression during seed and grain development has been carried out for *A. thaliana* (De Folter *et al.*, 2004; Le *et al.*, 2010; Ruuska *et al.*, 2002), barley (*Hordeum vulgare* L.) (Druka *et al.*, 2006), *Medicago truncatula* (Verdier *et al.*, 2008), soybean (*Glycine max* L.) (Libault *et al.*, 2010) and wheat (Capron et al., 2012; Drea et al., 2005; Laudencia-Chingcuanco et al., 2007). A study focusing on *A. thaliana* seed development allowed the identification of 48 TFs required for grain development and expressed exclusively in grain (Le *et al.*, 2010). Among them, seven were already known to be key regulators of grain development including three Leafy Cotyledon TFs.

DNA microarrays are particularly helpful for crop species where little genome information is available both to study a specific biological process or to identify candidate target genes for breeding (Rensink *et al.*, 2005). In this studied we used A 40k custom expression microarray for wheat based on the UniGene wheat set build #55 (Rustenholz *et al.*, 2011) to follow the expression of 40,642 wheat genes, including 2,891 TFs (Romeuf *et al.* 2010). Three microarray experiments were used to shed light on TFs involved in the regulation of wheat grain development and their response to N and S availability. A first experiment referred as “plant development” (hereafter PD) allowed us to observe the expression of TFs from multiple

organs sampled throughout the plant life cycle (Rustenholz *et al.*, 2011). In a second experiment referred as "endosperm development" (hereafter ED), transcription profiling was performed across 11 grain developmental stages to better understand the transcriptional regulation underlying grain development. Finally an experiment referred as "N/S response" (hereafter NSR) was carried out to study the effect of different N and S supply on TFs expression at 14 stages of grain development.

MATERIALS AND METHODS

Plant growth conditions and experiments setup

"Plant development" (PD) experiment

The expression of TFs at early, mid and late stages of root, leaf, stem, ear and grain development was studied for the spring wheat (*Triticum aestivum* L.) cv. Chinese Spring. Seeds were shown in pots filled with a mixture of peat and pozzolana and were grown in a greenhouse. The plants received a mean daily photosynthetic photon flux density (PPFD) of 26 mol m^{-2} d^{-1} (80% of ambient solar radiation), with daily maximum/minimum air temperatures averaging 21°C/15°C and day/night relative humidity averaging 50%/60%.

Roots were sampled at growth stage (GS; Zadoks *et al.*, 1974) 10 (first leaf through coleoptile), 13 (three leaves unfolded), and 39 (flag leaf ligule just visible); leaves (laminae) at GS 10, 23 (main shoot and three tillers), and 71 (grain water ripe; i.e. 50°Cdays after anthesis); stem at GS 30 (ear at 1 cm), 32 (second node detectable), and 65 (anthesis half-way); ears at GS 32, 39, and 65; and grains at GS 71, 75 (medium milk; 300°Cdays after anthesis), and 85 (soft dough; 700°Cdays after anthesis). Tissues were immediately frozen in liquid N and then stored at -80°C prior to RNA extraction. Growth stages were determined following Tottam *et al.*, (1987). Two biological replicates were used.

"Endosperm development" (ED) experiment

The expression of TFs at 11 stages of grain development was studied for the winter wheat cv. Récital. Seeds were shown in 294-mL pots filled with a peat moss mixture and were kept in a greenhouse until they reach GS 13. Light/dark air temperatures in the greenhouse were controlled at 18°C/10°C and air relative humidity at 70%/50% (light/dark). Then plants were vernalized for 8 weeks in a growth chamber were the air temperature was maintained at 4°C ± 1°C, the air relative humidity at 40%, the average daily PPFD at the top of the plants was 43 mmol m^{-2} d^{-1} during the 8-h photoperiod. After vernalization, the plants were transplanted to soil beds in a greenhouse with a distance between plants of 30 cm and a row spacing of 0.5

m. Plants were watered as needed to maintain a soil water potential in the rooting zone higher than -0.1 MPa and received 30 g m^{-2} of nitrogen:phosphorus:potassium (17:17:17) fertilizer when transferred to the greenhouse. The conditions in the greenhouse were as described for the PD experiment.

Ears were tagged when the anthers of the central florets appeared (anthesis). Ears were sampled every about 50°Cd between 132 and 686°Cd after anthesis, thus covering the cellularization, cell division, and active grain filling periods of grain development. The basal grains on the spikelets of the central third of the ear were collected, the embryo and the external pericarp were rapidly removed, and the endosperm was frozen in liquid N and stored at -80°C prior to RNA extraction.

"N/S response" (NSR) experiment

The expression of TFs in response to varying N and S supply was studied for the cv. Récital. Seeds were sown in 50-mL plastic pots (two seeds per pot) filled with a peat moss mix and were kept in a greenhouse until they reach GS 13. After 8 weeks of vernalization (in the same conditions as for the ED experiment), the plants were transplanted into polyvinyl chloride (PVC) columns (1.9 L, i.d. 7.5 cm, length 50 cm) filled with a 1:1 (v:v) washed perlite:river sand mix and were arranged in a 10 m^{-2} walk-in growth chamber in a randomized complete block design with four N/S treatments and three blocks to form an homogeneous stand with a plant density of 261 plant m^{-2}. The conditions in the growth chamber were 20°C/15°C (day/night), 55%/75% air relative humidity (day/night), with an average PPFD of 550 µmol m^{-2} s^{-1} at top of the plants during the 16 h photoperiod.

Each PVC column received 68 mL of a modified Hoagland's nutrient solution (Castle and Randall, 1987) four times per day and 170 mL of demineralized water at the end of the dark period to eliminate any potential enrichment of nutrients in the substrate. All nutrient solutions were prepared with demineralized water. For the control treatment (termed NS) the nutrient solution contained 1 mM KH_2PO_4, 1 mM $Ca(NO_3)_2$, 0.5 mM NH_4NO_3, 2 mM $MgSO_4$, 3 mM $CaCl_2$, 5 mM KCl, 10 µM H_3BO_3, 0.7 mM $ZnCl_2$, 0.4 µM $CuCl_2$, 4.5 µM $MnCl_2$, 0.22 µM MoO_3, and 50 µM EDS-Fe. Three different combinations of N and S supply were applied with either 3 mM N or 2 mM S (treatment NS), 3 mM N and 0.02 mM S (normal N and low S, treatment NS-), or 15 mM N and 0.02 mM S (supra-optimal N and low S, treatment N+S-). Treatment NS- was obtained from treatment NS by reducing $MgSO_4$ concentration to 0.02 mM and adding 1.98 mM $MgCl_2$ to keep Mg concentration constant. Treatment N+S- was obtained from treatment NS- by increasing NH_4NO_3 to 1 mM and $Ca(NO_3)_2$ to 4 mM to replace

$CaCl_2$ and replacing KCl by 5 mM KNO_3. To obtain a large range of N:S ratio in grains, after anthesis all N containing compounds were excluded from the NS and NS- solutions while S concentration was kept unchanged. The nutrient solution of treatment N+S- was kept unchanged until ripeness maturity. Three additional treatments were applied to each of the three main treatments, complementing the nutrient solution with N and/or S from 490°Cdays after anthesis (i.e. midway through the effective grain filling phase) to grain ripeness maturity. Treatment NS was shifted to N+S, NS- to N+S-, and N+S- to N+S. Treatment N+S- was obtained from Treatment NS- by removing $MgCl_2$ and increasing $MgSO_4$ concentration to 2 mM.

Main-stem ears were tagged when the anthers of the central florets appeared and they were sampled every 90°Cd from 200°Cd after anthesis to ripeness maturity (ca. 1150°Cd after anthesis). To obtain representative grain samples, 8 (from 200 to 370°Cds after anthesis) to 4 (from 460 to 1150°Cd after anthesis) main-stem ears were sampled to form a biological replicate for each treatment. Three biological replicates were used unless indicated otherwise.

Total RNA extraction and double-stranded cDNA synthesis

Tissues were disrupted in liquid N using a TissueLyser® grinder (Quiagen). Total RNA was extracted as described in Khaled *et al.* (2005). A lysis buffer was added to tissue powder to disrupt cells. Extractions were made with phenol/chloroform and nucleic acids were precipitated with ethanol. DNAse treatment was performed with the DNA-free™ kit (ambion). Nucleic acids and RNA quality were controlled on agarose gel electrophoresis before and after DNAse treatment.

The SuperScript® Double-Stranded cDNA synthesis kit (invitrogen) was used to turn RNA into double-stranded cDNA. The first strand was transcribed with an oligo dT primer and the Superscript™ II Reverse Transcriptase. The synthesis of the second strand was performed according to the method of Gubler and Hoffman (Gubler & Hoffman, 1983) using RNAse H, DNA ligase, DNA polymerase I.

Dessign of the custom 40k NimbleGen® wheat microarray

Transcriptome analyses were carried out using a custom 40k NimbleGen® (Roche Diagnostics) wheat microarray (ArrayExpress ref. A-MEXP-1928) comprising 39,179 sequences from the *T. aestivum* National Center for Biotechnology Information (NCBI) UniGene (build no. 55) and 1,613 TFs from the wDBTF database (Romeuf et al., 2010). The design of the microarray is described elsewhere (Rustenholz *et al.*, 2011). Below we briefly describe the enrichment of the microarray in TF sequences.

For polyploid species such as wheat, the UniGene clustering method overshadows homoeologous sequences, while the TIGR assembly used to identify TFs in wDBTF distributes alternative splice variants and closely related sequences (potentially three homoeologous copies in wheat) in different contigs (Quackenbush *et al.*, 2001). Therefore we first compared the 7,112 TF sequences from the *in silico* study of TIGR assembly sequences to obtain only one group for homeologous copies (or alternative splicing). This step allowed distinguishing 4,538 groups of genes that were given a unique name. A sequence homology search (e-value $< 1 \times 10^{-100}$) showed that 62% of the 4,538 TF sequences were not present in the UniGene set. Sequences from 1,156 additional groups were added to those of UniGene. Therefore, the microarray used in this study provides the opportunity to follow the expression of 40,642 genes, among them, 64% of the TFs identified in wheat. Each transcript is represented on the microarray by three 60-mer probes. All TF families and subfamilies identified by *in silico* analysis have members on the microarray and the percentage of genes represented is different from one family to another (Supporting Information, Fig. S1 and Table S1).

The microarray was previously validated by comparing expression results with those obtainted by qRT-PCR (Rustenholz et al., 2011; Capron et al., 2012). Pearson's coefficient of correlation between the two methods is > 0.7 and genes have similar pattern of expression during development for the two methods.

Microarray hybridization, image analysis and data availability

For the PD experiment, dual-color hybridizations were performed for all the duplicate cDNAs making four repetitions for each of the 15 samples. cDNA labeling was carried out using the NimbleGen Dual-Color DNA labeling kit (Roche NimbleGen) according to the manufacturer's procedure for gene expression analysis. The pairs of samples were randomly chosen and allocated to a plex of the microarray. For the ED experiment, each biological replicate consisting of 11 samples was hybridized on independent microarrays. The cDNA was labeled with Cy3 by random priming using the NimbleGen® One-Color DNA Labeling Kit. For the NSR experiment, a partial dye swap was applied where two samples in each array were labeled with two colors. An ANOVA analysis ($\alpha = 0.001$) on the preprocessed data of the dual-colored samples did not detect any significant effects of dye color or array, assuring the reliability of other samples without dye swap.

All hybridizations were performed using NimbleGen® hybridization kit, according to the manufacturer's instructions. Microarrays were washed using NimbleGen® Wash Buffer Kit and NimbleGen® Array Processing Accessories. Microarray images of each slide were

obtained with an InnoScan900 AL microarray scanner (Innopsys) and read with Mapix program (Innopsys). Data were extracted from scanned images using NimbleScan 2.5 software (Roche NimbleGen), which allows for automated grid alignment, extraction, and generation of data files.

All expression data are available from ArrayExpress repository (http://www.ebi.ac.uk/microarray-as/ae/) and follow MIAME guidelines (http://www.mged.org/Workgroups/MIAME/miame.html). Data from the PD experiment were deposited under the accession no. E-TABM-1095 and those of the NSR experiment under the accession no. E-MTAB-1782 and E-MTAB-1920.

Data analysis

All data were analyzed using the statistical software R (R Development Core Team, 2010) and Bioconductor tools (Gentleman et al., 2004).

Microarray data processing and normalization

The procedure used to analyze the microarray data was adapted to the hybridization design and the aim of the experiment. First, each image was divided into 30 identical zones. The background intensity was estimated using the median of the intensities of the empty spots and of the spots made of random sequences for each zone and subtracted to each spot intensity within the zone. For each experiments, data were log2 transformed.

For the PD experiment, two successive lowess corrections were undertaken to correct the dye and the duplicate biases. Afterward, the corrected intensity values were checked for each gene, and the aberrant values were removed. The corrected intensity values for each gene were compared with the corrected background intensity values using *t*-tests ($\alpha = 0.05$). The median of the corrected intensity values was calculated for each significantly expressed gene and was considered as the expression value of the gene for the sample. To standardize the expression values between the 15 samples, the expression values of each significantly expressed gene were subtracted by the median of the expression values of the significantly expressed genes for the sample and then divided by the standard deviation. Finally, the absolute values were taken and the expression the genes that were not significant expressed in a sample were set to zero.

For the ED experiment, expression data were normalized using the robust multi-array average (RMA) algorithm of R package oligo (Carvalho and Irizarry, 2010). After background correction and data normalization by the quantile method, outliers among the probesets

were detected using a Grubb's test (Grubbs, 1969) with a cut-off *P*-value of 0.05 and were replaced by the mean of the two other probes values for each outlier detected. Then, expression data were scaled to standardize experimental and technical variations between microarrays. Finally, the expression of each gene for each microarray was obtained using the algorithm median polish of the R package oligo.

For the NSR experiment, expression data were quality controlled using the R package ArrayQualityMetrics (Kauffmann et al., 2009) and expression intensities were normalized with the RMA algorithm of the R package oligo.

Multi-dimensional scaling analysis of expression data

To explore similarities between samples, non-metric multi-dimensional scaling (MDS) analysis was performed using the isoMDS function of the R package MASS (Venables *et al.,* 2002). In each case, data were reduced to a two-dimensional space. In the process, matrices of similarities between objects (samples) were calculated using Euclidean distance. The quality of the MDS representations was assessed by calculating the Kruskal stress values (Kruskal & Wish, 1978).

Differential expressed transcription factors and clustering anlaysis

For the ED experiment, microarray and biological replicate effects for each sampling date were analyzed by one-way ANOVAs (α = 0.05). Differentially expressed TFs between two developmental stages (ED experiment) or among the six N and S treatments (NSR experiment) were identified using moderate *t*-test of the R package LIMMA (Smyth *et al.*, 2012). A multiple correction test based on the Benjamini and Hochberg's method (Benjamini & Hochberg, 1995) was applied with a 5% false discovery rate. For the ED experiment, TFs expression clustering based on Pearson's correlations was performed using the Hierarchical Clustering Explorer software version 3.5 (http://www.cs.umd.edu/hcil/hce/).

Directed cluster and regulatory network inference

Directed clusters and regulatory networks were inferred with an algorithm describing rule semantics between attributes (Agier et al., 2007) using the RulNet platform (http://rulnet.isima.fr). This algorithm provides a unifying framework for the discovery of rules among entities and allows defining different semantics (biological meanings) for the relationship between a given pair of variables. Semantics (Supplementary Tables S1 and S2) were defined with each corresponding to a specific kind of rules. The semantics were intended to generate rules between variables according to their relative levels among

different N/S treatments. Thresholds corresponding to low and high relative levels were empirically set to -0.5 and 0.5, respectively. During the process of rule discovery, quality metrics were computed to filter discovered rules. The following threshold values were used: support > 0.3, confidence > 0.9, and lift > 1.5. Inferred directed clusters and regulatory networks were visualized and rearranged using the Cytoscape software version 3.0 (Smoot et al., 2011). Functional modules within the constructed network were identified using the Cytoscape plug-in Clust&see (Spinelli et al., 2013).

RESULTS

Organs transcriptional signature and global TFs expression

In the PD experiment, the transcriptome of roots, leaves, stems, ears, and grains was analyzed at three different developmental stages chosen to represent key stages of development for each organ. The percentage of genes for which expression was detected in at least one of the three stages of development was similar for the five organs, ranging from 64% to 69% (Fig. 1A). However, great differences were observed when each stage of development of each organ was analyzed independently. For instance, the percentage of genes expressed in roots at male meiosis was 37% while it was 62% in stems at the ear at 1 cm stage (Fig. 1A, Table 1). No expression was detected for 6,990 (17%) genes present on the microarray while their *in silico* expression is proven by their presence in EST libraries. Expression was detected in all organs and at all developmental stages for 8,464 genes (21%) and 20,459 genes (50%) were expressed in the five organs in at least one development stages. This percentage increased in pair-wise comparisons between organs. For instance, 31,449 (77%) genes were expressed in both ears and grains (Table 2).

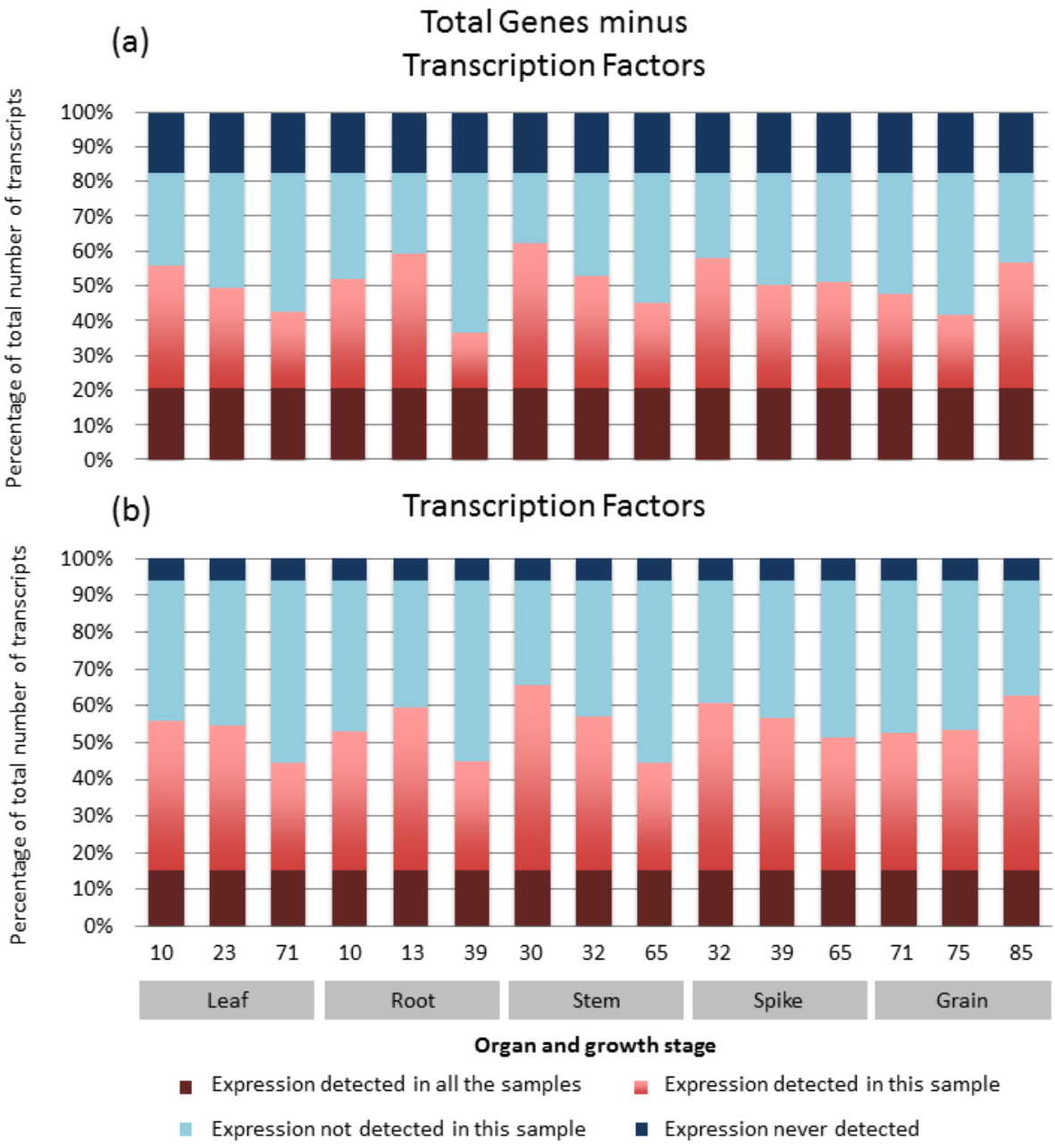


Fig. 1 Percentages of total transcripts excluding the transcription factors (a) or transcription factors only (b) expressed in different organs at different growth stages. *Growth stages are: 10, first leaf through coleoptile; 13, three leaves unfolded; 23, main shoot and three tillers; 30, ear at 1 cm; 32, second node detectable; 39, flag leaf ligule just visible; 65, anthesis half-way; 71, grain water ripe; 75, medium milk; 85, soft dough.*

This experiment allows highlighting genes with organ-specific expression whose number ranged from 508 (1.9% of expressed genes) in leaves to 874 (3.2%) in roots. Overall, 8.6% of expressed genes were organ-specific.

Table 1 Number of transcripts expressed in each organ at different stages of development. *Numbers between brackets are the number transcription factor transcripts. GS, growth stage.*

	Growth stage									
Organ	1st leaf through coleoptile (GS 10)	3 leaves unfolded (GS 13)	Main shoot and 3 tillers (GS 23)	Ear at 1 cm (GS 30)	2nd node detectable (GS 32)	Flag leaf ligule just visible (GS 39)	Anthesis half-way (GS 65)	Grain water ripe (GS 71)	Medium milk (GS 75)	Soft dough (GS 85)
Roots	21 112 (1 527)	24 105 (1 722)		-	-	14 907 (1 293)	-	-	-	-
Leaves	22 679 (1 615)		20 119 (1 583)	-	-	-	-	17 310 (1 284)	-	-
Stems	-	-		25 312 (1 893)	21 545 (1 646)	-	18 352 (1 291)	-	-	-
Ears	-	-		-	23 657 (1 638)	20 565 (1 761)	20 874 (1 488)	-	-	-
Grains	-	-		-	-	-	-	19 426 (1 524)	17 084 (1 548)	23 061 (1 812)

Ninty-two percent (2,722) of the TFs were expressed in at least one organ/development stage combination. This percentage was higher than the 82% of total expressed genes. The percentage of TFs among expressed genes varied only slightly (from 6.9% to 9.1%) from one condition to another (Table 1). There was no difference between the average level of expression of TFs and that of other genes from the microarray ($P > 0.1$).

Expression was detected in all organs and at all stages of development for 444 TFs (15%) and 1,547 TFs (53%) were expressed in the five organs in at least one stage of development. The latter percentage was comparable to the percentage of total genes expressed in the five organs (50%). In contrast to what was observed for TFs genes families in *A. thaliana* (Schmid *et al*., 2005), a great similarity in terms of family of the TFs expressed was observed between the five organs (Fig. 2) and was not different from a random distribution ($P > 0.1$). Among the 2,294 TFs expressed in grain, 960 were expressed in all three stages of grain development representing all TFs families but the NZZ family.

Only 233 TFs (8%) were specifically expressed in one organ (Table 2). Seventy TFs were specifically expressed in the grains. These TFs belonged to 19 TF families, among them, the Zinc finger, NAC, MYB, bHLH and HB were the most represented (Fig. 3A) and were more represented in this organ than in the four other organs (data not shown). The 70 TFs expressed

specifically in the grain were classified into families and subfamilies and their stage specificity was assessed in the PD experiment as well as by their profile of expression in the ED

experiment (Fig. 3).Five TFs were expressed specifically during early grain development (50°Cd after anthesis) and 27 and 29 were expressed specifically at 350 and 700°Cdays after anthesis, respectively. The transient expression of these TFs suggests that they play a specific role in the finely tuned development of grains.

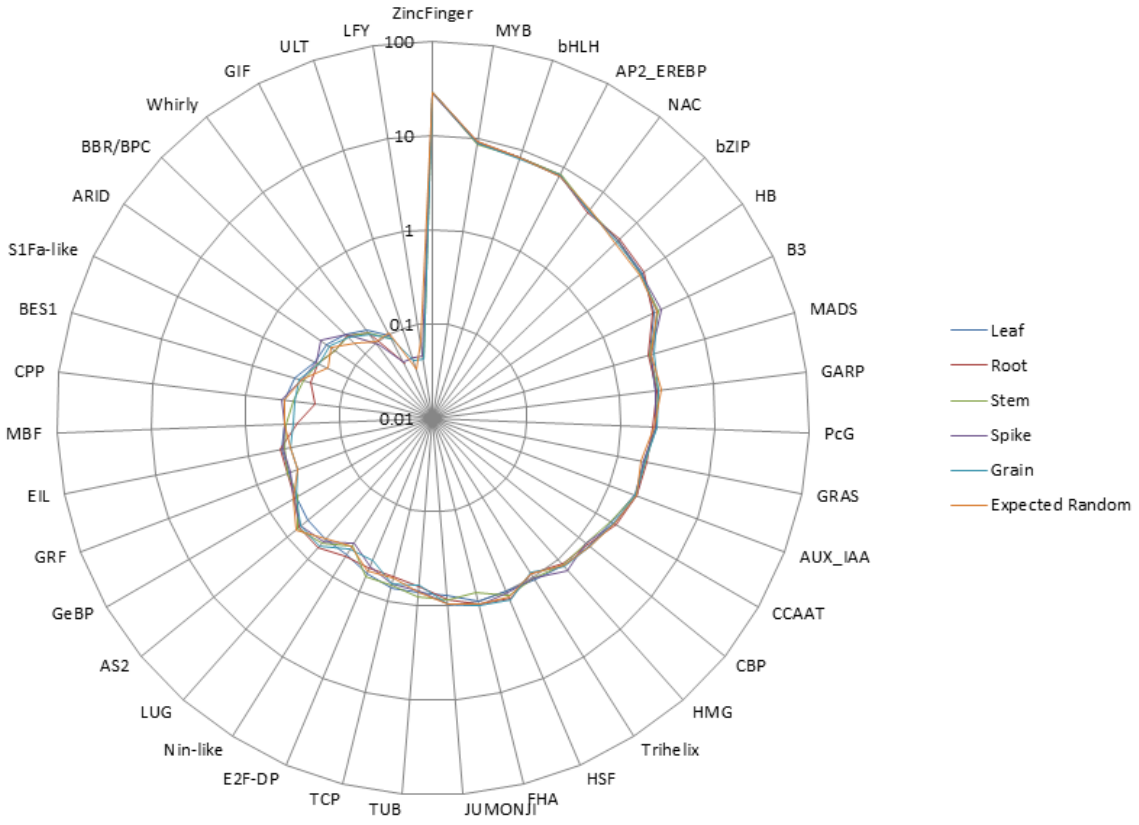


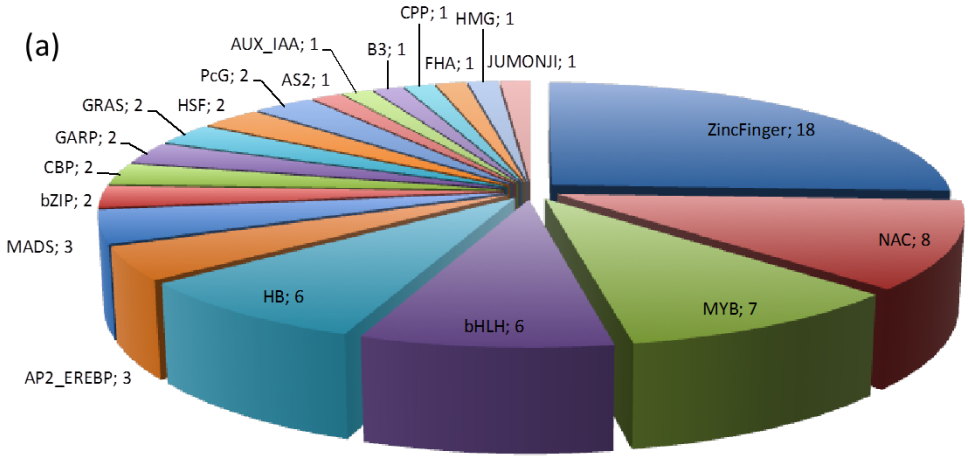
Fig. 2 Comparison of the transcription factors transcriptome of different organs during wheat plant development. *Radar chart of the families of transcription factors detected in five organs. Data are percent of total number of transcription factors expressed in each organ and the percent of total number of transcription factors spotted on the microarray. Data are plotted on a logarithmic scale.*

Roots, leaves, stem, ears, and grain have a strong transcriptional signature

The transcriptional signature of the five organs at different stages of development was assessed using non-metric MDS analysis (Fig. 4). The resultant two-dimensional plots show groups of samples with similar transcriptomes and highlight the strong transcriptional signature of organs, which is consistent with previous reports (Honys & Twell, 2003; Schmid et al., 2005).

Table 2 ***Number of genes and transcription factors specifically expresed in different organs.*** *Numbers between brackets are the number transcription factor transcripts.*

	Leaf	Root	Stem	Spike	Grain
No. of mRNAs detected specifically in each organ	508 (34)	874 (59)	579 (38)	736 (32)	804 (70)



(b)

TF name	Family	SubFamily	Time after flowering (°Cdays) 50	350	700	Development
Alfin_19	ZincFinger	Alfin				
AS2_12	AS2	AS2				
ASH1_5	PcG	ASH13				
AUX/IAA_93	AUX_IAA	AUX_IAA				
bHLH_141	bHLH	bHLH				
bHLH_239	bHLH	bHLH				
bHLH_282	bHLH	bHLH				
bHLH_33	bHLH	bHLH				
bHLH_87	bHLH	bHLH				
bHLH_89	bHLH	bHLH				
bZIP_105	bZIP	bZIP				
bZIP_77	bZIP	bZIP				
C2C2-CO-like_12	ZincFinger	C2C2-CO-like				
C2C2-CO-like_3	ZincFinger	C2C2-CO-like				
C2C2-CO-like_97	ZincFinger	C2C2-CO-like				
C2C2-GATA_52	ZincFinger	C2C2-GATA				
C2H2_112	ZincFinger	C2H2				
C3H_16	ZincFinger	C3H				
C3H_47	ZincFinger	C3H				
C3H_70	ZincFinger	C3H				
C3H_76	ZincFinger	C3H				
C3H_85	ZincFinger	C3H				
C3H_89	ZincFinger	C3H				
CPP_3	CPP	CPP				
EREBP_193	AP2_EREBP	EREBP				
FRFBP_215	AP2_FRFBP	FRFBP				
EREBP_235	AP2_EREBP	EREBP				
FHA_22	FHA	FHA				
GARP-G2-like_22	GARP	GARP-G2like				
GARP-G2-like_89	GARP	GARP-G2like				
HD-ZIP_117	HB	HD-ZIP				
HD-ZIP_123	HB	HD-ZIP				
HD-ZIP_23	HB	HD-ZIP				
HD-ZIP_4	HB	HD-ZIP				
HD-ZIP_75	HB	HD-ZIP				

TF name	Family	SubFamily	Time after flowering (°Cdays) 50	350	700	Development
HD#ZIP_78	HB	HD#ZIP				
HMG_12	HMG	HMG				
HSF_52	HSF	HSF				
HSF_8	HSF	HSF				
JUMONJI_8	JUMONJI	JUMONJI				
LAV_1	B3	LAV				
M#type_3	MADS	M#type				
MIKC#type_29	MADS	MICK				
MIKC#type_48	MADS	MICK				
MYB#related_111	MYB	related				
MYB#related_159	MYB	related				
MYB#related_166	MYB	related				
NAC_13	NAC	NAC				
NAC_14	NAC	NAC				
NAC_173	NAC	NAC				
NAC_18	NAC	NAC				
NAC_185	NAC	NAC				
NAC_223	NAC	NAC				
NAC_249	NAC	NAC				
NAC_41	NAC	NAC				
PAT1_30	GRAS	PAT				
PLATZ_1	ZincFinger	PLATZ				
R1R2R3#MYB_1	MYB	R1R2R3				
R1R2R3#MYB_5	MYB	R1R2R11				
R2R3#MYB_194	MYB	R2R87				
R2R3#MYB_36	MYB	R2R160				
RBCMT_21	PcG	RBCMT				
SBP_1	ZincFinger	SBP				
SHR_1	GRAS	SHR				
TAZ_10	CBP	TAZ				
TAZ_34	CBP	TAZ				
WRKY_112	ZincFinger	WRKY				
WRKY_15	ZincFinger	WRKY				
WRKY_64	ZincFinger	WRKY				
ZIM_96	ZincFinger	ZIM				

Fig. 3 Classification and expression of transcription factor transcripts expressed specifically in grains. *(a) Classification by family of the 70 grain-specific transcription factors. The number of representatives for each family is indicated after the family name. (b) Name, family, sub-family, and pattern of expression of the 70 grain-specific transcription factors during grain development in the PD and ED experiment.*

Development has a stronger transcriptional signature than nitrogen and sulfur supply for grains

Transcriptional reprogramming was further studied for endosperm in the ED experiment. The similarities and dissimilarities between the 11 stages of endosperm development were first assessed using the non-metric MDS analysis (Fig. 5AB). Both the TFs and the rest of the genes highlighted the same groups of samples. The first group consisted of samples taken at 132 and 172°Cdays after anthesis. The second group consisted of samples taken from 289 to 527°Cdays after anthesis and the last one was composed of samples taken from 577 to 686°Cdays after anthesis. Thus each groups represented the three main phases of grain development.

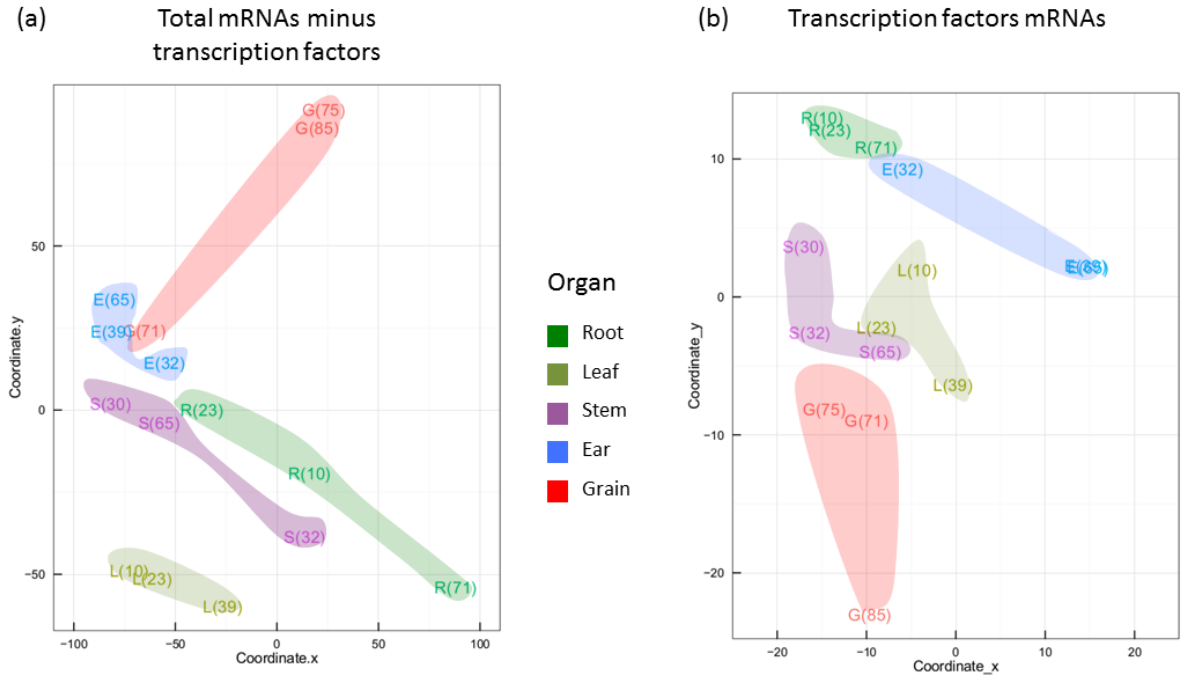


Fig 4 Visualization of organ transcriptional signature by multi-dimensional scaling. *(a) Total transcript excluding the transcription factors; (b) transcription factors. Numbers between brackets are growth stages. The 33,652 genes and 2,294 TFs expressed in grains in at least one sample in the PD experiment were used. The Kruskal stress test values were 0.12 and 0.05 for the TF and other genes, respectively.*

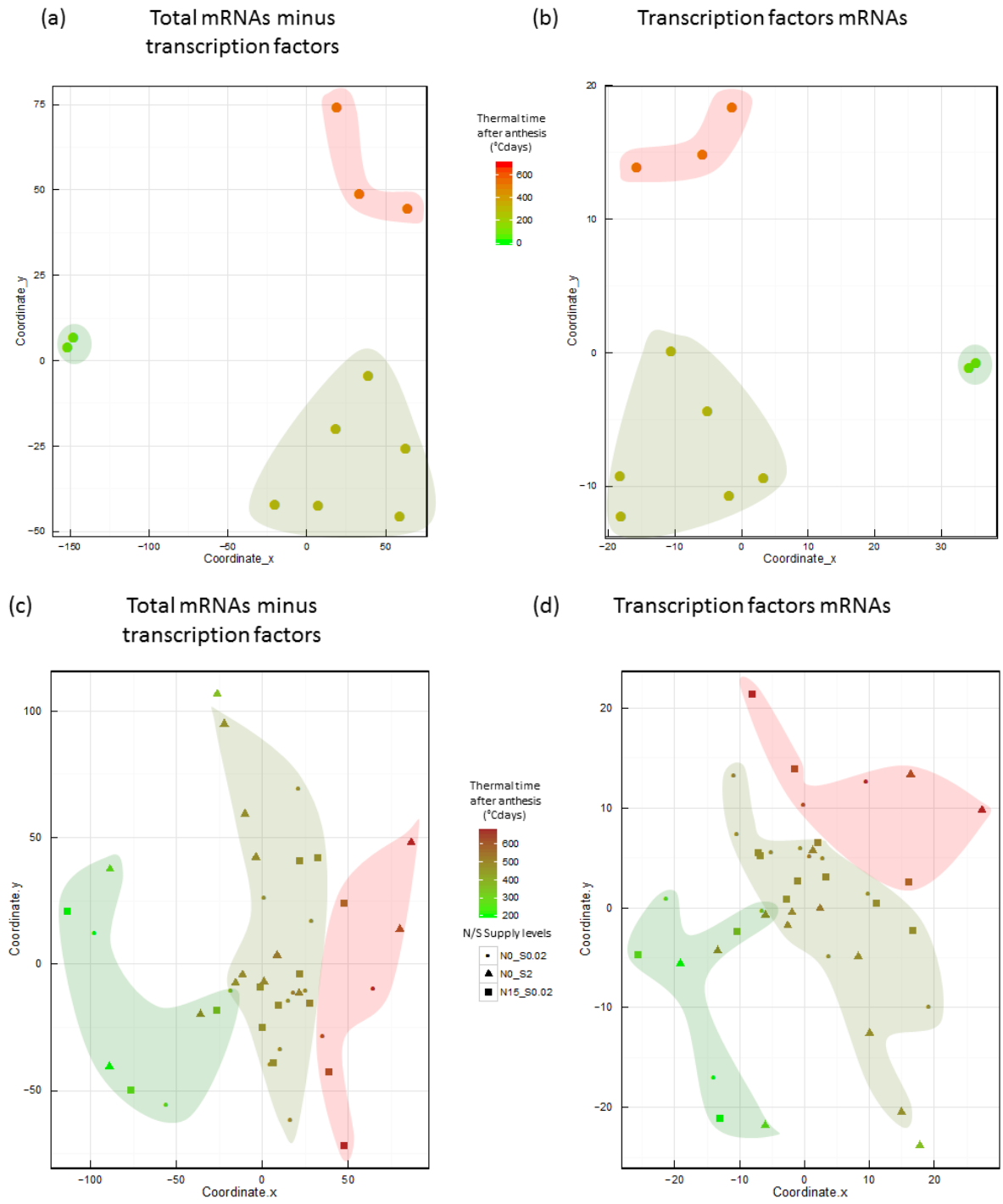


Fig. 5 Visualization by multi-dimensional scaling of transcriptional reprogramming during grain development in response to N and S supply. *(a) and (b) changes in endosperm transcriptome during grain development in the ED experiment. (c) and (d) changes in grain transcriptome in response to N and S supply. (a) and (c) total transcript excluding the transcription factors. (b) and (d) transcription factors. In (a) and (b) number above the symbols are thermal time after anthesis. Data dimension reduction was performed using a non-metric Multi-Dimensional Scaling (MDS) method. The 33,652 genes and 2,294 TFs expressed in grains in at least one sample in the PD experiment were used. Kruskal stress values were 0.13 in (a) and (b) and 0.03 and 0.04 in (c) and (d), respectively.*

The effect of N/S supply on the transcriptional reprogramming of grains was assessed for NSR experiment using the non-metric MDS analysis (Fig. 5CD). Although a larger dispersion is observed, due to the N/S effect, the transcriptional reprogramming corresponding to the three major stages of grain development can still be distinguished with higher similarities between samples from similar stages of development rather than similar treatments. The low N and high S treatments had the more marked transcriptional signature.

Differentially expressed TFs in the developing endosperm

In the ED experiment, 505 (17%) TFs were differentially expressed during endosperm development. The most represented families among these differentially expressed TFs were AP2-EREBP, MYB and NAC (each of these three families represents 9% of differentially expressed TFs), the bZIP, bHLH, B3, and C3H (5% each), then 3 the Zinc finger subfamilies (C2H2, PHD, WRKY; 4% each) and finally the MADS family (3%). Among these 505 TFs, 13 were specifically expressed in the grain (AS2_12, bHLH_282, EREBP_193, HD-ZIP_78, HSF_8, MIKC-type_29 M-type_3, MYB-related_111, NAC_13, NAC_14, NAC_223, PLATZ_1, and R1R2R3-MYB_5).

The expression pattern of the 505 TFs differentially expressed during endosperm development is shown in Figure 6. We observe major differences in TF family distribution between clusters and therefore in the proportion of different TF families expressed at different stages of wheat development. The largest cluster of TFs consists of 156 TFs that are up-regulated during the early phase of grain development. It can be grouped with cluster II, most TFs from these clusters, corresponding to the embryogenesis phase, belong to the MYB and ZyngFinger families. Eighty-two TFs show expression profiles similar to storage proteins, being up-regulated from the start of the grain-filling phase to its end.

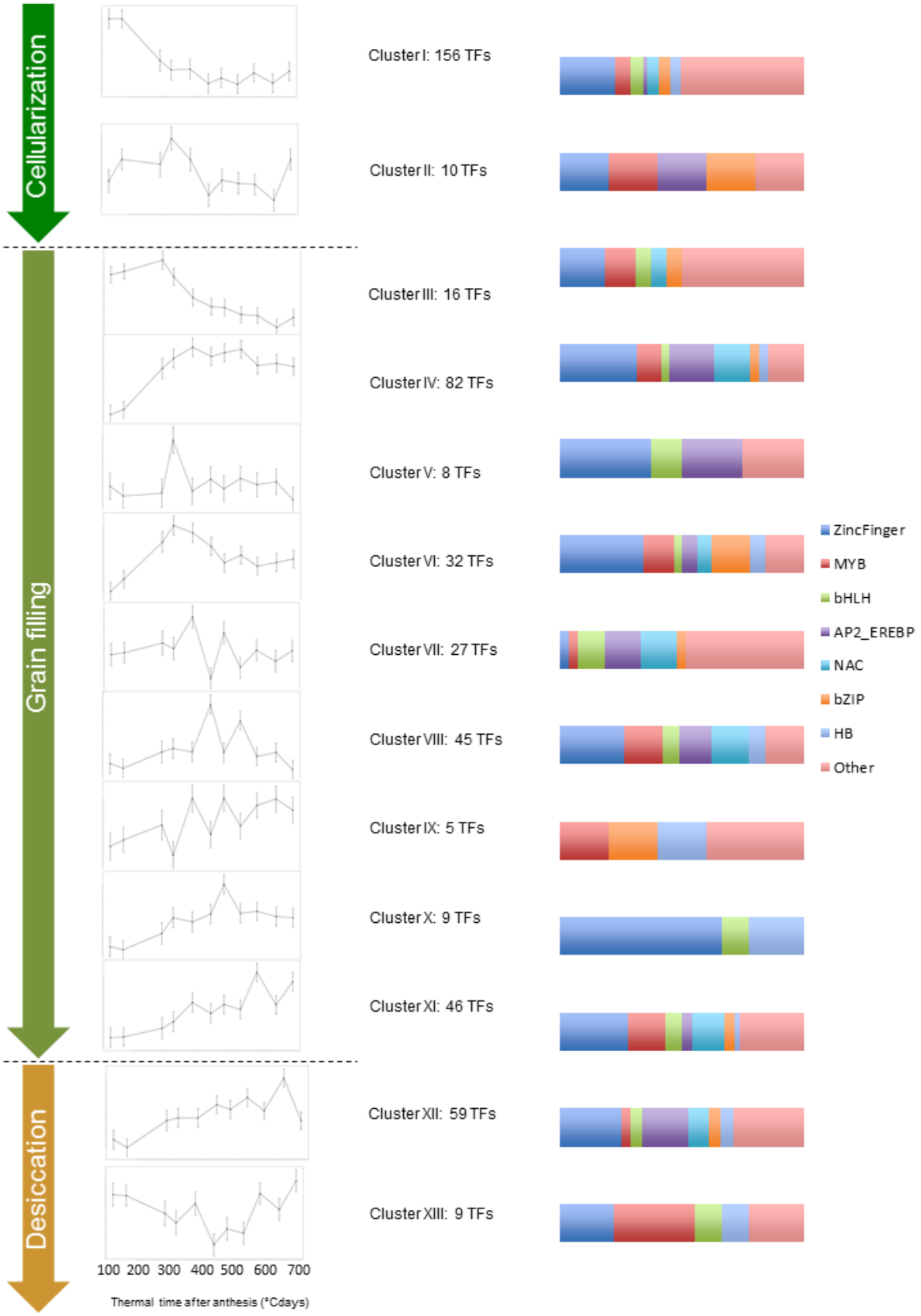


Fig. 6 Classification and expression of transcription factor transcripts expressed specifically in grains. *(a) Classification by family of the 70 grain-specific transcription factors. The number of representatives for each family is indicated after the family name. (b) Name, family, sub-family, and pattern of expression of the 70 grain-specific transcription factors during grain development.*

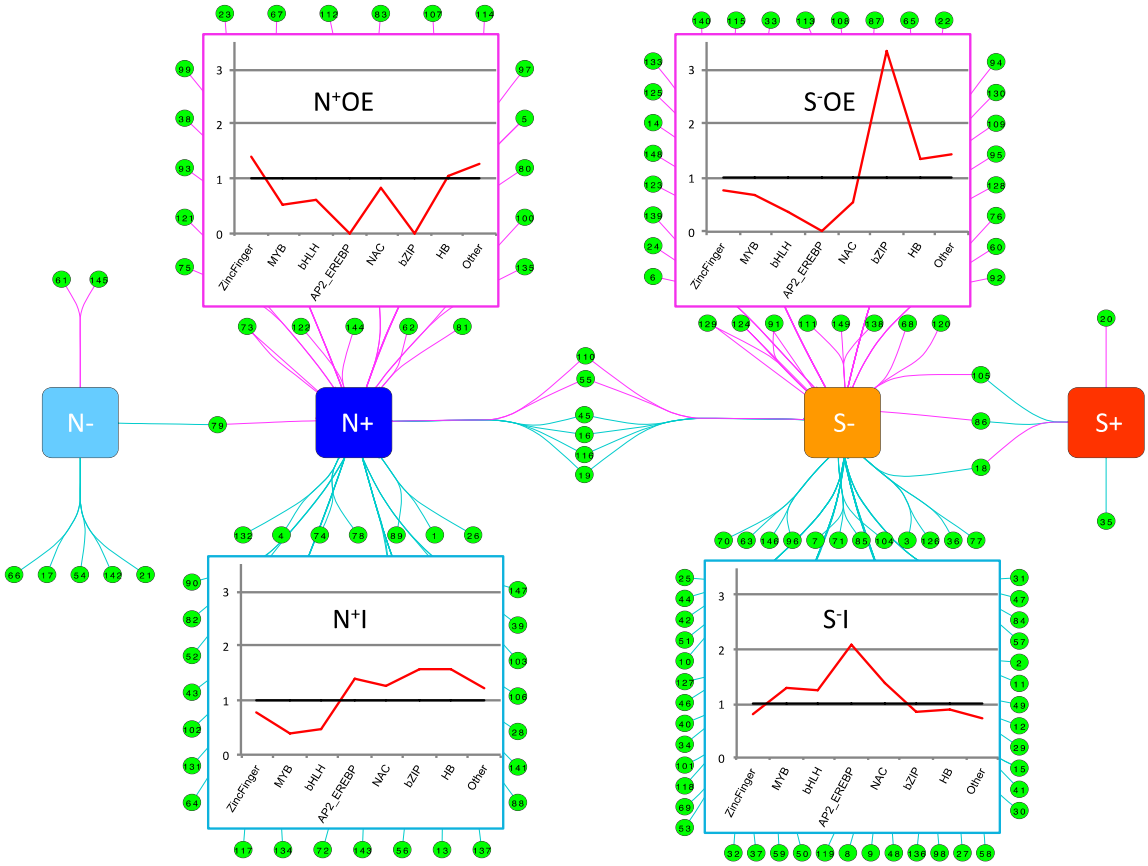


Fig. 7 Directed clustering of transcription factors associated with N and/or S supply. *Pink and light blue links indicate rules discovered with the QNS1 and QNS2 queries (see Supplmental Table S1), respectively. X-Y plots show the ratio of the percentage of transcription factor family members in each cluster to the percentage of transcription factor family members on the microarray. The correspondence between node IDs and TFs names is given in Supplemental Table S3.*

Revealing TFs involved wheat grain response to N and S supply through a network inference approach

A directed clustering anlaysis was used to highlight grain TFs sensitive to N/S supply in the NSR experiment (Fig. 8). A total of 153 TFs were associated with N and/or S supply levels. S deficiency and N excess had a greater impact than S excess and N deficiency in terms of the number of TFs affected. Fifty-six and 90 TFs were differential expressed under N excess and S deficiency, respectively, while only 5 and 8 TFs were sensitive to S excess and N deficiency, respectively. Four clusters, thereafter referred as $N^{+}OE$, $N^{+}I$, $S^{-}OE$ and $S^{-}I$ can be analyzed consistently as their size are significantly higher than for other clusters of TFs. A greater number of TFs were down-regulated in each condition compared to the number of overexpressed TFs. We sought to highlight the bias that could exist regarding the families of TFs involved in the response to N/S supply in these four clusters (insert panels in Fig. 7). The

AP2_EREBP family was the most affected by N/S supply as none of them were found in the N^{+}OE and S^{-}OE clusters while they were overrepresented in both N^{+}I and S^{-}I clusters. The bZIP family was highly overrepresented in the S^{-}OE cluster and underrepresented in the N^{+}OE cluster.

Six TFs were down- or over-regulated both in conditions of N excess and S deficiency (Fig; 7). These TFs belong to the Zinc Finger (*PHD_ZNFG_146* and *WRKY_109*), MYB (*MYB_related_115*), NAC (*NAC_246*), bZIP (*bZIP_33*) and MADS (*MIKC_type_101*) families. None of these TFs belong to the grain-specific TFs highlighted previously. Although five out of these six TFs do not have any known function in plants, *bZIP_33* is the wheat orthologous gene of *A. thaliana* At*bZIP11* TF, a sucrose-dependent regulator of asparagine and proline metabolism (Rook *et al*., 1998; Wiese *et al*., 2004; Hummel *et al.*, 2009).

Genes coregulated with these six TFs were identified using the RulNet platform. For this purpose, three queries were designed (Supplementary Table S2). The resulting regulatory network (Fig. 8) highlights three highly connected sub-networks. Sub-network 1 consists of bZIP_33, MYB_related_115 and genes belonging to different metabolisms, including stress response, energy metabolism, CHO glycolysis or hormone metabolism (Supplementary Table S3). Sub-network 2 consists mostly of genes involved in the translation process, storage of proteins (serine protease inhibitors) and S-metabolism. Sub-network 3 is a larger sub-network and contains genes involved in translation process, S-metabolism, sugar metabolism, lipid metabolism and amino acid metabolism.

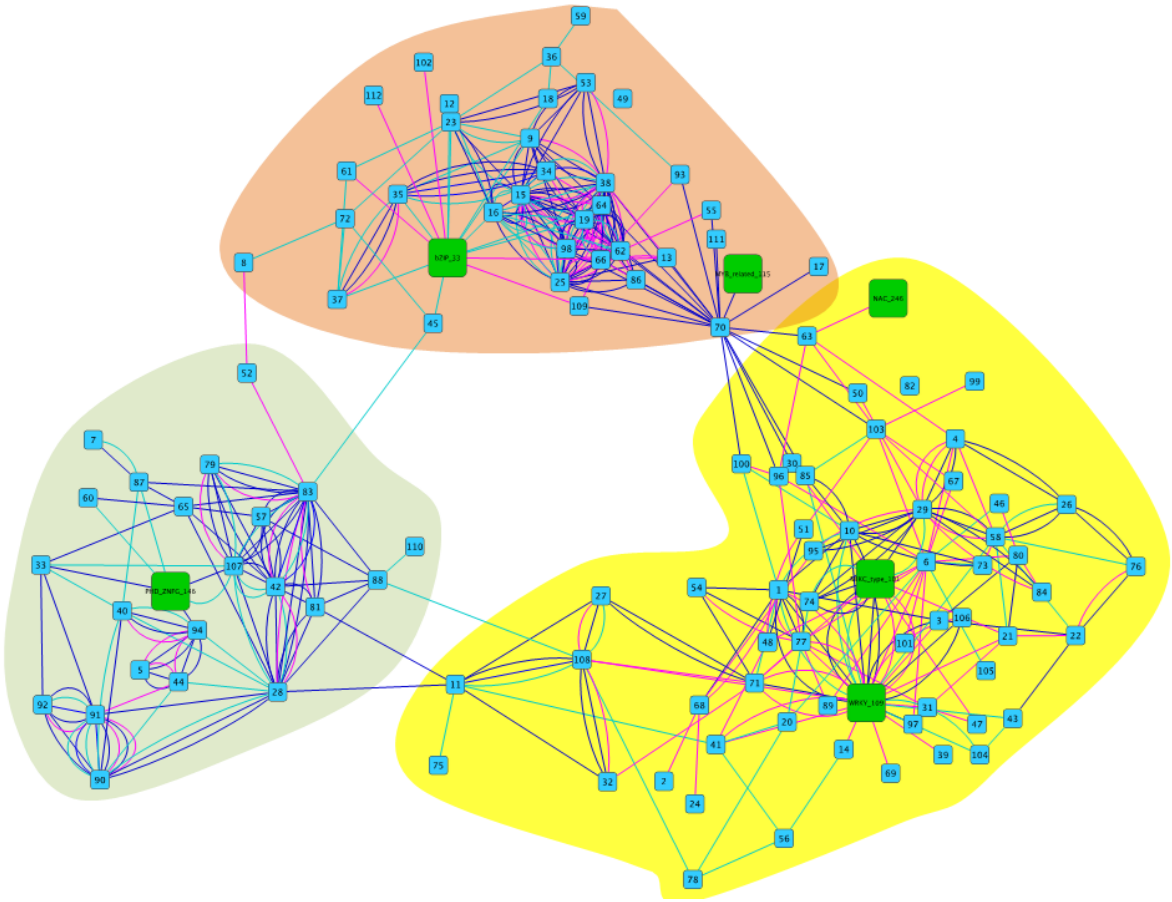


Fig. 8 Transcriptional regulatory network involving the six transcription factors associated with high nitrogen supply and low sulfur supply in Figure 8. *The color forms show the three main sub-networks. The red squares show the six transcription factors identified in Figure 8. Pink, light blue, and dark blue links indicate rules discovered with the Q1 Q2, and Q3 queries, respectively (see Supplementary Table S2).*

DISCUSSION

The analysis of gene expression in wheat using microarrays is complex considering the presence of homoeologous sequences almost identical (Poole *et al.*, 2007). Microarrays produced do not distinguish the contributions of each homoeologous. Expression profiles obtained represent the sum of individual profiles for about 90% homoeologous genes (Schreiber *et al*., 2009). Depending on sensibility of microarray experiments, all oligo-based microarray for wheat have been developed with affymetrix technology based on UniGene clustering that regroups homeologous transcripts copies. Several studies have been carried on developing wheat grains, most of them used Affymetrix Wheat GeneChip developed with UniGene build #38 that contains 19,729 UniGene sequences (Schreiber *et al*., 2009; Wan *et al*., 2008). The TIGR clustering used for TF identification, unlike the previous clustering described, distributes alternative splice variants and closely related sequences (thus potentially three homoeologous copies in wheat) in different contigs (Quackenbush *et al*., 2001). This is confirmed by our results, which indicate that 2,891 genes present on the NimbleGen® chip correspond to 5,206 sequences identified in TIGR assembly. Thus, a group of three 60-mers probes gave results for all homoeologous copies transcribed and probably various forms of alternative splicing of this gene. The same results have been described for tobacco TFs where 2,513 TFs were identified by Rushton *et al.* (2008) corresponding to 779 unique UniGenes (Edwards *et al*., 2010). Therefore, the custom NimbleGen® microarray used in this this studied is the most comprehensive tool to date to study the expression of wheat genes and TFs.

The use of microarray in this study has allowed us to measure the expression of 40,642 genes including 2,891 encoding wheat TFs. An atlas of expression of genes in five different organs has been generated and genes expressed in a organ specific manner were identified. More than 17% of the genes were never detected, in any organ and at any stage of development whereas their expression is demonstrated by their presence in EST libraries. A first explanation could be of a technical nature, directly linked to the design of the probes. However, it could also be explained biologically by the relatively low number of samples for each organ combined to the transient expression of some genes whose expression could have been missed. Considering the TFs independently brings a partial answer. Indeed, particular attention was given to the design of TF probes spotted on the microarray and only 6% of the TFs were not expressed in any organ/development stage combinations, which would tend to confirm the technical issue. On the contrary, more than a 60% (20,459/33,652) of the

expressed genes were expressed in at least 2 different organs and are therefore not organ-specific.

Out of all the genes in at least one organ/development stage combination, about 25% were expressed in all organs at all the stages of development while only 19% of the TFs expressed in at least one organ/development stage combination showed such aspecific expression and 8% of the TFs had an organ-specific expression. These results tend to confirm that TFs are key regulators of development in specific organs and at specific times of development. Interestingly, The percentage of TFs among expressed genes did not change significantly across all organs/development stages combinations, averaging 7.2%. These results are consistent with the results observed in a previous study of *A. thaliana* transcriptome (Le *et al*., 2010).

Classification of the wheat organs performed via a dimensional reduction analysis showed differences in the transcriptome between organs rather than between development stages, and suggest that plant organ identity is directly linked to their respective transcriptome. Similar results have been obtained in several studies carried out on *A. Thaliana*, soybean and medicago (Aceituno *et al.*, 2008; Libault *et al*., 2010; Schmid *et al*., 2005). Our study showed that the transcriptional signature of organs is also carried by TF. We here make an educated guess based on the well-studied function of TFs and make an assertion that such a transcriptional signature is initially carried by TFs. Even though TFs can act at very low level of expression, we found that the overall expression of TFs is not different than that of other genes, which is consistent with other studies of plant TFs. In comparison, for *A. thaliana*, contradicting results have been reported. In good agreement with our results Schmid et al., (2005) found not found difference in the average level of expression of TFs and that of other genes. In contrast Czechowski, *et al.* (2004) reported that TF genes expression was substantially lower than that of other genes in *A. thaliana*. In contrast with what is observed for multi-gene families (Dressler & Gruss, 1988), in our study there was no bias between the expressions of TFs from different families in the five organs. The organ-gene link may stems from the function of the genes rather than from a structural association.

The study of TF expression in different organs showed that 2,294 TFs are expressed in grain, suggesting that grain development requires many regulations process. Among them, 70 were grain specific. The results of microarray study showed that proportion of TFs involves in grain development in wheat (79%) is almost the same as in *A. Thaliana* (77%) (Le *et al*., 2010).

Grain global transcriptome and TF transcriptome both allowed identifying the major phases of endosperm development. Wheat grain development comprises distinct phases, starting with a syncytial, cellularization and differentiation phase characterized by active endosperm cell division, expansion and differentiation (ca. 0 to 250°Cdays after anthesis), a second phase characterized by a maximum rate of accumulation of storage polymers (ca. 250 to 550°Cdays after anthesis), and a maturation and desiccation phase (ca. 550 to 900°Cdays after anthesis) (Bennett *et al.*, 1973; Evers & Millar, 2002). Important transcriptomic reprogramming between these phases have been observed in several studies (Capron *et al*., 2012; Drea *et al*., 2005; Laudencia-Chingcuanco *et al*., 2007; Wan *et al*., 2008). Again, our results suggest that this reprogramming is carried by TFs transcriptome.

The expression of TFs was followed throughout endosperm development and a list of 505 differentially expressed TFs was established. In a comparison, in *A. Thaliana* 1,327 TFs are expressed during seed development and 48 TFs are expressed specifically in seeds (Le *et al*., 2010). In both species, these genes are active at different times of development that may correspond to transitions between two phases and therefore be important in grain development. In *A. Thaliana* transient expression experiments confirmed that these genes are active in specific organs and compartments of the grain (Le *et al*., 2010). Thus, we could make the hypothesis that the wheat 70 TFs expressed specifically in the grain identified in our study, may be key regulators of grain development. They represent 3% of TF expressed in grain. These results show that the same proportion of TF is specifically expressed in grain for *A. Thaliana* and wheat.

The most represented families in endosperm development are the NAC, MYB, bZIP, AP2-EREBP and Zinc finger, which is similar to what is observed in barley (Sreenivasulu *et al*., 2006). In addition, the TFs specifically expressed during the grain filling have been identified and belong to the families: Zinc finger (DOF, C3H C3HC4), MYB, bZIP and NAC (Sreenivasulu *et al*., 2006) suggesting a conservation regulatory mechanisms between the two species. These results are consistent with studies carried out by Schreiber *et al.* (2009), which have shown that wheat and barley transcriptome share great similarities. The involvement of many members of these families during the effective grain-filing period has been demonstrated both in mono-and dicotyledonous plants (reviewed in Verdier & Thompson, 2008).

Also, out of the 6 TFs involved in both N and S supply response; *bZIP_33* seems to be a serious candidate for further investigation. Little is known about TFs involved in different processes of wheat development, but plants use gene expression regulation to govern developmental

and growth process in response to a variety of stress. Knowledge of the function of *A. Thaliana* orthologous gene of bZIP_33 tend to confirm that it might be involved in the response to N and S supply through the regulation of amino acid metabolism.

Recent studies from closely related species highlight the relatively high modularity of TFs, being able to bind a wide range of gene promoters (Mitsuda *et al*., 2010; Ou *et al*., 2011). Such results suggest that TFs may operate as switches controlling organ- and/or condition-specific regulatory networks. Moreover, it seems that the number of TFs actually binding gene promoter under certain conditions is substantially lower than what could be expected relying exclusively on binding motifs. It is therefore not only important to study master regulators and their surrounding effectors but also more specific networks and their dynamic. Here expression data have been obtained for 2,891 TFs in 78 developmental conditions providing a precious resource for further studies of wheat TF transcriptome and as a model for other cereal species. We identified TFs likely to be involved in the response to N and S supply in grain, yet, it would be worth investigating whether these TFs act as a trigger of this particular stress response or how far upstream they stand in the signaling cascade leading to more global adjustments of the plant physiology. A systems approach could eventually provide scientists with a global regulatory network involved in suh a process.

REFERENCES

Aceituno, F. F., Moseyko, N., Rhee, S. Y., & Gutiérrez, R. a. (2008). The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in Arabidopsis thaliana. *BMC Genomics*, *9*, 438. doi:10.1186/1471-2164-9-438

Altenbach, S. B., Kothari, K. M., & Lieu, D. (2002). Environmental Conditions During Wheat Grain Development Alter Temporal Regulation of Major Gluten Protein Genes. *Cereal Chemistry*, *79*(2), 279–285. doi:10.1094/CCHEM.2002.79.2.279

Benedito, V. a, Torres-Jerez, I., Murray, J. D., Andriankaja, A., Allen, S., Kakar, K., … Udvardi, M. K. (2008). A gene expression atlas of the model legume Medicago truncatula. *The Plant Journal : For Cell and Molecular Biology*, *55*(3), 504–13. doi:10.1111/j.1365-313X.2008.03519.x

Benfey, P. N., & Weigel, D. (2001). Transcriptional networks controlling plant development. *Plant Physiology*, *125*(1), 109–11. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1539338&tool=pmcentrez&rendertype=abstract

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate -- a practical and powerful approach to multiple testing. *Journal of the Royal Statistics Society: Series B (Statistical Methodology)*, *57*, 289–300.

Bennett, M. D., Rao, M. K., Smith, J. B., & Bayliss, M. W. (1973). Cell Development in the Anther, the Ovule, and the Young Seed of Triticum aestivum L. Var. Chinese Spring. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* , *266* (875), 39–81. doi:10.1098/rstb.1973.0036

Boguski, M. S., Lowe, T. M. J., & Tolstoshev, C. M. (1993). dbEST [mdash] database for [ldquo]expressed sequence tags[rdquo]. *Nat Genet*, *4*(4), 332–333. Retrieved from http://dx.doi.org/10.1038/ng0893-332

Brivanlou, A. H., & Darnell, J. E. (2002). Signal transduction and the control of gene expression. *Science (New York, N.Y.)*, *295*(5556), 813–8. doi:10.1126/science.1066355

Capron, D., Mouzeyar, S., Boulaflous, A., Girousse, C., Rustenholz, C., Laugier, C., … Bouzidi, M. F. (2012). Transcriptional profile analysis of E3 ligase and hormone-related genes expressed during wheat grain development. *BMC Plant Biology*, *12*(1), 35. doi:10.1186/1471-2229-12-35

Czechowski, T., Bari, R. P., Stitt, M., Scheible, W.-R., & Udvardi, M. K. (2004). Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *The Plant Journal : For Cell and Molecular Biology*, *38*(2), 366–79. doi:10.1111/j.1365-313X.2004.02051.x

De Folter, S., Busscher, J., Colombo, L., Losa, A., & Angenent, G. C. (2004). Transcript profiling of transcription factor genes during silique development in Arabidopsis. *Plant Molecular Biology*, *56*(3), 351–66. doi:10.1007/s11103-004-3473-z

Drea, S., Leader, D. J., Arnold, B. C., Shaw, P., Dolan, L., & Doonan, J. H. (2005). Systematic Spatial Analysis of Gene Expression during Wheat Caryopsis Development. *The Plant Cell Online* , *17* (8), 2172–2185. doi:10.1105/tpc.105.034058

Dressler, G. R., & Gruss, P. (1988). Do multigene families regulate vertebrate development? *Trends in Genetics*, *4*(8), 214–219. doi:http://dx.doi.org/10.1016/S0168-9525(88)80003-9

Druka, A., Muehlbauer, G., Druka, I., Caldo, R., Baumann, U., Rostoks, N., … Waugh, R. (2006). An atlas of gene expression from seed to seed through barley development. *Functional & Integrative Genomics*, *6*(3), 202–11. doi:10.1007/s10142-006-0025-4

DuPont, F. M., Hurkman, W. J., Vensel, W. H., Chan, R., Lopez, R., Tanaka, C. K., & Altenbach, S. B. (2006). Differential accumulation of sulfur-rich and sulfur-poor wheat flour proteins is affected by temperature and mineral nutrition during grain development. *Journal of Cereal Science*, *44*(1), 101–112. doi:10.1016/j.jcs.2006.04.003

Edwards, K. D., Bombarely, A., Story, G. W., Allen, F., Mueller, L. a, Coates, S. a, & Jones, L. (2010). TobEA: an atlas of tobacco gene expression from seed to senescence. *BMC Genomics*, *11*, 142. doi:10.1186/1471-2164-11-142

Evers, T., & Millar, S. (2002). Cereal Grain Structure and Development: Some Implications for Quality. *Journal of Cereal Science*, *36*(3), 261–284. doi:10.1006/jcrs.2002.0435

Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W., Zheng, W., … Luo, J. (2006). DRTF: a database of rice transcription factors. *Bioinformatics (Oxford, England)*, *22*(10), 1286–7. doi:10.1093/bioinformatics/btl107

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *11*(1), 1–21. doi:10.1080/00401706.1969.10490657

Gubler, U., & Hoffman, B. (1983). A simple and very efficient method for generating cDNA libraries. *Gene*. Retrieved from http://www.sciencedirect.com/science/article/pii/0378111983902305

Gupta, R., Singh, N., & Shepherd, K. (1989). The cumulative effect of allelic variation in LMW and HMW glutenin subunits on dough properties in the progeny of two bread wheats. *Theoretical and Applied Genetics*, 57–64. Retrieved from http://link.springer.com/article/10.1007/BF00292316

Hummel, M., Rahmani, F., Smeekens, S., & Hanson, J. (2009). Sucrose-mediated translational control. *Annals of Botany*, *104*(1), 1–7. doi:10.1093/aob/mcp086

Jiao, Y., Tausta, S. L., Gandotra, N., Sun, N., Liu, T., Clay, N. K., … Nelson, T. (2009). A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nature Genetics*, *41*(2), 258–63. doi:10.1038/ng.282

Khaled, A.-S., Vernoud, V., Ingram, G. C., Perez, P., Sarda, X., & Rogowsky, P. M. (2005). Engrailed-ZmOCL1 fusions cause a transient reduction of kernel size in maize. *Plant Molecular Biology*, *58*(1), 123–39. doi:10.1007/s11103-005-5219-y

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. Sage Publications.

Laudencia-Chingcuanco, D. L., Stamova, B. S., You, F. M., Lazo, G. R., Beckles, D. M., & Anderson, O. D. (2007). Transcriptional profiling of wheat caryopsis development using cDNA microarrays. *Plant Molecular Biology*, *63*(5), 651–68. doi:10.1007/s11103-006-9114-y

Le, B. H., Cheng, C., Bui, A. Q., Wagmaister, J. a, Henry, K. F., Pelletier, J., … Goldberg, R. B. (2010). Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(18), 8063–70. doi:10.1073/pnas.1003530107

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., Franklin, L. D., … Stacey, G. (2010). An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *The Plant Journal : For Cell and Molecular Biology*, *63*(1), 86–99. doi:10.1111/j.1365-313X.2010.04222.x

Ma, L., Chen, C., Liu, X., Jiao, Y., Su, N., Li, L., … Wang, J. (2005). A microarray analysis of the rice transcriptome and its comparison to Arabidopsis. *Genome Research*, *15*(9), 1274–83. doi:10.1101/gr.3657405

Manickavelu, A., Kawaura, K., Oishi, K., Shin-I, T., Kohara, Y., Yahiaoui, N., … Ogihara, Y. (2012). Comprehensive Functional Analyses of Expressed Sequence Tags in Common Wheat (Triticum aestivum). *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes*, *19*(2), 165–77. doi:10.1093/dnares/dss001

Martre, P., Porter, J. R., & Jamieson, P. D. (2003). Modeling Grain Nitrogen Accumulation and Protein Composition to Understand the Sink / Source Regulations of Nitrogen Remobilization for Wheat, *133*(December), 1959–1967. doi:10.1104/pp.103.030585.Therefore

Mitsuda, N., Ikeda, M., Takada, S., Takiguchi, Y., Kondou, Y., Yoshizumi, T., … Ohme-Takagi, M. (2010). Efficient yeast one-/two-hybrid screening using a library composed only of transcription factors in Arabidopsis thaliana. *Plant & Cell Physiology*, *51*(12), 2145–51. doi:10.1093/pcp/pcq161

Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y., & Shinozaki, K. (2009). TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiology*, *150*(3), 1135–46. doi:10.1104/pp.109.138214

Ou, B., Yin, K.-Q., Liu, S.-N., Yang, Y., Gu, T., Wing Hui, J. M., … Qu, L.-J. (2011). A high-throughput screening system for Arabidopsis transcription factors and its application to Med25-dependent transcriptional regulation. *Molecular Plant*, *4*(3), 546–55. doi:10.1093/mp/ssr002

Poole, R., Barker, G., Wilson, I. D., Coghill, J. a, & Edwards, K. J. (2007). Measuring global gene expression in polyploidy; a cautionary note from allohexaploid wheat. *Functional & Integrative Genomics*, *7*(3), 207–19. doi:10.1007/s10142-007-0046-7

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., … White, J. (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, *29*(1), 159–64. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=29813&tool=pmcentrez&rendertype=abstract

Rensink, W. A., Lee, Y., Liu, J., Iobst, S., Ouyang, S., & Buell, C. R. (2005). Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts. *BMC Genomics*, *6*, 124. doi:10.1186/1471-2164-6-124

Riechmann, J. L. (2000). Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. *Science*, *290*(5499), 2105–2110. doi:10.1126/science.290.5499.2105

Romeuf, I., Tessier, D., Dardevet, M., Branlard, G., Charmet, G., & Ravel, C. (2010). wDBTF: an integrated database resource for studying wheat transcription factor families. *BMC Genomics*, *11*, 185. doi:10.1186/1471-2164-11-185

Rook, F., Gerrits, N., Kortstee, a, van Kampen, M., Borrias, M., Weisbeek, P., & Smeekens, S. (1998). Sucrose-specific signalling represses translation of the Arabidopsis ATB2 bZIP transcription factor gene. *The Plant Journal : For Cell and Molecular Biology*, *15*(2), 253–63. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9721683

Rushton, P. J., Bokowiec, M. T., Han, S., Zhang, H., Brannock, J. F., Chen, X., … Timko, M. P. (2008). Tobacco transcription factors: novel insights into transcriptional regulation in the Solanaceae. *Plant Physiology*, *147*(1), 280–95. doi:10.1104/pp.107.114041

Rustenholz, C., Choulet, F., Laugier, C., Safár, J., Simková, H., Dolezel, J., … Paux, E. (2011). A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant Physiology*, *157*(4), 1596–608. doi:10.1104/pp.111.183921

Ruuska, S. A., Girke, T., Benning, C., & Ohlrogge, J. B. (2002). Contrapuntal Networks of Gene Expression during Arabidopsis Seed Filling, *14*(June), 1191–1206. doi:10.1105/tpc.000877.teins.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., … Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. *Nat Genet*, *37*(5), 501–506. Retrieved from http://dx.doi.org/10.1038/ng1543

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., … Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. *Nature Genetics*, *37*(5), 501–6. doi:10.1038/ng1543

Schreiber, A. W., Sutton, T., Caldo, R. a, Kalashyan, E., Lovell, B., Mayo, G., … Baumann, U. (2009). Comparative transcriptomics in the Triticeae. *BMC Genomics*, *10*, 285. doi:10.1186/1471-2164-10-285

Smyth, G. K., Ritchie, M., & Thorne, N. (2012). Linear Models for Microarray Data User ' s Guide (Now Including RNA-Seq Data Analysis), (September).

Sreenivasulu, N., Radchuk, V., Strickert, M., Miersch, O., Weschke, W., & Wobus, U. (2006). Gene expression patterns reveal tissue-specific signaling networks controlling programmed cell death and ABA- regulated maturation in developing barley seeds. *The Plant Journal : For Cell and Molecular Biology*, *47*(2), 310–27. doi:10.1111/j.1365-313X.2006.02789.x

Triboi, E., Abad, A., Michelena, A., Lloveras, J., Ollier, J. L., & Daniel, C. (2000). Environmental effects on the quality of two wheat genotypes: 1. quantitative and qualitative variation of storage proteins. *European Journal of Agronomy*, *13*(1), 47–64. doi:http://dx.doi.org/10.1016/S1161-0301(00)00059-9

Verdier, J., Kakar, K., Gallardo, K., Signor, C., Aubert, G., Schlereth, A., … Thompson, R. (2008). Gene expression profiling of M. truncatula transcription factors identifies putative regulators of grain legume seed filling. *Plant Molecular Biology*, *67*(6), 567–580. doi:10.1007/s11103-008-9320-x

Verdier, J., & Thompson, R. D. (2008). Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant & Cell Physiology*, *49*(9), 1263–71. doi:10.1093/pcp/pcn116

Wan, Y., Poole, R. L., Huttly, A. K., Toscano-Underwood, C., Feeney, K., Welham, S., … Mitchell, R. A. (2008). Transcriptome analysis of grain development in hexaploid wheat. *BMC Genomics*, *9*, 121. doi:10.1186/1471-2164-9-121

Wiese, A., Elzinga, N., Wobbes, B., & Smeekens, S. (2004). A Conserved Upstream Open Reading Frame Mediates Sucrose-Induced Repression of Translation, *16*(July), 1717–1729. doi:10.1105/tpc.019349.Rolland

Wieser, H., Gutser, R., & von Tucher, S. (2004). Influence of sulphur fertilisation on quantities and proportions of gluten protein types in wheat flour. *Journal of Cereal Science*, *40*(3), 239–244. doi:10.1016/j.jcs.2004.05.005

Zhao, F., Salmon, S. E., Withers, P. J. a, Evans, E. J., Monaghan, J. M., Shewry, P. R., & McGrath, S. P. (1999). Responses of breadmaking quality to sulphur in three wheat varieties. *Journal of the Science of Food and Agriculture*, *79*(13), 1865–1874. doi:10.1002/(SICI)1097-0010(199910)79:13<1865::AID-JSFA449>3.3.CO;2-I

Zhu, J., Khan, K., & Huang, S. (1999). Allelic Variation at Glu-D1 Locus for High Molecular Weight (HMW) Glutenin Subunits : Quantification by Multistacking SDS-PAGE of Wheat Grown Under, (C), 4–8.

Zhu, Q.-H., Guo, A.-Y., Gao, G., Zhong, Y.-F., Xu, M., Huang, M., & Luo, J. (2007). DPTF: a database of poplar transcription factors. *Bioinformatics (Oxford, England)*, *23*(10), 1307–8. doi:10.1093/bioinformatics/btm113

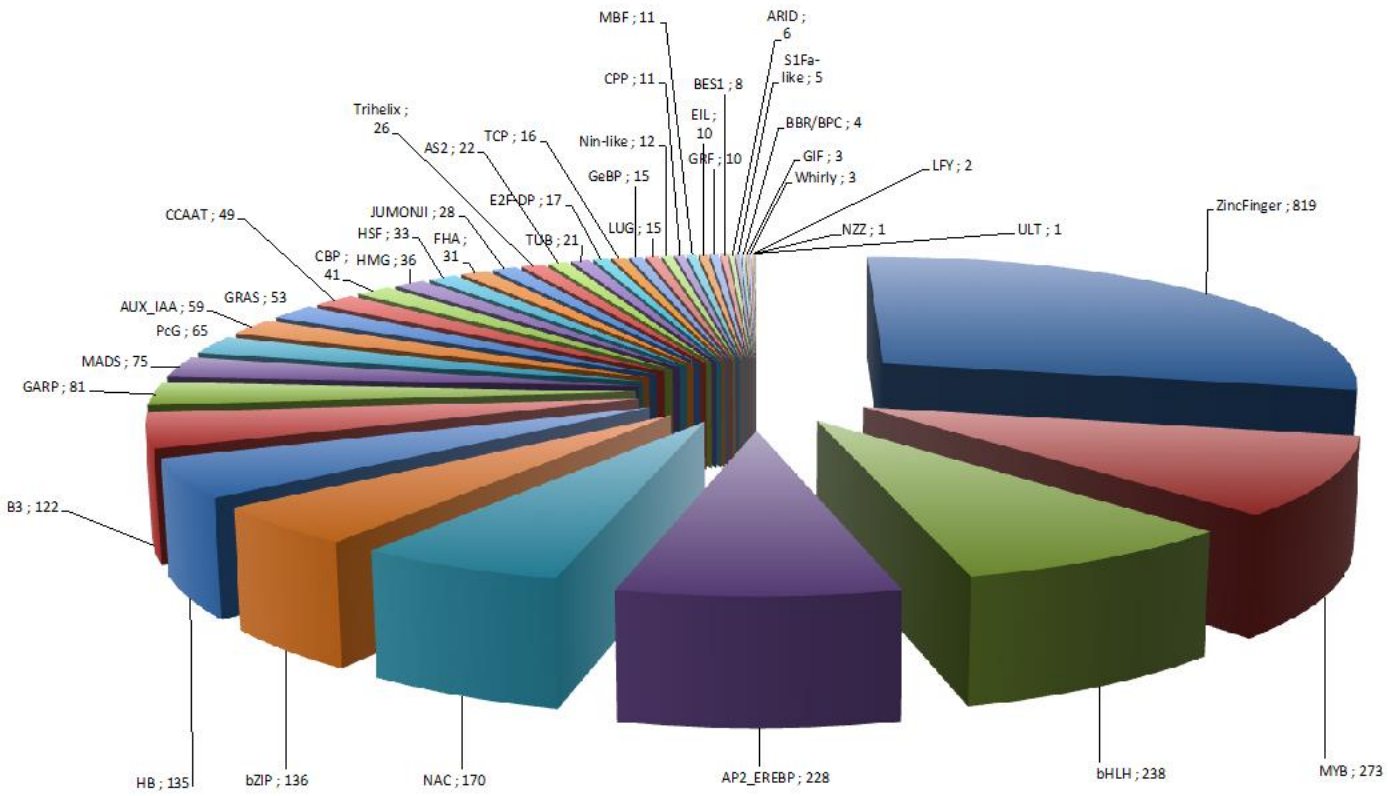


Fig. S1 Classification by family of the 2,891 transcription factors spotted on the NimbleGen® wheat microarray. **The number of representatives for each family is indicated after the family name.**

Table S1 Semantics designed to infer the clustering network.

QNS1: Discover rules between highly abundant TFs and N/S supply levels.

```
FINDRULES
SCOPE t1 IN TF_NS_data
HAVING high : t1.ATT > 0.5 OVER ALL MINUS sample, Nitrogen, Sulfur
AND NoNitrogen : t1.ATT = 0 OVER Nitrogen
AND Nitrogen : t1.ATT = 15 OVER Nitrogen
AND NoSulfur : t1.ATT = 0.2 OVER Sulfur
AND Sulfur: t1.ATT = 2 OVER Sulfur;
```

QNS2: Discover rules between not abundant TFs and N/S supply levels.

```
FINDRULES
SCOPE t1 IN TF_NS_data
HAVING high : t1.ATT < -0.5 OVER ALL MINUS sample, Nitrogen, Sulfur
AND NoNitrogen : t1.ATT = 0 OVER Nitrogen
AND Nitrogen : t1.ATT=15 OVER Nitrogen
AND NoSulfur : t1.ATT = 0.2 OVER Sulfur
AND Sulfur: t1.ATT = 2 OVER Sulfur;
```

Table S2 Semantics designed to infer the regularory network.

Q1: Discover links overexpression between genes

```
FINDRULES
SCOPE t1 IN TF_NS_data
HAVING high : t1.ATT < -0.5 OVER ALL MINUS sample, Nitrogen, Sulfur;
```

Q2: Discover links under-expression between genes

```
FINDRULES
SCOPE t1 IN TF_NS_data
HAVING high : t1.ATT > 0.5 OVER ALL MINUS sample, Nitrogen, Sulfur;
```

Q3: Discover genes with similar profiles of expression

```
FINDRULES
SCOPE t1 IN TF_NS_data
HAVING same : ABS(t2.ATT – t1.ATT) < 0.3 OVER ALL MINUS sample, Nitrogen, Sulfur;
```

Table S3 Correspondence of transcription factors identifiers and names used in Figure 8. The names are those used in the wheat transcription factor database wDBTF (http://wwwappli.nantes.inra.fr:8180/wDBFT/).

ID	TF name	TF family	TF Subfamily	ID	TF name	TF family	TF Subfamily
1	RAV_B3_2	B4	RAV	39	C2C2_GATA_11	ZincFinger	C2C2_GATA
2	AS2_20	AS2	AS2	40	EREBP_66	AP2_EREBP	EREBP
3	bHLH_92	bHLH	bHLH	41	C2H2_121	ZincFinger	C2H2
4	LISCL_10	GRAS	LISCL	42	R2R3_MYB_231	MYB	R2R132
5	Nin_like_10	Nin_like	Nin_like	43	EREBP_86	AP2_EREBP	EREBP
6	REM_46	B42	REM	44	NAC_203	NAC	NAC
7	REM_26	B28	REM	45	NAC_246	NAC	NAC
8	R2R3_MYB_227	MYB	R2R129	46	R2R3_MYB_193	MYB	R2R86
9	AP2_9	AP2_EREBP	AP2	47	NAC_187	NAC	NAC
10	C2H2_55	ZincFinger	C2H2	48	SBP_14	ZincFinger	SBP
11	C3H_119	ZincFinger	C3H	49	AUX_IAA_12	AUX_IAA	AUX_IAA
12	EREBP_21	AP2_EREBP	EREBP	50	EREBP_16	AP2_EREBP	EREBP
13	NAC_237	NAC	NAC	51	R2R3_MYB_84	MYB	R2R218
14	bZIP_163	bZIP	bZIP	52	MYB_related_29	MYB	MYB_related
15	bHLH_283	bHLH	bHLH	53	MYB_related_44	MYB	MYB_related
16	bZIP_33	bZIP	bZIP	54	CCAAT_Dr1_2	CCAAT	CCAAT_Dr6
17	bHLH_232	bHLH	bHLH	55	PHD_ZNFG_146	ZincFinger	PHD
18	EREBP_163	AP2_EREBP	EREBP	56	GRF_12	GRF	GRF
19	WRKY_109	ZincFinger	WRKY	57	REM_53	B52	REM
20	HD_ZIP_99	HB	HD_ZIP	58	Trihelix_37	Trihelix	Trihelix
21	FHA_28	FHA	FHA	59	C3H_102	ZincFinger	C3H
22	bZIP_19	bZIP	bZIP	60	bZIP_146	bZIP	bZIP
23	MYB_related_138	MYB	MYB_related	61	NAC_7	NAC	NAC
24	SUVAR_53	PcG	SUVAR	62	E2F_DP_4	E2F_DP	E2F_DP
25	PLATZ_15	ZincFinger	PLATZ	63	AUX_IAA_35	AUX_IAA	AUX_IAA
26	LIM_10	ZincFinger	LIM	64	WRKY_31	ZincFinger	WRKY
27	R2R3_MYB_61	MYB	R2R187	65	bHLH_254	bHLH	bHLH
28	MIKC_type_76	MADS	MICK	66	WRKY_7	ZincFinger	WRKY
29	bHLH_189	bHLH	bHLH	67	GeBP_11	GeBP	GeBP
30	HD_ZIP_25	HB	HD_ZIP	68	PHD_ZNFG_149	ZincFinger	PHD
31	WRKY_148	ZincFinger	WRKY	69	bZIP_2	bZIP	bZIP
32	HAM_15	GRAS	HAM	70	RAV_AP2EREBP_7	AP2_EREBP	RAV
33	JUMONJI_1	JUMONJI	JUMONJI	71	bHLH_64	bHLH	bHLH
34	bZIP_37	bZIP	bZIP	72	AS2_5	AS2	AS2
35	E2F_DP_5	E2F_DP	E2F_DP	73	PHD_ZNFG_140	ZincFinger	PHD
36	WRKY_181	ZincFinger	WRKY	74	Trihelix_34	Trihelix	Trihelix
37	FHA_7	FHA	FHA	75	ZIM_2	ZincFinger	ZIM
38	HD_ZIP_5	HB	HD_ZIP	76	PHD_HB_8	HB	PHD(HB)

Table S3. Continued

ID	TF name	TF family	TF Subfamily	ID	TF name	TF family	TF Subfamily
77	NAC_35	NAC	NAC	114	PHD_ZNFG_147	ZincFinger	PHD
78	TUB_29	TUB	TUB	115	WRKY_204	ZincFinger	WRKY
79	TRX_3	PcG	TRX	116	MIKC_type_101	MADS	MICK
80	FHA_27	FHA	FHA	117	EREBP_236	AP2_EREBP	EREBP
81	C2H2_23	ZincFinger	C2H2	118	EREBP_227	AP2_EREBP	EREBP
82	C2H2_10	ZincFinger	C2H2	119	WRKY_186	ZincFinger	WRKY
83	PHD_ZNFG_301	ZincFinger	PHD	120	EIL_19	EIL	EIL
84	HD_ZIP_67	HB	HD_ZIP	121	NAC_244	NAC	NAC
85	GRF_8	GRF	GRF	122	bHLH_119	bHLH	bHLH
86	MBF1_9	MBF	MBF32	123	PHD_ZNFG_248	ZincFinger	PHD
87	Trihelix_28	Trihelix	Trihelix	124	MYB_related_6	MYB	MYB_related
88	ZIM_90	ZincFinger	ZIM	125	GRF_14	GRF	GRF
89	GeBP_16	GeBP	GeBP	126	WRKY_90	ZincFinger	WRKY
90	EREBP_176	AP2_EREBP	EREBP	127	WRKY_169	ZincFinger	WRKY
91	TRX_4	PcG	TRX	128	PHD_ZNFG_299	ZincFinger	PHD
92	EIL_17	EIL	EIL	129	HD_ZIP_80	HB	HD_ZIP
93	E2F_DP_6	E2F_DP	E2F_DP	130	WRKY_167	ZincFinger	WRKY
94	C3H_123	ZincFinger	C3H	131	bHLH_291	bHLH	bHLH
95	bZIP_157	bZIP	bZIP	132	HD_ZIP_32	HB	HD_ZIP
96	HSF_8	HSF	HSF	133	MIKC_type_98	MADS	MICK
97	TRX_1	PcG	TRX	134	bZIP_35	bZIP	bZIP
98	EREBP_223	AP2_EREBP	EREBP	135	R2R3_MYB_90	MYB	R2R224
99	LISCL_1	GRAS	LISCL	136	AP2_27	AP2_EREBP	AP2
100	C3H_117	ZincFinger	C3H	137	BHL_6	HB	BHL
101	bHLH_3	bHLH	bHLH	138	PHD_ZNFG_187	ZincFinger	PHD
102	LFY_1	LFY	LFY	139	NAC_218	NAC	NAC
103	PHD_ZNFG_230	ZincFinger	PHD	140	MYB_related_133	MYB	MYB_related
104	bHLH_192	bHLH	bHLH	141	GARP_G2_like_60	GARP	GARP_G2like
105	BHL_10	HB	BHL	142	bZIP_85	bZIP	bZIP
106	C2C2_CO_like_102	ZincFinger	C2C2_CO_like	143	NAC_155	NAC	NAC
107	PHD_ZNFG_216	ZincFinger	PHD	144	FHA_24	FHA	FHA
108	REM_30	B32	REM	145	C2C2_DOF_36	ZincFinger	C2C2_DOF
109	Trihelix_20	Trihelix	Trihelix	146	NAC_74	NAC	NAC
110	MYB_related_115	MYB	MYB_related	147	bZIP_43	bZIP	bZIP
111	HMG_78	HMG	HMG	148	bZIP_162	bZIP	bZIP
112	PHD_ZNFG_83	ZincFinger	PHD	149	CCAAT_HAP3_22	CCAAT	HAP3
113	GeBP_8	GeBP	GeBP				

CONCLUSION PUBLICATION 3

La puce à ADN NimbleGen blé contient plus 40 642 gènes, dont 2 891 FT, ce qui en fait l'outil le mieux adapté pour observer l'expression des FT du blé. Nous retirons de nombreux enseignements de cette étude centrée sur leur expression, tout d'abord, nous remarquons, que malgré la caractéristique des FT d'être capables d'agir à des niveaux d'expression très faibles, leur niveau d'expression général n'est pas différent de celui des gènes du métabolisme ou de la signalisation.

Cette partie de l'étude a également mis en évidence des similarités entre les transcriptomes du blé et d'*A. thaliana*, espèce dicotylédone de référence. Tout d'abord concernant l'expression globale des FT par rapport à l'ensemble des gènes exprimés. Quel que soit l'organe, le stade de développement ou les conditions expérimentales, les FT représentent toujours la même proportion des gènes exprimés. Le proportion de gènes exprimés dans chaque organe est équivalent, néanmoins il est très variable en fonction du stade de développement considéré. Une hypothèse est qu'il pourrait s'agir de phases physiologiques plus « calmes » au niveau du développement de l'organe.

L'analyse comparative des transcriptomes des différents organes montre aussi que près d'un quart des gènes du métabolisme et de signalisation sont exprimés dans tous les organes. Concernant les FT, une telle aspécificité n'est observée que pour 16% d'entre eux. Nous avons également remarqué que plus de 8% des FT sont exprimés spécifiquement dans un organe, ce qui est supérieur aux autres gènes et montre bien qu'ils doivent intervenir dans des processus particuliers. Nous avons également mis en évidence la signature transcriptionnelle des organes de la plante, qui avait déjà été identifiée dans plusieurs études chez les dicotylédones, mais nous précisons qu'elle est d'abord portée par les FT.

Cette analyse comparative permet aussi de mettre en exergue 70 FT spécifiquement exprimés dans le grain et qui pourraient jouer des rôles clés dans son développement. De plus, similairement à ce qui a été observé chez l'orge, les FT spécifiquement exprimés au cours de la phase de remplissage du grain appartiennent aux familles des Zinc Finger, MYB, bZIP et NAM. Cette similarité entre nos résultats et des études menées chez l'orge étayent l'hypothèse d'une conservation des mécanismes de régulation impliqués dans le développement du grain et selon laquelle le réseau mis en évidence par Rubio-Somoza *et al.* (2006a et 2006b) chez l'orge pourrait être en partie conservé chez le blé.

De même que pour l'identité transcriptionnelle des organes, nous remarquons que la reprogrammation transcriptomique observée aux transitions entre les grandes phases de

développement du grain (cellularisation, accumulation des réserves, maturation/dessiccation) est également portée par les FT.

Dans cette étude, nous avons utilisé l'approche d'inférence de réseaux développée comme une approche de clustering dirigée, elle nous a permis de mettre en exergue 6 FT qui semblent impliqués dans la réponse au niveaux d'N et S. Parmi ces 6 FT, bZIP33, semble être un bon candidat pour de futures recherches. Son orthologue chez *A. thaliana* est impliqué dans la régulation du métabolisme de l'asparagine et nous pouvons supposer qu'il est impliqué dans le réajustement de la synthèse des acides aminés en fonction de la disponibilité en N et S. La suite de l'étude (***publication 4***) permet notamment d'étudier les réseaux globaux impliquant les FT, les gènes des PR et les acides aminés.

RÉFERENCES

Rubio-Somoza, I., Martinez, M., Diaz, I., & Carbonero, P. (2006a). HvMCB1, a R1MYB transcription factor from barley with antagonistic regulatory functions during seed development and germination. *The Plant Journal*, *45*(1), 17–30. doi:10.1111/j.1365-313X.2005.02596.x

Rubio-Somoza, I., Martinez, M., Abraham, Z., Diaz, I., & Carbonero, P. (2006b). Ternary complex formation between HvMYBS3 and other factors involved in transcriptional control in barley seeds. *The Plant Journal*, *47*(2), 269–281. doi:10.1111/j.1365-313X.2006.02777.x

PUBLICATION 4

Nitrogen to sulfur supply ratio modulates the composition and accumulation kinetics of grain storage protein in association with metabolism rebalancing and transcriptome reprogramming for wheat

Zhanwu Dai[1,2,†], Anne Plessis[1,2,††], Jonathan Vincent[1,2,3], Nathalie Duchateau[1,2], Alicia Besson[1,2,¥], Mireille Dardevet[1,2], Duyen Prodhomme[4], Yves Gibon[4], Ghislaine Hilbert[5], Marie Pailloux[3], Catherine Ravel[1,2] and Pierre Martre[1,2]

[1]INRA, UMR1095 Génétique, Diversité et Ecophysiologie des Céréales, 5 chemin de Beaulieu, F-63 039 Clermont-Ferrand, France; [2]Université Blaise Pascal, UMR1095 Diversité et Ecophysiologie des Céréales, F-63 177 Aubière, France; [3]Université Blaise Pascal, UMR6158 CNRS LIMOS Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, F-63 173 Aubière, France; [4]INRA, UMR1332 Biologie du Fruit et Pathologie, F-33 882 Villenave d'Ornon, France; [5]INRA, UMR1287 Ecophysiologie et Génomique Fonctionnelle de la Vigne, Institut des Sciences de la Vigne et du Vin, F-33 882 Villenave d'Ornon, France;[†]Current address: INRA, UMR1287 Ecophysiologie et Génomique Fonctionnelle de la Vigne, Institut des Sciences de la Vigne et du Vin, F-33 882 Villenave d'Ornon, France ; [††]Current address: Department of Biology, New York University, New York, NY 10003. [¥]Current address: Neuroscience Research Center, Université Claude Bernard Lyon 1, F-69 622, Lyon, France.

ABSTRACT

Grain storage proteins (GSPs) are an essential source of nitrogen (N) and sulfur (S) during germination. In wheat (*Triticum aestivum* L.), GSP concentration and composition are major determinants of flour end-use value. Molecular mechanisms underlying the developmental and nutritional regulation of GSP accumulation in cereals are as yet poorly understood. Here, we conducted an integrative temporal analysis of GSP, metabolite pools, and transcriptome of wheat grains in response to N and S supply. We found that the N:S supply ratio modulates the duration of accumulation of S-rich GSPs and the rate of accumulation of S-poor GSPs. These changes can be interpreted as the result of distinct relationships of N and S allocation, depending on GSP S content. We show that a major part of both developmental and nutritional modifications in GSP synthesis can be explained by the transcriptional regulation of their structural genes. Changes in the expression of transport and metabolism genes regulate the concentrations of different free amino acids under various plant nutritional status and these amino acids seem to play essential roles in the regulation of GSP expression. These results provide comprehensive insights into the regulation of wheat GSP accumulation in response to N and S supply and contribute to improve N and S use efficiency by developing new cultivars with balanced GSP composition under low S supply conditions or high N:S supply ratio.

INTRODUCTION

Grain storage proteins (GSPs) play a central role in successful plant propagation by providing nitrogen (N) and sulfur (S) resources during seed germination. In wheat (*Triticum aestivum* L.), GSPs are also essential determinants shaping grain end-use value. Wheat grains contain two predominant types of GSPs, also known as prolamins, the monomeric gliadins and the polymeric glutenins, which account for 60% to 80% of total proteins in mature grains (Shewry and Halford, 2002). The gliadins have been further classified into ω1,2-, ω5-, α/β-, and γ-gliadins, and the glutenins into high-molecular-weight (HMW-GS) and low-molecular-weight (LMW-GS) subunits (Wieser, 2007). These proteins differ by their levels of S-containing amino acids (Cys and Met), dividing GSPs into S-poor (HMW-GS and ω1,2- and ω5-gliadins) and S-rich (LMW-GS, α/β-, and γ-gliadins) proteins (Shewry et al., 1997). HMW-GS and some LMW-GS and α/β- and γ-gliadins have free –SH groups allowing them to form inter-molecular disulfide bonds resulting in the formation of very large gluten macropolymers during the late stage of grain development (Carceller and Aussenac, 1999). The S concentration and N:S ratio of grains are important indicator of the size distribution of the gluten macropolymers that govern dough rheological properties (Randall and Wrigley, 1986; Zhao et al., 1999b; Zörb et al., 2009).

Nowadays, wheat industry has to face the increased incidence of S deficiency (Zhao et al., 1999a), caused by the depletion of S in soil due to the reduction of industrial SO_2 emission and the current practice of applying large amounts of N fertilizers without concomitant S supply (Eriksen and Donald, 2009). This raises the need to improve our understanding of the dynamic regulation of plant N and S metabolism in order to develop novel fertilization strategies that could take into account the concurrent management of N and S supplies to optimize fertilizer use efficiency. This is particularly crucial to maintain wheat grain quality, which relies on GSP quantity and composition, in conditions where S is limiting.

The concentration and composition of GSPs in mature grains of cereals are strongly affected by N and S supply. N supply largely determines the concentration of GSPs and generally, a higher N supply leads to a higher GSP concentration (see review of Shewry et al., 2001; Triboi et al., 2003; Chope et al., 2014). On the other hand, S-deficiency decreases the concentration of S-rich proteins in mature grains, while it increases of the concentration of S-poor proteins, maintaining a steady level of total GSPs under S-limitation (Higashi et al., 2006; Zörb et al., 2010). Moreover, a too high N supply at a given S supply induces grain S deficiency responses, with similar influences on GSP composition (Wieser and Seilmeier, 1998). Most of these

results have been obtained on mature grains, and, therefore, little is known about the impact of S shortage on the time-course (i.e. rate and duration) of the accumulation of the different GSPs that lead to the changes in GSP composition in mature grains. However, it is obvious that the rate and duration of GSP accumulation is influenced by high temperature, water deficit, and N supply (Triboi et al., 2003; Altenbach, 2012; Hurkman et al., 2013). Thus a temporal analysis along grain development is essential to gain more insights into the dynamic adaptations to nutrient availability that underlie and define the final grain composition.

Despite the importance of an optimal N:S ratio for bread-making quality, there is a relative lack of information on the biological mechanisms underlying the interplay between N and S metabolism and intake and its influence on GSP synthesis in cereal grains. Cereal GSPs are specifically synthesized and accumulated in the endosperm and their biosynthesis is under both developmental and nutritional regulation. In wheat, transcript levels of ω- and α/β-gliadins and HMW-GS can be increased by post-anthesis N supply resulting in increased concentration of the corresponding proteins, while those of LMW-GS and γ-gliadins showed little change (Dupont et al., 2006; Altenbach and Kothari, 2007; Wan et al., 2014). However, a novel family of γ-gliadins was recently identified as highly regulated by N with dramatically repressed gene expression under N deficiency (Wan et al., 2013). N probably exerts its influence on GSP genes expression, at least in part, through transcriptional regulation mediated by conserved cis-element in GSP promoters (Müller and Knudsen, 1993).

Very little is known about the S regulation of GSP gene expression. Early studies demonstrated that S-deficiency can increase the level of S-poor protein, the β-subunit of soybean β-conglycinin, through transcriptional activation of its gene expression (Hirai et al., 1995). Oppositely, transcriptomic analysis showed that the expression of *Arabidopsis thaliana* S-rich At2S3 and At12S3 GSP genes was not decreased under S-deficient conditions, whereas the accumulation of the encoded proteins was decreased (Higashi et al., 2006), therefore suggesting a translational and/or post-translational regulation mechanism. Surprisingly, there is almost no information available about the effect of S supply on GSP gene expression in wheat grain. A few recent studies have highlighted that S deficiency can induce the expression of genes related to S transport and assimilation in wheat (Buchner et al., 2010; Ciaffi et al., 2013), as occurs in many other plant species (reviewed by Lewandowska and Sirko, 2008). In addition to these S metabolism-related genes, several S deficiency-induced genes, often with yet unknown functions, have been identified through transcriptomic analysis in *A. thaliana* (Maruyama-Nakashita et al., 2003) and tobacco (Nicotiana tabacum; Wawrzyńska et al., 2005). Detailed analysis of these S responsive genes enabled researchers to identify two *cis*-

elements (Maruyama-Nakashita et al., 2005; Wawrzyńska et al., 2010), one *trans*-factor (Maruyama-Nakashita et al., 2006) and one microRNA (Kawashima et al., 2009; Matthewman et al., 2012) involved in S mediated regulation of gene expression. To date, in wheat only one S deficiency-induced gene (*sdi1*) has been identified in vegetative organs of seedlings (Howarth et al., 2009) and no trans-acting factor has been reported. Therefore, it is still not clear whether/how S deficiency modulates the metabolism and transcriptome of wheat endosperm that are responsible for maintaining a constant quantity of total GSPs under fluctuating S availability conditions.

Here, we performed an integrative analysis of the temporal response of wheat grain to N and S supply. Grains from four N and S supply conditions were sampled at various developmental stages and subjected to detailed GSP, targeted metabolites, and transcriptome analyses. We found that the accumulation of S-rich and S-poor GSPs are differentially affected by S deficiency and that the rates of accumulation of LMW-GS, HMW-GS, and ω5-, ω1,2-, and α/β-gliadins are significantly correlated with their transcripts. These modifications in GSP synthesis were associated with rebalanced grain metabolisms that caused temporal modifications in grain C (sucrose, organic acids), N (Asn) and S (GSH) pools. Moreover, specific phases of gene response were identified and coordinated modifications were observed in genes related to S transport and S metabolism, S response and storage compound synthesis.

MATERIALS AND METHODS

Plant Material and Growth Conditions

Plants of the bread wheat (*Triticum aestivum* L.) cv. Récital were transplanted just after vernalization into polyvinyl chloride (PVC) columns (1.9 L, i.d. 7.5 cm, length 50 cm) filled with a 1:1 (v:v) washed perlite:river sand mix and were arranged in a 10 m^{-2} walk-in growth chamber in a randomized complete block design with four N/S treatments (see below) and three blocks to form an homogeneous stand with a plant density of 261 plant m^{-2}. The conditions in the growth chamber were 20°C/15°C (day/night), 55%/75% air relative humidity (day/night), with an average PPFD of 550 µmol m^{-2} s^{-1} at top of the plants during the 16 h photoperiod.

Each PVC column received 68 mL of a modified Hoagland's nutrient solution (Castle and Randall, 1987) 4 times per day and 170 mL of demineralized water at the end of the dark period to eliminate any potential enrichment of nutrients in the substrate. All nutrient solutions were prepared with demineralized water. Four different combinations of N and S supply were applied with either 3 mM N and 2 mM S (control treatment NS), 3 mM N and 0.02 mM S (normal N and low S, treatment NS-), 15 mM N and 0.02 mM S (supra-optimal N treatment and low S, treatment N+S-), or replacing treatment N+S- with 2 mM S was applied from 25 days (490 °Cdays) after anthesis (i.e. midway through the effective grain filling phase) to grain ripeness maturity (termed N+S-/N+S). For the control treatment (NS) the nutrient solution contained 1 mM KH_2PO_4, 1 mM $Ca(NO_3)_2$, 0.5 mM NH_4NO_3, 2 mM $MgSO_4$, 3 mM $CaCl_2$, 5 mM KCl, 10 µM H_3BO_3, 0.7 mM $ZnCl_2$, 0.4 µM $CuCl_2$, 4.5 µM $MnCl_2$, 0.22 µM MoO_3, and 50 µM EDS-Fe. Treatment NS- was obtained from treatment NS by reducing $MgSO_4$ concentration to 0.02 mM and adding 1.98 mM $MgCl_2$. Treatment N+S- was obtained from Treatment NS- by increasing NH_4NO_3 to 1 mM and $Ca(NO_3)_2$ to 4 mM to replace $CaCl_2$ and replacing KCl by 5 mM KNO_3. The shift from N+S- to N+S was obtained by removing $MgCl_2$ and increasing $MgSO_4$ concentration to 2 mM in N+S-. To obtain a large range of N:S ratio in grains, after anthesis all N containing compounds were excluded from the NS and NS- solutions while S concentration was kept unchanged. The nutrient solution of treatment N+S- was kept unchanged until ripeness maturity.

Main-stem ears were tagged when the anthers of the central florets appeared and they were sampled every 90 degree-days (°Cdays, equivalent to ca. 11 days) from 200 °Cdays after anthesis to ripeness maturity (ca. 1150 °Cdays after anthesis). To obtain representative grain samples 8 (from 200 to 370 °Cdays after anthesis) to 4 (from 460 to 1150 °Cdays after

anthesis) main-stem ears were sampled to form a biological replicate for each treatment. Three biological replicates were used unless indicated otherwise.

Grain Processing for Transcriptional and Biochemical Analysis

Two to three basal grains of the central third of each ears sampled between 200 and 650 °Cdays were first taken, the embryo was removed with a fresh razor blade, and the remaining tissues (i.e. the outer layers and the endosperm) were immediately frozen in liquid N and then stored at -80 °C for RNA extraction. The ears with remaining grains (the whole ears after 650 °Cdays after anthesis) were frozen in liquid N and then freeze-dried. The ears where then threshed and the number and dry mass of grains were determined. Grains were ground to wholemeal flour with a MM 200 mixer mill (Retsch). The wholemeal flour was then stored at -80°C under N_2 atmosphere before biochemical analysis.

Determination of Total Grain Sulfur and Nitrogen Concentration

An aliquot of 5 mg flour was weighed in tin capsules and the total N and S concentrations were determined with the Dumas combustion method using a FlashEA 1112 NC Analyzer (Thermo Electron) following the manufacturer recommendations. Grain protein content was calculated by multiplying grain N concentration by 5.62 (Mossé et al., 1985).

Sequential Extraction and Quantification of the Non-Prolamin, Gliadin and Glutenin Protein Fractions and Separation and Quantification of the Gliadin classes and Glutenin Subunits

Non-prolamin, gliadin, and glutenin protein fractions were sequential extracted from 100 mg of wholemeal flour as described in (Plessis et al., 2013). The total N concentration of each protein fraction was determined with the Dumas combustion method as described above.

Gliadin classes and glutenin subunits were separated and quantified by HPLC using an Agilent 1290 Infinity LC system (Agilent Technologies). The gliadin extracts obtained by sequential extraction were used. Glutenins were extracted with a protocol adapted from Fu and Kovacs (1999). The gliadin and glutenin extracts were then filtered through regenerated cellulose syringe filters (0.45 µm pore diameter, UptiDisc, Interchim) and 4 µL (gliadin) or 2 µL (glutenin) or protein extract were injected onto a C8 reversed-phase Zorbax 300 StableBound column (2.1 × 100 mm, 3.5 µm, 300 Å, Agilent Technologies) maintained at 50°C. The eluents used were ultra-pure water (solvent A) and acetonitrile (solvent B) each containing 0.1% trifluoroacetic. The flow rate was 1 mL min^{-1}. Proteins were separated by using a linear gradient from 24% to 50% B in 13 min for gliadin and from 23% to 42% B in 25 min for glutenin. Proteins were detected by UV absorbance at 214 nm. After the gradient, the column was

washed with 80% B for 2 min and then equilibrated at 24% (gliadin) or 23% (glutenin) B for 2 min, at the same flow rate.

Chromatograms were processed with ChemStation 10.1 software (Agilent Technologies). The signal obtained from a blank injection was subtracted from the chromatograms before integrating the data. The HPLC peaks corresponding to each of the four gliadin classes were identified following Wieser et al. (1998). The quantity of each gliadin class or glutenin subunit as a percentage of total gliadin or glutenin was calculated by dividing the areas under each HPLC peak by the total area under the chromatogram trace. The quantity of each gliadin or glutenin class per grain was calculated by multiplying the proportion of each gliadin class in total gliadin or glutenin subunit in total glutenin by the total quantity of gliadin or glutenin per grain as quantified by Dumas analysis.

Free Amino Acid Separation and Quantification

Free amino acids were extracted from 20 mg of whole meal flour and quantified by HPLC using a Alliance 2695 system (Waters) equipped with a Nova-Pack C18 AccQ-Tag column (WAT052885, Waters) and a fluorescence scanning detector (Waters 474) after derivatization with 6-aminoquinolyl-N-hydroxysuccinimidyl carbamate as described in (Cohen and Michaud, 1993).

Metabolite Assays

Metabolites were extracted from about 20 mg of wholemeal flour. Soluble sugars (glucose, fructose and sucrose) were assayed enzymatically according to (Jelitto et al., 1992), malate was assayed according to (Nunes-Nesi et al., 2007) and citrate according to (Tompkins and Toffaletti, 1982). Starch was measure as in (Hendriks et al., 2003). Oxidized glutathione (GSSG) and reduced glutathione (GSH) were extracted and assayed enzymatically with glutathione reductase and 5,5'-dithiobis-2-nitrobenzoic acid as described in (Griffith, 1980). Assays were prepared in 96-well microplates and absorbance was read in an MP96 microplate reader (SAFAS, Monaco). Results were visualized in metabolic pathways using MapMan 3.5 software (Thimm et al., 2004).

RNA Extraction and Gene Expression Analysis

Grains from two biological replicates were ground in liquid N_2 and total RNA was extracted as described in (Romeuf et al., 2010) and then treated with Ambion® DNaseI (Life Technologies). cDNA was synthesized from total RNA using the SuperScript Double-stranded cDNA Synthesis Kit (Invitrogen, Life Technologies) and labeled with fluorescent Cy3 and Cy5 dyes using a NimbleGen® two-color DNA labeling kit (Roche Diagnostics).

Grain transcriptome was analyzed using a custom 40k NimbleGen® (Roche Diagnostics) wheat microarray (ArrayExpress ref. A-MEXP-1928) comprising 39,179 sequences from the *T. aestivum* National Center for Biotechnology Information NCBI UniGene (build no. 55) and 1,613 transcription factors from the wDBTF database (Romeuf et al., 2010). Hybridization was performed using NimbleGen® hybridization kit, according to the manufacturer's instructions. Samples from the two biological replicates were labeled with the two different dyes and were randomly hybridized. Microarray images of each slide were obtained with an InnoScan900 AL microarray scanner (Innopsys) and read with Mapix program (Innopsys).

The expression of 11 GSP and 8 regulatory genes were further quantified by real-time qPCR with an ABI PRISM 7900HT sequence detection system using Power SYBR Green PCR Master Mix (Applied Biosystems). The specificity and efficiency of the primer pairs were verified as described in (Ravel et al., 2009). The genes coding for glyceraldehyde 3-phosphate dehydrogenase and elongation factor 1α were used as internal controls and later used for normalizing the expression (Pfaffl et al., 2004). Primer pairs amplifying a cyclophilin gene and spanning one intron were used to check for the absence of DNA contamination.

Statistical Analysis

Microarray

Microarray data were analyzed using R (R Development Core Team, 2010) and Bioconductor tools (Gentleman et al., 2004). After quality control, expression intensities were normalized with the RMA function from the oligo package (Carvalho and Irizarry, 2010). Differentially expressed genes among the four N and S treatments were identified using moderate *t*-test in the LIMMA package (Smyth, 2005). Genes with more than 2-folds change in expression and an adjusted *P*-value < 0.05 (Benjamini and Hochberg, 1995) after false discovery rate correction were considered significantly differentially expressed. The whole dataset has been submitted in MIAME-compliant form to the ArrayExpress database (Accession no. E-MTAB-1782). Functional annotation of the transcripts spotted on the microarray was conducted using the dbWFA database (Vincent et al., 2013) using threshold coverage and identity of 50 and 45%, respectively. The probe set and functional annotation of the significant genes and their clustering are given in Supplemental Table 4.

Correlation between the Accumulation Rate of Grain Storage Proteins and the Expression of their Structural genes

To determine the rate and duration of accumulation of each GSP fraction, data were fitted with a 3-parameter logistic function (Triboi et al., 2003). Standardized major axis linear

regressions were then used to quantify the relationship between the accumulation rate of each GSP fraction and the expression of their structural genes and to compare the slopes of the linear regressions among N and S treatments (Warton et al., 2012).

Network Analysis

Regulatory networks among transcripts, metabolites and proteins were inferred with an algorithm describing rule semantics between attributes (Agier et al., 2007) using the RulNet platform (http://rulnet.isima.fr). This algorithm provides a unifying framework for the discovery of rules within–omics data and allows defining different semantics (different biological meanings) for the relationship between a given pair of variables, therefore forming valuable complementary and novel knowledge in comparison with classical unsupervised techniques (such as correlation based network analysis). Data were first scaled and the semantics intended to generate rules between variables according to their relative levels among different treatments. Out of all the rules that were discovered a filter set was applied using the following thresholds: support > 0.3, confidence > 0.9, and lift > 1.5. Functional modules were identified within the constructed network with the Cytoscape 2.8.2 software (Smoot et al., 2011) using the Clust&see plug-in (Spinelli et al., 2013).

RESULTS

Nitrogen and Sulfur Supply Causes Profound Modifications in Mature Grains

To elucidate the developmental, metabolic, and transcriptional components of GSP synthesis regulation in response to S and N supply, winter wheat (*Triticum aestivum cv.* Récital) plants were grown with four treatments combining high and low rates of N and S supply: either 3 mM N and 2 mM S (control treatment, termed NS), 3 mM N and 0.02 mM S (normal N and low S, treatment NS-), 15 mM N and 0.02 mM S (supra-optimal N and low S, treatment N+S-), or by transferring plants from treatment N+S- to a nutrient solution containing 15 mM N and 2 mM S (high S and supra-optimal N, treatment N+S-/N+S) midway through the effective grain filling period (i.e. at 490 °Cdays after anthesis).

The treatments led to a large range of grain total S concentration (from 0.11 for NS- to 0.16 g S 100 g^{-1} DM for N+S-/N+S) and N:S ratio (14.5 for NS to 24.5 for N+S-) in mature grains (Figures 1B and 1F, Supplemental Table 1). S deficiency affects flour commercial value and dough functional properties when grain S concentration is lower than 0.12% or the N:S ratio is higher than 17 (Randall et al., 1981). Based on these criteria, N+S- grains suffered from severe S deficiency and NS- grains experienced marginal S deficiency. Grain total protein concentration was mainly modified by N supply (from 11.5 g protein 100 g^{-1} DM for NS and NS- to 15.7 g protein 100 g^{-1} DM for N+S- and N+S-/N+S) with less influence of S supply (Figure 1D, Supplemental Table 1).

Multiple HPLC peaks were detected within each class of GSP, including 5 peaks for HMW-GS, 16 LMW-GS, 5 ω5-gliadins, 5 ω1,2-gliadins, 19 α/β-gliadins, and 6 γ-gliadins (Supplemental Figures 1 and 2). In mature grains, the percentage of each individual peak within a given class of GSP was not significantly modified by the treatments (Supplemental Figures 1 and 2), suggesting that proteins of the same class are under coordinated regulation in response to nutrient availability. Common accumulation profiles for GSPs of the same class have also been observed under various temperature and fertilization scenarios (Hurkman et al., 2013). Because of this co-regulation, we analyzed the kinetics of the total amount of protein of each class (Figures 2 and 3).

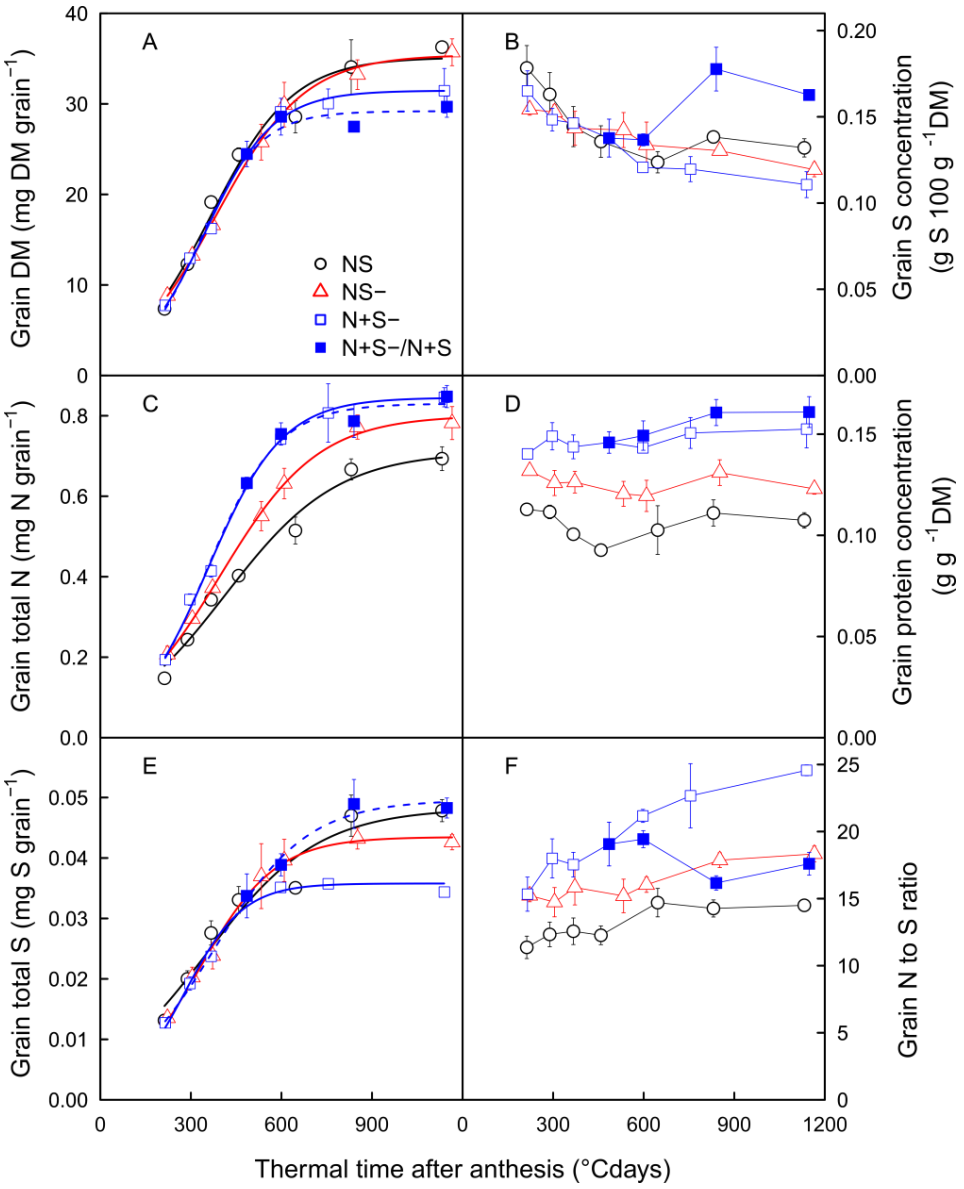


Fig. 1. ***Grain dry mass (DM), total nitrogen (N) and sulfur (S) mass and concentration, and N:S ratio versus thermal time after anthesis for wheat grains grown with different rates of N and S supply.*** ***(A)*** *Grain dry mass.* ***(B)*** *Grain total S concentration.* ***(C)*** *Grain total N mass per grain.* ***(D)*** *Grain total protein concentration .* ***(E)*** *Grain total S mass per grain.* ***(F)*** *Grain N:S ratio. The four treatments were NS (3 mM N and 2 mM S), NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and N+S-/N+S (shift at mid-grain filling from N+S- to 15 mM N and 2 mM S). Data are means ± 1 SE for n = 3 biological replicates. Lines in* ***(A)****,* ***(C)*** *and* ***(E)*** *were fitted to the data using the parameters of a logistic function reported in Supplemental Table 2.*

S supply finely tuned GSP composition at maturity, with S-poor proteins (HMW-GS and ω5- and ω1,2-gliadins) favored under S deficiency at the expense of S-rich proteins (LMW-GS and γ-gliadins; Figures 2 and 3, Supplemental Table 1). At maturity, the proportions of HMW-GS

and ω5- and ω1,2-gliadins were all significantly increased by severe S deficiency (N+S-) in comparison with NS, while those of LMW-GS and γ-gliadins were decreased (Supplemental Table 1), in agreement with previous reports (MacRitchie and Gupta, 1993; Zhao et al., 1999b; Zörb et al., 2009). Initially classified as S-rich proteins (Shewry et al., 1997), α/β-gliadins exhibited a response to S more similar to S-poor proteins (Figure 3C), as previously found by (Wieser et al., 2004). Therefore, hereafter α/β-gliadins are classified as S-poor proteins.

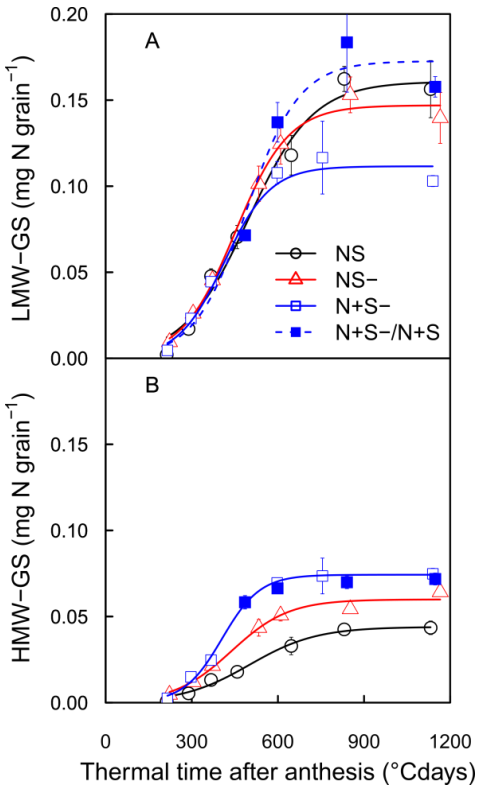


Fig. 2. Nitrogen (N) and sulfur (S) have distinct effects on the kinetics of the low- (LMW-GS) and high-molecular weight (HMW-GS) glutenin subunits. (A) *LMW-GS mass per grain.* ***(B)*** *HMW-GS mass per grain. The four treatments were NS (3 mM N and 2 mM S), NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and N+S-/N+S (shift at mid-grain filling from N+S- to 15 mM N and 2 mM S). Data are means ± 1 SE for n = 3 biological replicates. Lines were fitted to the data using the parameters of a logistic function reported in Supplemental Table 2.*

The Effect of Nitrogen and Sulfur Supply on Grain Protein Composition Can Be Broken Down into Modifications of Rate and Duration of Accumulation

The total quantity of a storage compound at maturity is the product of the rate and duration of its accumulation. To quantify the effects of N and S supply on these two parameters, data were fitted to a logistic function equation (Supplemental Table 2). Severe S-deficiency (N+S-) decreased grain dry mass at maturity because of a shorter duration of grain-filling (Figure 1A, Supplemental Table 2). A similar response to S-deficiency was observed for grain total S mass (Figure 1E): it was 28% lower for N+S- than for NS (Supplemental Table 1) as a result of a duration of accumulation 340 °Cdays (ca. 17 days) shorter for N+S- than for NS (Supplemental Table 2). S supply had less influence on grain total N, which was largely determined by N

supply (Figure 1C), modifying both its rate and duration of accumulation (Supplemental Table 2).

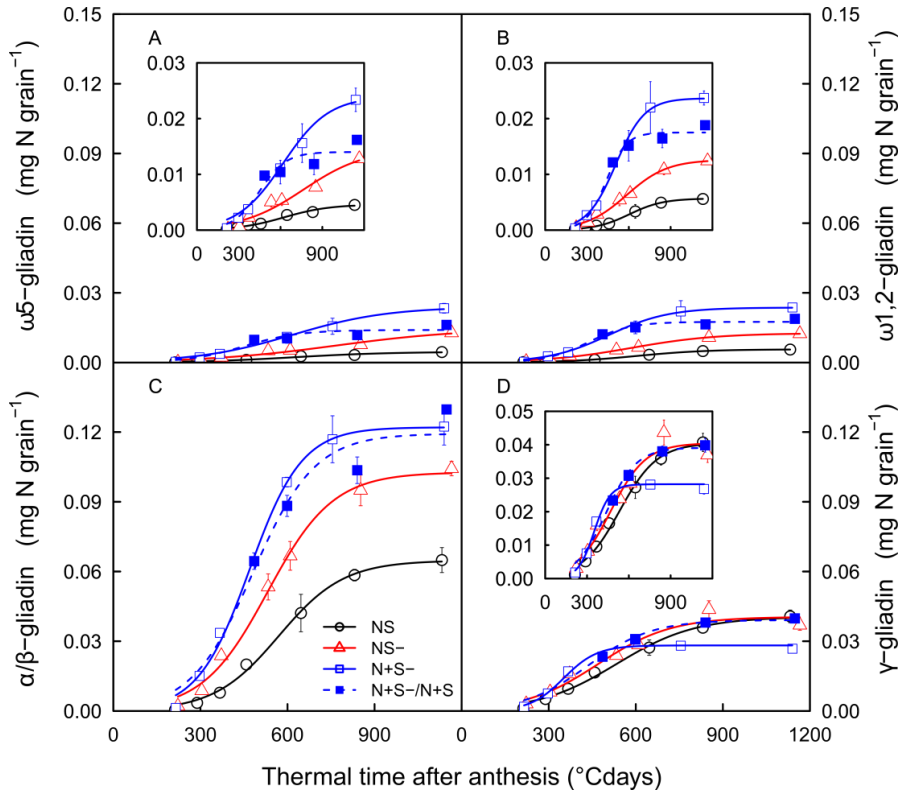


Fig. 3. Nitrogen (N) and sulfur (S) have distinct effects on the kinetics of the four classes of gliadins. ***(A)*** *ω5-gliadin mass per grain.* ***(B)*** *ω1,2-gliadin mass per grain.* ***(C)*** *α/β-gliadin mass per grain.****(D)*** *γ-gliadin mass per grain. Plants were grown with one of the following four combinations of N and S supply: NS (3 mM N and 2 mM S), NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and N+S-/N+S (shift at mid-grain filling from N+S- to 15 mM N and 2 mM S). Data are means ± 1 SE for n = 3 biological replicates. Lines were fitted to the data using the parameters of a logistic function reported in Supplemental Table 2. The same y-axis scale was used to facilitate the comparison among different gliadin fractions. The insets show the data with expanded y axis.*

Kinetic analysis of GSP uncovered different responses of S-rich and S-poor GSPs to N and S supply (Figures 2 and 3). On the one hand, the rates of accumulation of S-rich GSPs changed only to a limited extent between the four treatments, the differences in their mass per grain at maturity being mainly due to changes in the duration of accumulation (Figures 2A and 3D). For instance, the duration of accumulation of LMW-GS was 170 °Cdays shorter for N+S- than for NS (Supplemental Table 2). On the other hand, both the duration and rate of accumulation of S-poor GSPs were modified by N and S treatments (Supplemental Table 2). The most significant change was however the increase in the rate of accumulation under high N:S

supply ratio (N+S-), which partly compensated for the shorter duration, always resulting in higher mass of S-poor GSP per grain at maturity (Figure 2B and 3A-C).

Supplying S midway through grain filling (treatment N+S-/N+S) increased the mass per grain and the concentration of total S within 113 ˚Cdays (ca. 6 days), while grain DM and total N mass did not respond (Figure 1). The rapid increase in grain total S mass after the nutritional shift rebalanced grain total N and S mass, ultimately relieving S deficiency, as showed by a N:S ratio of 17.6 and the highest S concentration (0.16 g S g S 100 g^{-1} DM) at maturity (Supplemental Table 1). After the nutritional shift, the mass of S-rich GSPs increased gradually and at maturity it was similar to NS and NS- (Figures 2A and 3D). In contrast, the accumulation of S-free proteins (ω5- and ω1,2-gliadins) stopped after the nutritional shift (Figures 3A and 3B), resulting in lower mass per grain than for N+S- at maturity. The kinetics of S-poor GSP (HMW-GS and α/β-gliadins) was not affected by the supply of S at mid-grain filling (Figures 2B and 3C).

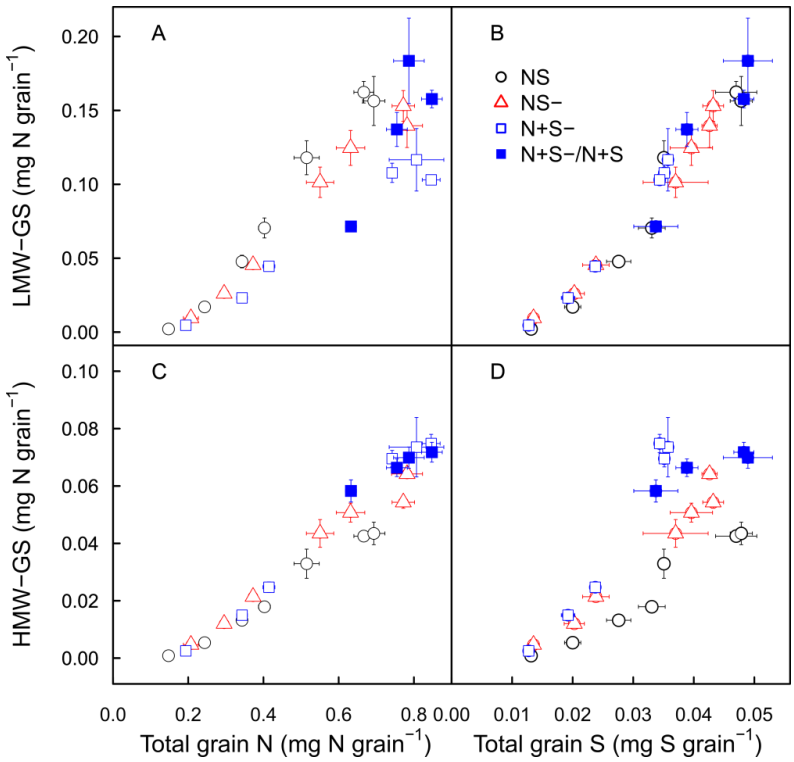


Fig. 4. The mass of low-molecular weight glutenin subunits (LMW-GS) per grain scaled with the mass of total sulfur (S) per grain, while the mass of high-molecular weight glutenin subunits (HMW-GS) scaled with the mass of total nitrogen (N) per grain. (A) *Mass of LMW-GS per grain versus mass of total N per grain.* ***(B)*** *Mass of LMW-GS per grain versus mass of total S per grain.* ***(C)*** *Mass of HMW-GS per grain versus mass of total N per grain.* ***(D)*** *Mass of HMW-GS per grain versus mass of total S per grain. Data are means ± 1 SE for n = 3 biological replicates.*

Sulfur-rich and Sulfur-poor Grain Storage Proteins Follow Distinct Allocation Relationships

We assessed how the N and S accumulated in the grain allocate to different groups of GSP by quantifying the correlation between the mass per grain of total N or S and GSP (Figures 4 and 5). We found unique correlations between the mass of total grain N and of these protein groups regardless of N and S supply treatments (Figures 4C, 5A 5C, and 5E). In contrast, the proportion of N allocated to S-rich GSPs changed with N and S treatments (Figures 4A and 5G). For instance, a higher proportion of N was allocated to LMW-GS for NS than for N+S- (Figures 4A). However, S allocation to S-rich proteins was independent of S and N supply (Figures 4B and 5H), even if the accumulation duration of these groups of GSPs was largely modified by the treatments (Figures 2A and 3D).

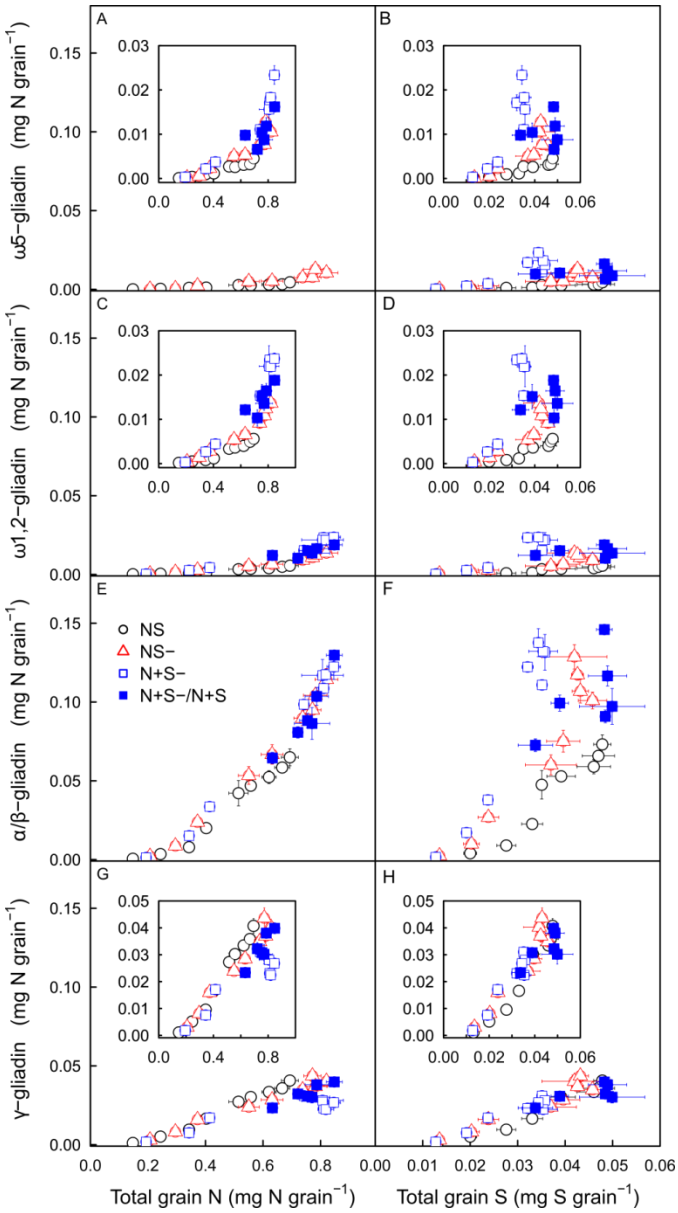


Fig. 5. Allocation of grain N is independent of the treatment for ω5-, ω1,2- and α/β-gliadins while allocation of grain S is independent of treatment for γ-gliadins. ***(A)*** *Mass of ω5-gliadins per grain versus mass of total N per grain.* ***(B)*** *Mass of ω5-gliadins per grain versus mass of total S per grain.* ***(C)*** *Mass of ω1,2-gliadins per grain versus mass of total N per grain.* ***(D)*** *Mass of ω1,2-gliadins per grain versus mass of total S per grain.* ***(E)*** *Mass of α/β -gliadins per grain versus mass of total N per grain.* ***(F)*** *Mass of α/β -gliadins per grain versus mass of total S per grain.* ***(G)*** *Mass of γ -gliadins per grain versus mass of total N per grain.* ***(H)*** *Mass of γ -gliadins per grain versus mass of total S per grain The same y-axis scale was used to facilitate the comparison among different gliadin fractions. The insets show the data with expanded y axis.*

S-poor Grain Storage Proteins are the Most Transcriptionally Regulated Grain Storage Proteins

The kinetic analysis of GSP showed that N and S supply had a major effect on the rate of accumulation of S-poor GSPs. To evaluate the importance of transcriptional control on this aspect of GSP synthesis regulation, we analyzed the relationship between the accumulation rate of a given protein and its gene expression (Figure 6, Supplemental Figures 3 and 4). The S supply shift treatment (N+S-/N+S) was excluded from this analysis because of its limited number of time points. We found significant linear relationships ($P < 0.002$) between gene expression and rates of protein accumulation, independently of treatments, for LMW-GS, HMW-GS, and ω5-, ω1,2-, and α/β-gliadins (Figure 6, Supplemental Table 3), suggesting a strong transcriptional regulation of these proteins. The intercepts of the regressions were not significantly different from zero and the slopes were similar for the LMW-GS, HMW-GS, and α/β-gliadins, and were all much lower than those of the S-free ω5- and ω1,2-gliadins (Supplemental Table 3).

The correlation coefficient was higher for the S-poor HMW-GS, and ω5-, ω1,2-, and α/β-gliadins (all $r^2 > 0.68$) than for the S-rich LMW-GS ($r^2 = 0.59$). For the γ-gliadins, the correlation between the rate of protein accumulation and gene expression was not significant (Figure 6F and Supplemental Table 3). These results suggest that S-poor GSPs are more transcriptionally regulated than S-rich GSPs in response to N and S supply, with the γ-gliadins at the end of this spectrum.

Nitrogen and Sulfur Availabilities Affect Free Amino Acids and their Precursors

As transcriptional regulation does not explain all the modifications in GSP accumulation in response to N and S supply, we investigated the possibility that the availability of their precursors (i.e. free amino acids) could affect GSP biosynthesis. We thus analyzed 20 free amino acids and some of the main markers of C and S metabolisms: starch, glucose, fructose, sucrose, citrate, malate, and two important S-containing tripeptides (glutathione and glutathione disulfide). The variation of each metabolite pool over time and treatments is displayed on a schematic representation of the related metabolic pathways to provide a straightforward view (Figure 7).

The most striking effect of N and S supplies on amino acid amount was observed for Asn. Its mass per grain reached a 12.7-fold increase under severe S deficiency (N+S-) when compared to NS (Figure 7 and Supplemental Figure 5). In the same conditions, the masses of Asp and Lys that belong to the Asn pathway were also increased. A similar tendency was

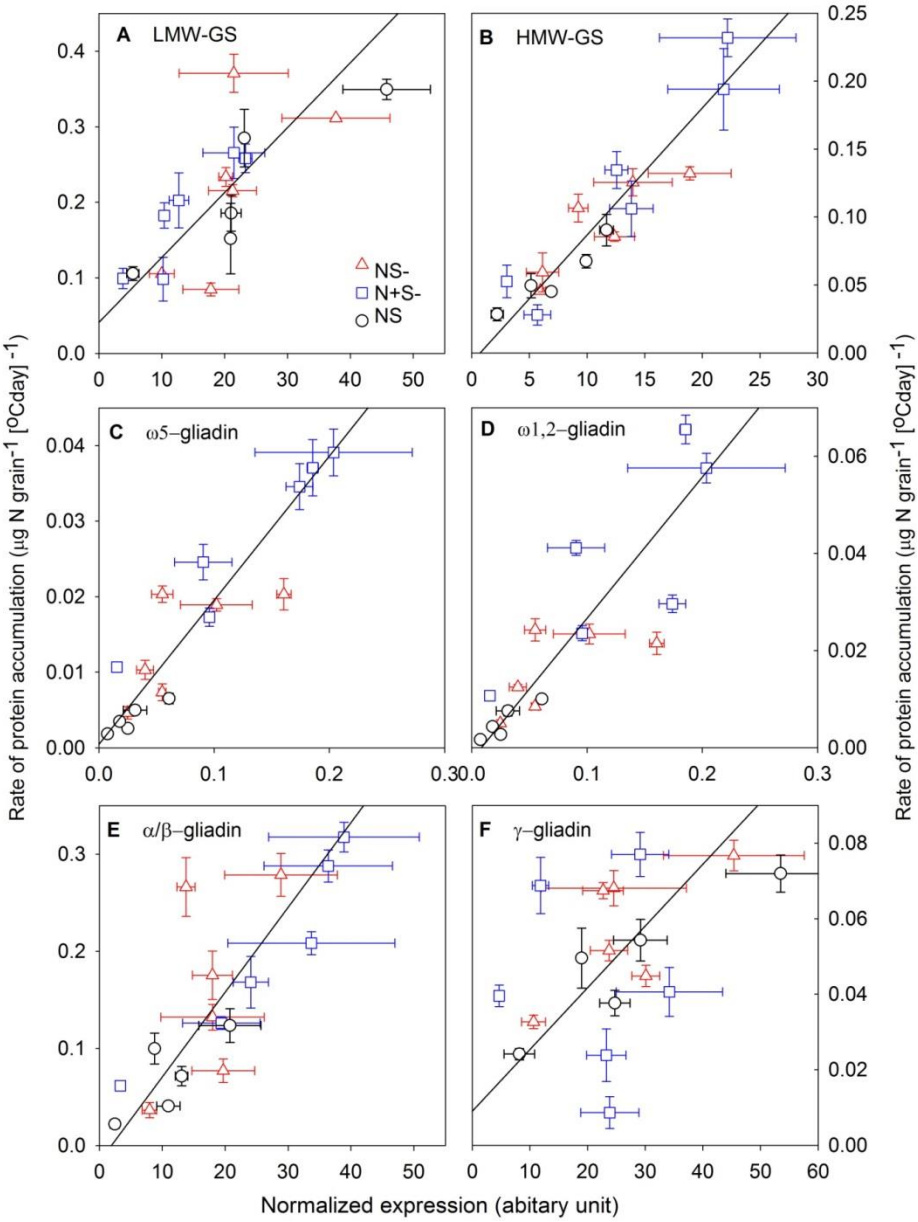


Fig. 6. The rate of accumulation of LMW-GS, HMW-GS, ω5-, ω1,2- and α/β-gliadins was correlated with the expression of their structural genes. (A) *Accumulation rate versus expression level for LMW-GS.* ***(B)*** *Accumulation rate versus expression level for HMW-GS.* ***(C)*** *Accumulation rate versus expression level for ω5-gliadins.* ***(D)*** *Accumulation rate versus expression level for ω1,2-gliadins.* ***(E)*** *Accumulation rate versus expression level for α/β-gliadins.* ***(F)*** *Accumulation rate versus expression level for γ-gliadins. Lines are standardized major axis regressions. Statistics of the regressions are given in Supplemental Table 3.*

observed for Tyr, Leu, Val, Arg, and Gln, while the remaining amino acids were hardly influenced by the treatments.Citrate and malate are organic acids necessary to the synthesis of the amino acid precursor 2-oxoglutarate via the tricarboxylic acid cycle. The mass of citrate per grain was increased from mid-grain filling to maturity while the mass of malate per grain was higher under S-deficiency only around its accumulation peak, at mid-grain filling (Figure 7 and Supplemental Figure 6). Sucrose, glutathione (GSH), and glutathione disulfide (GSSG) were all decreased by S deficiency, while starch, glucose, and fructose were not influenced by the treatments.

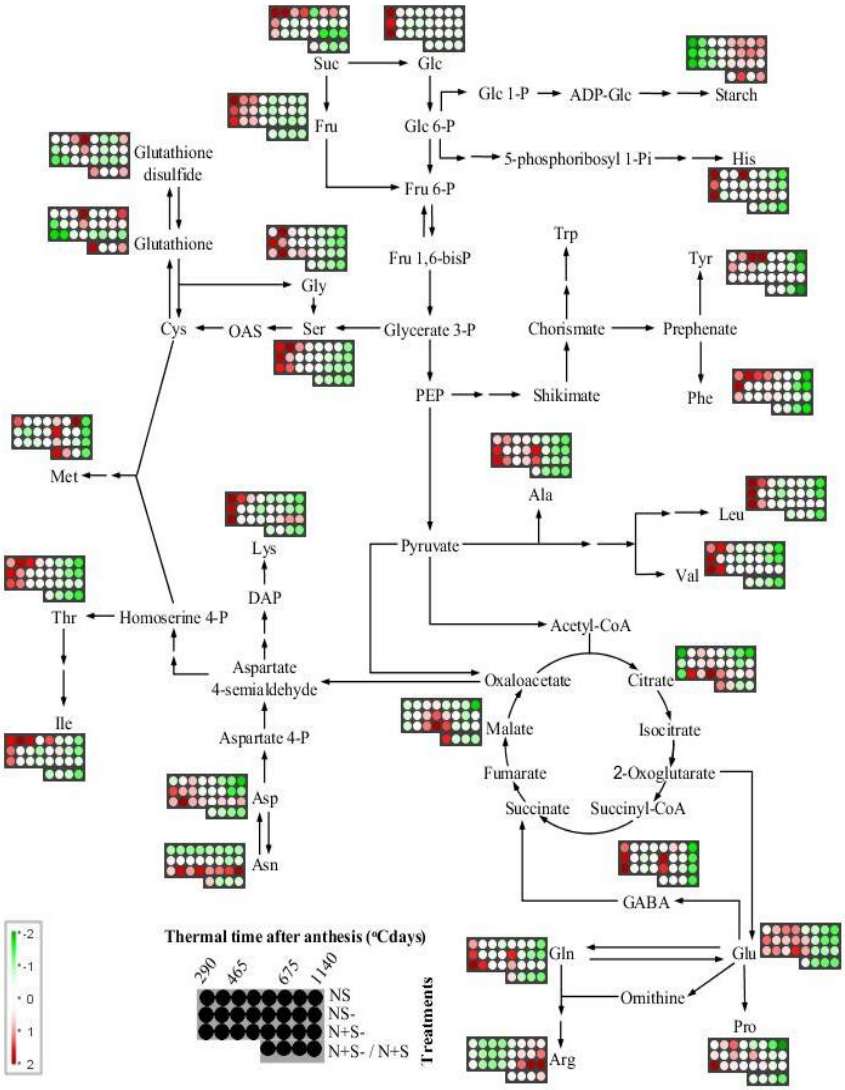


Fig. 7. Effect of nitrogen (N) and sulfur (S) supply on selected metabolite pools during grain development. *The mass per grain of sugars, organic acids and amino acids are represented on a 4 × 8 grid in their respective pathways. Each row of the grid corresponds to a treatment and each column to a thermal time after anthesis. For each metabolite, the means for n = 3 biological replicates were scaled over all 28 data points and represented on a color scale with red for higher mass per grain and green for lower mass per grain.*

The supply of S at mid-grain filling (N+S-/N+S) reversed the effect of S deficiency on metabolite pools. The mass of Asn, Asp, Lys, Arg, Tyr, Leu, Val, citrate, and malate quickly decreased, while the mass of GSH and GSSG increased to NS levels. Only for sucrose was the recovery partial, it increased to a lower level than NS (Figure 7).

Sulfur Deficiency Induces Transcriptome Reprogramming in Metabolic and Transport Genes and Regulates Non-Prolamin and Grain Storage Protein Genes

Our previous analyses suggest that the import and metabolism of N and S could be involved in regulating GSP synthesis under variable N and S supply. We therefore analyzed the transcriptome of grains from the same ears as those used for protein and metabolite analyses. The reliability of the microarray was assured by high Pearson's coefficients of correlation (from 0.51 to 0.90, Supplemental Figure 7) between the expression levels of 11 selected genes determined by microarray and qPCR.

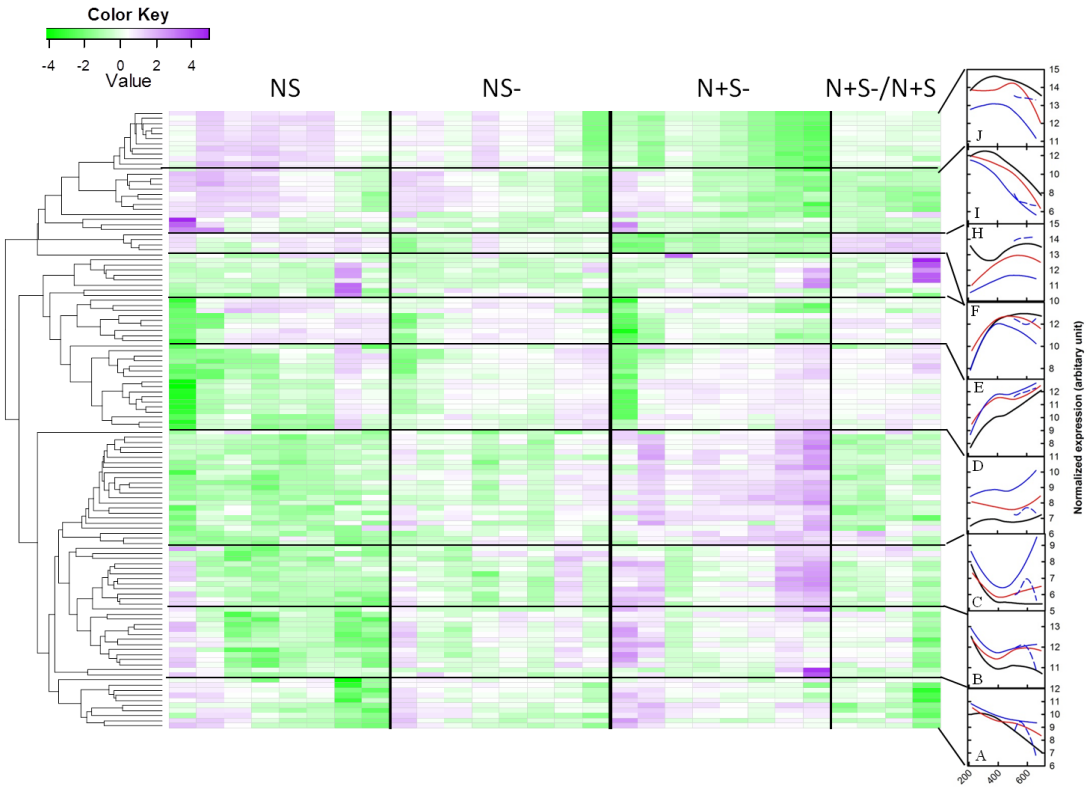


Fig. 8. Heat map of differentially expressed genes among the four nitrogen (N) and sulfur (S) supply treatments. *Data were scaled and log2-transformed and are represented in a color scale with purple for higher expression and green for lower expression. The average expression levels for each cluster is shown on the right panels, with black solid line for NS, red for NS-, blue for N+S-, and blue dashed line for N+S-/N+S.*

A total of 122 genes showed a significant change in expression between at least two of the four N and S treatments and for at least one time point during grain development (Supplemental Table 4). These genes were grouped into 10 clusters according to their temporal expression profiles and responses to N and S treatments (Figure 8). Clusters 1 to 5 contained 47 genes that were all repressed by S deficiency, while clusters 6-10 contained 75 genes that were induced by S deficiency.

Most of the genes that were significantly down-regulated by S deficiency were GSP or non-prolamin proteins. For instance, twelve putative genes of alpha-amylase/trypsin inhibitors in cluster 1 and four venin-like proteins in cluster 3 were strongly repressed by S deficiency. In contrast, four putative serpin genes and one globulin-3A and vicilin-like proteins in cluster 6 were up-regulated by S deficiency.

Among the genes that were significantly up-regulated by S deficiency, a major part was involved in the metabolism and transport of N and S. Three genes involved in S metabolism (adenosine-5-phosphosulfate kinase, cystathionine beta-lyase, and adenosine-5-phosphosulfate reductase) and several known S-deficiency induced genes (Sdi1, Sdi2, and UP9A) in other species (Wawrzyńska et al., 2005; Howarth et al., 2009) were significantly up-regulated by S shortage in our conditions (cluster 7 and 10). In clusters 7 and 9 we found three genes coding for amino acid metabolism enzymes. Genes involved in C metabolism were also represented. Three of them coded for glycolysis enzymes: a triose phosphate isomerase in cluster 10 and a phosphofructokinase and a diphosphate-fructose-6-phosphate 1-phosphotransferase in cluster 9, and a fourth one in cluster 8 encoded an ADP-glucose pyrophosphorylase (AGPase). We found that two amino acid transporters and three sulfate transporters were up-regulated by S-deficiency (cluster 7).

Our transcriptomic analysis also pointed out putative candidate genes for GSP gene expression regulation. We found two high-mobility group (HMG) transcriptional factors in cluster 8 and one APETALA2-ethylene-responsive element-binding protein (AP2-EREBP) in cluster 10, which were all up-regulated by S-deficiency.

Regulatory Network Analysis Uncovers Novel Protein-Metabolite-Gene Linkages

In addition to the above correlation analyses, we further explored our dataset using a directed multi-level regulatory network (Agier et al., 2007), screening for potential associations among transcripts, proteins, and metabolites.

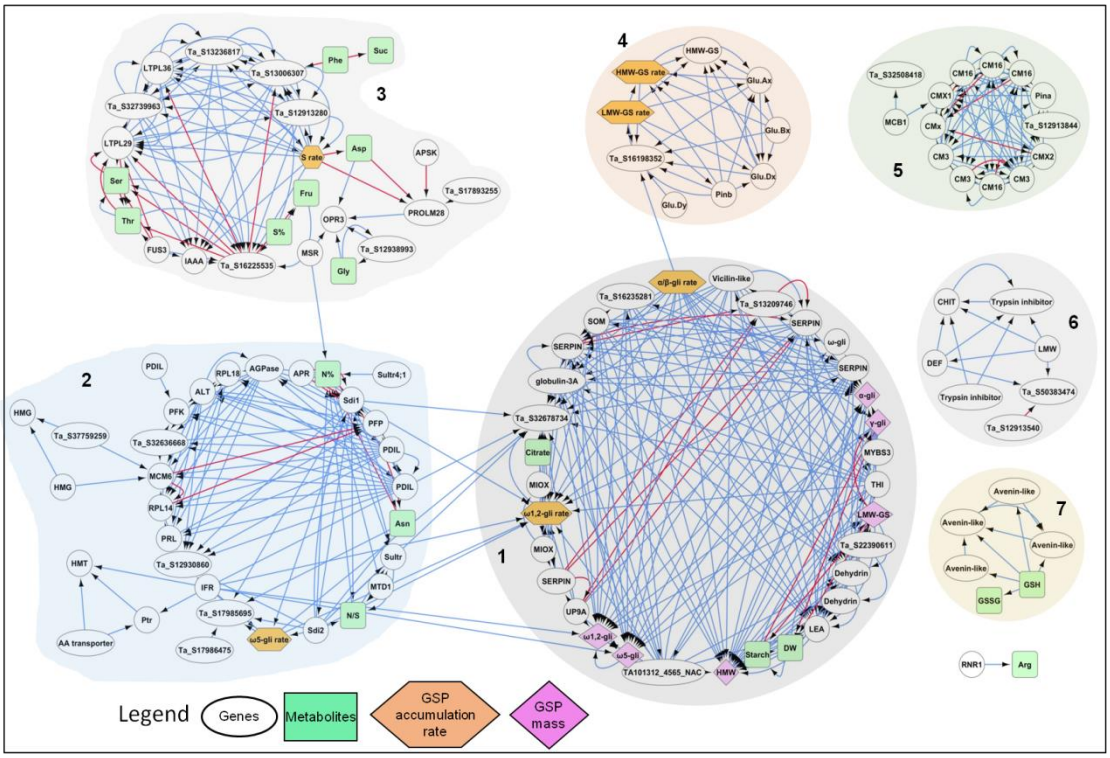


Fig. 9. Directed multi-level regulatory network for grain response to nitrogen (N) and sulfur (S) supply. *Arrows indicates the direction of the linkage from source to target nodes. Node colors and shapes reflect variable categories: genes, open circles; mass of GSP groups per grain, pink diamonds; rate of accumulation of GSP groups, orange hexagons; and metabolites, green squares. Edge colors reflect the biological meaning of the linkage between source and target: blue indicates that when the level of source was low, a low level of target was observed; red indicates that when the level of source was high, a high level of target was observed. For transcripts, either UniGene identifier or abbreviated GSP names are shown (for details, see Supplemental Table 4).*

The resulting network comprised 126 variables (nodes) and 536 significant linkages (edges), which were distributed into 8 modules (Figure 9, Supplemental Table 4). The two largest modules contained 33 (module 1) and 32 (module 2) nodes representing different types of functional variables. Modules 1 comprised mainly storage compounds, including LMW-GS, HMW-GS, ω5-, ω1,2-, α/β- and γ-gliadin mass per grain and the rate of accumulation of ω1,2- and α/β-gliadin, and genes encoding ω-gliadins and non-prolamin GSPs. Module 2 comprised highly responsive variables, notably the most responsive amino acid Asn, the N:S ratio, the rate of accumulation of ω5-gliadins and genes involved in S-deficiency response (APR, Sdi1, Sdi2, Sultr, Sultr4;1, and IFR), amino acid metabolism (HMT, ALT, amino acid transporter), translation (RPL14 and RPL18), and DNA replication (MCM6, PRL, and HMG). This module may pinpoint the underlying processes of a concerted response required by grains to maintain their N and S balance under S-deficiency.

The third module (module 3) contained 24 nodes, including Gly, Phe, Ser, Thr, Asp, Fru, and Suc mass per grain and total S mass per grain and rate of accumulation. The remaining modules were mainly composed of transcripts and their associations agreed well with the clustering analysis (Figure 8). Interestingly, in module 7 GSH showed high associations with the expression of genes encoding avenin-like proteins.

The network also contained three transcription factors known to regulate GSP genes in cereals. Surprisingly these transcription factors were not directly linked to GSP genes. MYBS3 (module 1) was linked to the rate of accumulation α/β-gliadins and the mass per grain of γ-gliadins (Figure 9). FUSCA3 (module 3) was linked to the mass per grain of Ser and Thr, and three transcripts encoding cysteine-rich proteins (Ta_S12902004, Ta_S12902632, Ta_S16225535). Finally, MCB1 (module 5) was linked to a gene with unknown function and to CMX1 a trypsin/alpha-amylase inhibitor. CMX1/CMX3 were involved in a highly connected module composed of genes of a mutligenic family of trypsin/alpha-amylase inhibitor proteins and puroindoline-a (Pina), a small Cyst-rich protein controlling grain hardness (Figure 9).

DISCUSSION

Allometric Allocation of Nitrogen and Sulfur to Grain Storage Proteins is Mainly Dependent on Transcriptional Regulation

Even though the effect of S deficiency on the distribution of GSPs between S-poor and S-rich proteins is well documented for mature wheat grain (Wieser et al., 2004; Zörb et al., 2009), the response of GSP composition to S supply during grain development is largely unknown. Here, we found that, independently of developmental stage and N and S supply, the mass per grain of S-poor GSP scaled allometrically with total N mass per grain, while the mass per grain of S-rich GSP scaled allometrically with total S mass per grain. This suggests that the control of GSP synthesis relies on an elegant regulation mechanism evoking communicating vessels, tightly linked to grain N and S status, in order to maintain total GSP mass per grain under S deficiency by increasing S-poor proteins like HMW-GS, ω5-, ω1,2- and α/β-gliadins and decreasing S-rich proteins like LMW-GS and γ-gliadins in a coordinated manner.

This strict regulation most likely involves the modulation of GSP gene expression. Nucleotide polymorphisms in the promoter region of the wheat transcription factor Storage Protein Activator (Spa-A) were associated with the scaling coefficient of the allometric relationship between total N and gliadin mass per grain (Ravel et al., 2009). Furthermore, a genome-wide association study identified several candidate transcription factors involved in this regulatory mechanism (Plessis et al., 2013). Both studies indicate that the allometric relationships of grain nitrogen allocation underlie the dynamics structuring these regulatory networks in which transcriptional regulation is dominant. Here, we found significant correlations between gene expression and the rate of protein accumulation for all GSPs but γ-gliadins. Thus, transcriptional regulation explained, at least in part, the observed changes in the rate of accumulation of these GSP groups in response to N and S supply. The slope of the relationship for the S-free ω-gliadins was much larger than that for LMW-GS, HMW-GS, and α/β-gliadins, suggesting that the former proteins have a higher translation efficiency.

The signal transduction between nutrient availability and GSP gene expression is still unknown. Our transcriptomic and network analyses pointed out several transcriptional factors (two HMG and one AP2-EREBP, MYBS3, FUSCA3 and MCB1) that may be involved in this signal transduction. HMG transcriptional factors can promote the binding of Dof (DNA-binding with one finger) domain proteins to DNA (Yanagisawa, 1997). Because two Dof domain proteins (PBF and SAD) are known to be involved in the transcriptional regulation of GSP genes (Mena et al., 1998; Diaz et al., 2005), these two HMGs may also play a role in this

regulation by interacting with PBF or SAD. AP2-EREBP may link the crosstalk between ethylene signaling and S metabolism as recently highlighted in plants (Iqbal et al., 2013). MYBS3, whose barley ortholog interacts with BLZ2 and BPBF to regulate hordein gene expression (Rubio-Somoza et al., 2006), was linked with the rate of accumulation of α/β-gliadins. FUSCA3 and MCB1 were linked to cystein-rich proteins and trypsin/alpha-amylase inhibitors, respectively. FUSCA3 is a key regulator of *A. thaliana* seed development (Wang and Perry, 2013) and its barley ortholog also trans-activates the expression of hordein genes (Moreno-Risueno et al., 2008). MCB1 has been implicated in the circadian clock-regulated expression in leaves, is a negative regulator of the GA-induced hydrolase genes in germinating aleurone cells of barley brains, and is a repressor of HvMYBS3 expression (Rubio-Somoza et al., 2006). These transcriptional factors provide good candidates for further functional characterization.

We were also able to generalize our findings on prolamin regulation to other GSPs: serpins and alpha-amylase/trypsin inhibitors. They contain respectively 2 (Rosenkrands et al., 1994) and 8 mol % of Cys and Met, which makes them respectively S-poor and S-rich GSPs. Proteomic analyses have shown that in wheat grain alpha-amylase/trypsin inhibitors can be decreased by N supply (Altenbach et al., 2011a) and S deficiency (Grove et al., 2009), while serpins can be increased by N supply (Altenbach et al., 2011b). Our gene expression results are in tight agreement with these proteomic studies, indicating that these proteins are, at least partly, under transcriptional regulation in response to N and S supply. Furthermore, this means that alpha-amylase/trypsin inhibitors and serpin genes are regulated as prolamins in response to S deficiency.

Adjustments of Amino Acid Transport and Metabolism Are Necessary to Regulate Grain Storage Protein Accumulation in Response to Sulfur Deficiency

The dynamics of metabolite pools, either transported by the phloem or synthesized *de novo* in the developing grains (Tabe et al., 2002), may also play a key role in the signal transduction between nutrient availability and GSP accumulation (Weber et al., 2010; Veliz et al., 2014). Good candidate metabolites for such a role are free amino acids. The availability of free amino acids can have direct effects on the translation efficiency, but it could also have a signaling role (Hernandez-Sebastia et al., 2005; Ohkama-Ohtsu et al., 2008). Amino acid precursors (e.g. O-acetylserine) or small peptides (e.g. GSH) could also serve as signal in the regulation of grain N and S allocation of GSP (Sogawa et al., 2005; Ohkama-Ohtsu et al., 2008). In accordance with the hypothesis that metabolites have an essential role in the regulation of

GSP accumulation in response to nutrient availability, we observed coordinated changes between N, S and C metabolisms and GSP accumulation.

As can be expected, S deficiency had a major effect on S metabolism. Three S transporters and APR, a core regulator of S metabolism, were strongly up-regulated. This concomitant up-regulation of S transport and metabolism upon S deficiency is a relatively well-established response (Maruyama-Nakashita et al., 2003; Nikiforova et al., 2003). The activation of these pathways may improve S uptake and compensate partly for S deficiency.

The concentration of Cys in our samples was undetectable, but the kinetics of Met mass per grain was not modified by the treatments. Two genes (homocysteine S-methyltransferase and cystathionine beta-lyase) involved in Met biosynthesis from Cys were up-regulated under S deficiency, which may explain the observed Met homeostasis under variable S supply. We suggest that the grain may give priority to S utilization for Cys and Met biosynthesis in order to meet the need for GSP accumulation, at the expense of other S-containing compounds. This hypothesis is supported by the observed decrease in GSH mass per grain under S deficiency conditions, in agreement with previous studies (Tea et al., 2005; Reinbold et al., 2008; Zörb et al., 2009).

The effects of S deficiency are not limited to the metabolism of S-containing amino acids. In our study, S deficiency induced a coordinated increase in most free amino acids, with Asn showing the strongest increase of all amino acids. This result is in agreement with previous reports for mature grains of wheat (Granvogl et al., 2007) and could be partly explained by the up-regulation of genes involved in Ser, Gly, and Ala metabolisms. Several amino acid transporters were also up-regulated by S deficiency, suggesting an increased import of amino acid from the mother plant. The import rate of amino acids from phloem to the grain has been shown to be positively correlated with the expression of B- and C-hordeins in barley (Veliz et al., 2014). The critical role of amino acids supply in GSP synthesis regulation is also supported by the increase in GSPs observed in narbon bean (*Vicia narbonensis* L.) and pea (*Pisum sativum* L.) seeds overexpressing a broad bean (*V. faba* L.) amino acid transporter (VfAAP1) (Weigelt et al., 2008).

Changes in primary C metabolism have been found to be linked to modifications in GSP synthesis. For example, the antisense-inhibition of AGPase, the enzyme of the first committed step in starch biosynthesis, in developing seeds of narbon bean resulted in an increase in GSP concentration with a moderate decrease of starch concentration (Weber et al., 2000), and the overexpression of barley sucrose transporter HvSUT1 in wheat endosperm stimulated

GSP synthesis (Weichert et al., 2010). The modifications in C metabolism observed in our study could also be related to GSP accumulation and N metabolism through the glycolysis, which provides precursors for amino acid synthesis. The most important changes caused by S deficiency on C metabolism were an increase in malate and citrate and a decrease in sucrose. This is consistent with a recent metabolomic analysis of mature wheat grain (Zörb et al., 2012) but in contradiction with an older study that found no effect of S supply on sucrose concentration (Granvogl et al., 2007). The decrease in sucrose could be due to the up-regulation of genes coding for enzymes of the glycolytic (PFP, PFK, and TPI) and starch (AGPase) pathways. These pathways can use sucrose imported into the grain as precursor and their up-regulation probably caused more sucrose to be channeled into pyruvate and/or starch, which results in decreased sucrose mass per grain and unchanged starch under S deficiency.

Sulfur Supplied at Mid-Grain Filling Can Compensates for Grain Sulfur Deficiency

Our results emphasize the importance of temporal analyses in order to gain better understanding in the occurrence of important regulatory events in response to N and S supply. We found for example that the increase in most of the free amino acids upon S deficiency started midway through grain filling (ca. 600 °Cdays after anthesis), which indicates that grains were able to maintain their amino acid homeostasis at the early stages of grain development regardless of N and S supply. This developmental stage is also the one when DM and most GSPs stopped accumulating in conditions of S deficiency. Those results pinpoint a critical stage in the grain response to S deficiency, which could help efficiently scheme S fertilizer application in the field.

Recent studies have shown that S fertilization at ear emergence stage can prevent S deficiency at later stages (Zörb et al., 2009; Steinfurth et al., 2012). Here, we show that S supply at 490°Cdays after anthesis (i.e. shortly before the critical stage of S deficiency effects identified here) can also efficiently mitigate S deficiency. This shift in nutrition led to a fast recovery of total S, S-rich GSPs and several metabolites (e.g. GSH, citrate, and free amino acids) that were concomitant with changes in the expression of S transporter and metabolism, amino acid metabolism, oxidative stress response genes.

S transporters are good candidates to explain the swiftness of this recovery, as their expression can respond to S availability within hours (Smith et al., 1997). Our results indicate

that wheat plants possess an efficient S sensing system that tunes S transport and metabolism in response to S availability.

REFERENCES

Agier, M., Petit, J.M., and Suzuki, E. (2007). Unifying framework for rule semantics: Application to gene expression data. Fundamenta Informaticae 78, 543-559.

Altenbach, S. (2012). New insights into the effects of high temperature, drought and post-anthesis fertilizer on wheat grain development. J. Cereal Sci. 56, 39 - 50.

Altenbach, S., Vensel, W., and Dupont, F. (2011a). The spectrum of low molecular weight alpha-amylase/protease inhibitor genes expressed in the US bread wheat cultivar Butte 86. BMC Research Notes 4, 242.

Altenbach, S., Tanaka, C., Hurkman, W., Whitehand, L., Vensel, W., and Dupont, F. (2011b). Differential effects of a post-anthesis fertilizer regimen on the wheat flour proteome determined by quantitative 2-DE. Proteome Science 9, 46.

Altenbach, S.B., and Kothari, K.M. (2007). Omega gliadin genes expressed in Triticum aestivum cv. Butte 86: Effects of post-anthesis fertilizer on transcript accumulation during grain development. J. Cereal Sci. 46, 169-177.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 57, 289 - 300.

Buchner, P., Parmar, S., Kriegel, A., Carpentier, M., and Hawkesford, M.J. (2010). The sulfate transporter family in wheat: tissue-specific gene expression in relation to nutrition. Mol. Plant.

Carceller, J.L., and Aussenac, T. (1999). Accumulation and changes in molecular size distribution of polymeric proteins in developing grains of hexaploid wheats: role of the desiccation phase. Aust. J. Plant Physiol. 26, 301-310.

Carvalho, B.S., and Irizarry, R.A. (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics 26, 2363-2367.

Castle, S.L., and Randall, P.J. (1987). Effects of sulfur deficiency on the synthesis and accumulation of proteins in the developing wheat seed. Aust. J. Plant Physiol. 14, 503-516.

Chope, G.A., Wan, Y., Penson, S.P., Bhandari, D.G., Powers, S.J., Shewry, P.R., and Hawkesford, M.J. (2014). Effects of genotype, season, and nitrogen nutrition on gene expression and protein accumulation in wheat grain. J. Agric. Food Chem. 62, 4399-4407.

Ciaffi, M., Paolacci, A.R., Celletti, S., Catarcione, G., Kopriva, S., and Astolfi, S. (2013). Transcriptional and physiological changes in the S assimilation pathway due to single or combined S and Fe deprivation in durum wheat (Triticum durum L.) seedlings. J. Exp. Bot. 64, 1663-1675.

Cohen, S.A., and Michaud, D.P. (1993). Synthesis of a fluorescent derivatizing reagent, 6-Aminoquinolyl-N-Hydroxysuccinimidyl Carbamate, and its application for the analysis of hydrolysate amino acids via high-performance liquid chromatography. Anal. Biochem. 211, 279-287.

Diaz, I., Martinez, M., Isabel-LaMoneda, I., Rubio-Somoza, I., and Carbonero, P. (2005). The DOF protein, SAD, interacts with GAMYB in plant nuclei and activates transcription of endosperm-specific genes during barley seed development. Plant J. 42, 652-662.

Dupont, F.M., Hurkman, W.J., Vensel, W.H., Tanaka, C., Kothari, K.M., Chung, O.K., and Altenbach, S.B. (2006). Protein accumulation and composition in wheat grains: effects of mineral nutrients and high temperature. Eur. J. Agron. 25, 96-107.

Eriksen, J., and Donald, L.S. (2009). Soil sulfur cycling in temperate agricultural systems. Adv. Agron. 102, 55-89.

Fu, B.X., and Kovacs, M.I.P. (1999). Rapid single-step procedure for isolating total glutenin proteins of wheat flour. J. Cereal Sci. 29, 113-116.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5, R80.

Granvogl, M., Wieser, H., Koehler, P., VonTucher, S., and Schieberle, P. (2007). Influence of sulfur fertilization on the amounts of free amino acids in wheat. Correlation with baking properties as well as with 3-aminopropionamide and acrylamide generation during baking. Journal of Agricultural Food Chemistry 55, 4271-4277.

Griffith, O.W. (1980). Determination of glutathione and glutathione disulfide using glutathione reductase and 2-vinylpyridine. Anal. Biochem. 106, 207-212.

Grove, H., Hollung, K., Moldestad, A., Færgestad, E.M., and Uhlen, A.K. (2009). Proteome changes in wheat subjected to different nitrogen and sulfur fertilizations. J. Agric. Food Chem. 57, 4250-4258.

Hendriks, J.H.M., Kolbe, A., Gibon, Y., Stitt, M., and Geigenberger, P. (2003). ADP-glucose pyrophosphorylase is activated by posttranslational redox-modification in response to light and to sugars in leaves of Arabidopsis and other plant species. Plant Physiol. 133, 838-849.

Hernandez-Sebastia, C., Marsolais, F., Saravitz, C., Israel, D., Dewey, R.E., and Huber, S.C. (2005). Free amino acid profiles suggest a possible role for asparagine in the control of storage-product accumulation in developing seeds of low- and high-protein soybean lines. J. Exp. Bot. 56, 1951-1963.

Higashi, Y., Hirai, M.Y., Fujiwara, T., Naito, S., Noji, M., and Saito, K. (2006). Proteomic and transcriptomic analysis of Arabidopsis seeds: molecular evidence for successive processing of seed proteins and its implication in the stress response to sulfur nutrition. Plant J. 48, 557-571.

Hirai, M.Y., Fujiwara, T., Chino, M., and Naito, S. (1995). Effects of sulfate concentrations on the expression of a soybean seed storage protein gene and its reversibility in transgenic Arabidopsis thaliana. Plant Cell Physiol. 36, 1331-1339.

Howarth, J.R., Parmar, S., Barraclough, P.B., and Hawkesford, M.J. (2009). A sulphur deficiency-induced gene, sdi1, involved in the utilization of stored sulphate pools under sulphur-limiting conditions has potential as a diagnostic indicator of sulphur nutritional status. Plant Biotechnology Journal 7, 200-209.

Hurkman, W., Tanaka, C., Vensel, W., Thilmony, R., and Altenbach, S. (2013). Comparative proteomic analysis of the effect of temperature and fertilizer on gliadin and glutenin accumulation in the developing endosperm and flour from Triticum aestivum L. cv. Butte 86. Proteome Science 11, 8.

Iqbal, N., Masood, A., Khan, M., Asgher, M., Fatma, M., and Khan, N. (2013). Cross-talk between sulfur assimilation and ethylene signaling in plants. Plant Signaling & Behavior 8, e22478.

Jelitto, T., Sonnewald, U., Willmitzer, L., Hajirezeai, M., and Stitt, M. (1992). Inorganic pyrophosphate content and metabolites in potato and tobacco plants expressing E. coli pyrophosphatase in their cytosol. Planta 188, 238-244.

Kawashima, C.G., Yoshimoto, N., Maruyama-Nakashita, A., Tsuchiya, Y.N., Saito, K., Takahashi, H., and Dalmay, T. (2009). Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different cell types. Plant J. 57, 313-321.

Lewandowska, M., and Sirko, A. (2008). Recent advances in understanding plant response to sulfur-deficiency stress. Acta Biochim. Pol. 55, 457-471.

MacRitchie, F., and Gupta, R.B. (1993). Functionality-composition relationships of wheat flour as a result of variation in sulfur availability. Aust. J. Agric. Res. 44, 1767-1774.

Maruyama-Nakashita, A., Inoue, E., Watanabe-Takahashi, A., Yamaya, T., and Takahashi, H. (2003). Transcriptome profiling of sulfur-responsive genes in Arabidopsis reveals global effects of sulfur nutrition on multiple metabolic pathways. Plant Physiol. 132, 597 - 605.

Maruyama-Nakashita, A., Nakamura, Y., Tohge, T., Saito, K., and Takahashi, H. (2006). Arabidopsis SLIM1 is a central transcriptional regulator of plant sulfur response and metabolism. Plant Cell 18, 3235-3251.

Maruyama-Nakashita, A., Nakamura, Y., Watanabe-Takahashi, A., Inoue, E., Yamaya, T., and Takahashi, H. (2005). Identification of a novel cis-acting element conferring sulfur deficiency response in Arabidopsis roots. Plant J. 42, 305-314.

Matthewman, C.A., Kawashima, C.G., Húska, D., Csorba, T., Dalmay, T., and Kopriva, S. (2012). miR395 is a general component of the sulfate assimilation regulatory network in Arabidopsis. FEBS Lett. 586, 3242-3248.

Mena, M., Vicente-Carbajosa, J., Schmidt, R.J., and Carbonero, P. (1998). An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. Plant J. 16, 53-62.

Moreno-Risueno, M.A., González, N., Díaz, I., Parcy, F., Carbonero, P., and Vicente-Carbajosa, J. (2008). FUSCA3 from barley unveils a common transcriptional regulation of seed-specific genes between cereals and Arabidopsis. Plant J. 53, 882-894.

Mossé, J., Huet, J.C., and Baudet, J. (1985). The amino acid composition of wheat grain as a function of nitrogen content. J. Cereal Sci. 3, 115-130.

Müller, M., and Knudsen, S. (1993). The nitrogen response of a barley C-hordein promoter is controlled by positive and negative regulation of the GCN4 and endosperm box. Plant J. 4, 343-355.

Nikiforova, V., Freitag, J., Kempa, S., Adamik, M., Hesse, H., and Hoefgen, R. (2003). Transcriptome analysis of sulfur depletion in Arabidopsis thaliana: interlacing of biosynthetic pathways provides response specificity. Plant J. 33, 633 - 650.

Nunes-Nesi, A., Carrari, F., Gibon, Y., Sulpice, R., Lytovchenko, A., Fisahn, J., Graham, J., Ratcliffe, R.G., Sweetlove, L.J., and Fernie, A.R. (2007). Deficiency of mitochondrial fumarase activity in tomato plants impairs photosynthesis via an effect on stomatal function. The Plant Journal 50, 1093-1106.

Ohkama-Ohtsu, N., Kezuka, A., Onouchi, H., Fujiwara, T., and Naito, S. (2008). Promoter region of the β subunit gene of β-conglycinin responds to methionine and glutathione in transient assays using Arabidopsis protoplasts. Soil Sci. Plant Nutr. 54, 128-132.

Pfaffl, M., Tichopad, A., Prgomet, C., and Neuvians, T. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper -Excel-based tool using pair-wise correlations. Biotechnol. Lett. 26, 509 - 515.

Plessis, A., Ravel, C., Bordes, J., Balfourier, F., and Martre, P. (2013). Association study of wheat grain protein composition reveals that gliadin and glutenin composition are trans-regulated by different chromosome regions. J. Exp. Bot. 64, 3627-3644.

R Development Core Team. (2010). R: A language and environment for statistical computing. (Vienna, Austria: R Foundation for Statistical Computing).

Randall, P., Spencer, K., and Freney, J. (1981). Sulfur and nitrogen fertilizer effects on wheat. I. Concentrations of sulfur and nitrogen and the nitrogen to sulfur ratio in grain, in relation to the yield response. Aust. J. Agric. Res. 32, 203-212.

Randall, P.J., and Wrigley, C.W. (1986). Effects of sulfur supply on the yield, composition, and quality of grain from cereals, oilseeds, and legumes. In Advances in Cereal Science and Technology, Y. Pomeranz, ed (St. Paul, Minnesota USA: American Association of Cereal Chemists), pp. 171-206.

Ravel, C., Martre, P., Romeuf, I., Dardevet, M., El-Malki, R., Bordes, J., Duchateau, N., Brunel, D., Balfourier, F., and Charmet, G. (2009). Nucleotide polymorphism in the wheat transcriptional activator Spa influences its pattern of expression and has pleiotropic effects on grain protein composition, dough viscoelasticity, and grain hardness. Plant Physiol. 151, 2133-2144.

Reinbold, J., Rychlik, M., Asam, S., Wieser, H., and Koehler, P. (2008). Concentrations of total glutathione and cysteine in wheat flour as affected by sulfur deficiency and correlation to quality parameters. J. Agric. Food Chem. 56, 6844-6850.

Romeuf, I., Tessier, D., Dardevet, M., Branlard, G., Charmet, G., and Ravel, C. (2010). wDBTF: an integrated database resource for studying wheat transcription factor families. BMC Genomics 11, 185.

Rosenkrands, I., Hejgaard, J., Rasmussen, S.K., and Bjørn, S.E. (1994). Serpins from wheat grain. FEBS Lett. 343, 75-80.

Rubio-Somoza, I., Martinez, M., Abraham, Z., Diaz, I., and Carbonero, P. (2006). Ternary complex formation between HvMYBS3 and other factors involved in transcriptional control in barley seeds. Plant J. 47, 269-281.

Shewry, P.R., and Halford, N.G. (2002). Cereal seed storage proteins: structures, properties and role in grain utilization. J. Exp. Bot. 53, 947-958.

Shewry, P.R., Tatham, A.S., and Lazzeri, P. (1997). Biotechnology of wheat quality. J. Sci. Food Agric. 73, 397-406.

Shewry, P.R., Tatham, A.S., and Halford, N.G. (2001). Nutritional control of storage protein synthesis in developing grain of wheat and barley. Plant Growth Regul. 34, 105-111.

Smith, F.W., Hawkesford, M.J., Ealing, P.M., Clarkson, D.T., Vanden Berg, P.J., Belcher, A.R., and Warrilow, A.G.S. (1997). Regulation of expression of a cDNA from barley roots encoding a high affinity sulphate transporter. The Plant Journal 12, 875-884.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27, 431-432.

Smyth, G.K. (2005). Limma: linear models for microarray data. In: '. In Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, eds (New York: Springer), pp. 397-420.

Sogawa, Y., Ohkama-Ohtsu, N., Hayashi, H., Yoneyama, T., and Fujiwara, T. (2005). Independent roles of glutathione and O-acetyl-L-serine in regulation of sulfur-responsive gene expression in Arabidopsis thaliana. Plant Biotechnology 22, 51-54.

Spinelli, L., Gambette, P., Chapple, C.E., Robisson, B., Baudot, A., Garreta, H., Tichit, L., Guénoche, A., and Brun, C. (2013). Clust&See: A Cytoscape plugin for the identification, visualization and manipulation of network clusters. Biosystems 113, 91-95.

Steinfurth, D., Zörb, C., Braukmann, F., and Mühling, K.H. (2012). Time-dependent distribution of sulphur, sulphate and glutathione in wheat tissues and grain as affected by three sulphur fertilization levels and late S fertilization. J. Plant Physiol. 169, 72-77.

Tabe, L., Hagan, N., and Higgins, T.J.V. (2002). Plasticity of seed protein composition in response to nitrogen and sulfur availability. Curr. Opin. Plant Biol. 5, 212-217.

Tea, I., Genter, T., Violleau, F., and Kleiber, D. (2005). Changes in the glutathione thiol-disulfide status in wheat grain by foliar sulphur fertilization: consequences for the rheological properties of dough. J. Cereal Sci. 41, 305-315.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 37, 914-939.

Tompkins, D., and Toffaletti, J. (1982). Enzymic determination of citrate in serum and urine, with use of the Worthington "ultrafree" device. Clin. Chem. 28, 192-195.

Triboi, E., Martre, P., and Triboi-Blondel, A.-M. (2003). Environmentally-induced changes of protein composition for developing grains of wheat are related to changes in total protein content. J. Exp. Bot. 54, 1731-1742.

Veliz, C.G., Criado, M.V., Roberts, I.N., Echeverria, M., Prieto, P., Gutierrez Boem, F.H., and Caputo, C. (2014). Phloem sugars and amino acids as potential regulators of hordein expression in field grown malting barley (Hordeum vulgare L.). J. Cereal Sci. (in press), DOI: 10.1016/j.jcs.2014.1004.1008.

Vincent, J., Dai, Z., Ravel, C., Choulet, F., Mouzeyar, S., Bouzidi, M.F., Agier, M., and Martre, P. (2013). dbWFA: a web-based database for functional annotation of Triticum aestivum transcripts. Database 2013, doi: 10.1093/database/bat1014.

Wan, Y., Shewry, P.R., and Hawkesford, M.J. (2013). A novel family of γ-gliadin genes are highly regulated by nitrogen supply in developing wheat grain. J. Exp. Bot. 64, 161-168.

Wan, Y., Gritsch, C.S., Hawkesford, M.J., and Shewry, P.R. (2014). Effects of nitrogen nutrition on the synthesis and deposition of the ω-gliadins of wheat. Ann. Bot. 113, 607-615.

Wang, F., and Perry, S.E. (2013). Identification of direct targets of FUSCA3, a key regulator of Arabidopsis seed development. Plant Physiol. 161, 1251-1264.

Warton, D.I., Duursma, R.A., Falster, D.S., and Taskinen, S. (2012). smatr 3- an R package for estimation and inference about allometric lines. Methods in Ecology and Evolution 3, 257-259.

Wawrzyńska, A., Lewandowska, M., and Sirko, A. (2010). Nicotiana tabacum EIL2 directly regulates expression of at least one tobacco gene induced by sulphur starvation. J. Exp. Bot. 61, 889-900.

Wawrzyńska, A., Lewandowska, M., Hawkesford, M.J., and Sirko, A. (2005). Using a suppression subtractive library-based approach to identify tobacco genes regulated in response to short-term sulphur deficit. J. Exp. Bot. 56, 1575-1590.

Weber, H., Sreenivasulu, N., and Weschke, W. (2010). Molecular physiology of seed maturation and seed storage protein biosynthesis. In Plant Developmental Biology - Biotechnological Perspectives, E.C. Pua and M.R. Davey, eds (Springer Berlin Heidelberg), pp. 83-104.

Weber, H., Rolletschek, H., Heim, U., Golombek, S., Gubatz, S., and Wobus, U. (2000). Antisense-inhibition of ADP-glucose pyrophosphorylase in developing seeds of Vicia narbonensis moderately decreases starch but increases protein content and affects seed maturation. Plant J. 24, 33-43.

Weichert, N., Saalbach, I., Weichert, H., Kohl, S., Erban, A., Kopka, J., Hause, B., Varshney, A., Sreenivasulu, N., Strickert, M., Kumlehn, J., Weschke, W., and Weber, H. (2010). Increasing sucrose uptake capacity of wheat grains stimulates storage protein synthesis. Plant Physiol. 152, 698-710.

Weigelt, K., Küster, H., Radchuk, R., Müller, M., Weichert, H., Fait, A., Fernie, A.R., Saalbach, I., and Weber, H. (2008). Increasing amino acid supply in pea embryos reveals specific interactions of N and C metabolism, and highlights the importance of mitochondrial metabolism. Plant J. 55, 909-926.

Wieser, H. (2007). Chemistry of gluten proteins. Food Microbiology 24, 115-119.

Wieser, H., and Seilmeier, W. (1998). The influence of nitrogen fertilisation on quantities and proportions of different protein types in wheat flour. J. Sci. Food Agric. 76, 49-55.

Wieser, H., Antes, S., and Seilmeier, W. (1998). Quantitative determination of gluten protein types in wheat flour by reversed-phase high-performance liquid chromatography. Cereal Chem. 75, 644-650.

Wieser, H., Guster, R., and von Tucher, S. (2004). Influence of sulphur fertilisation on quantities and proportions of gluten protein types in wheat flour. J. Cereal Sci. 40, 239-244.

Yanagisawa, S. (1997). Dof DNA-binding domains of plant transcription factors contribute to multiple protein-protein interactions. Eur. J. Biochem. 250, 403-410.

Zhao, F.J., Hawkesford, M.J., and McGrath, S.P. (1999a). Sulphur assimilation and effects on yield and quality of wheat. J. Cereal Sci. 30, 1-17.

Zhao, F.J., Salmon, S.E., Withers, P.J.A., Evans, E.J., Monaghan, J.M., Shewry, P.R., and McGrath, S.P. (1999b). Responses of breadmaking quality to sulphur in three wheat varieties. J. Sci. Food Agric. 79, 1865-1874.

Zörb, C., Grover, C., Steinfurth, D., and Hermann Mühling, K. (2010). Quantitative proteome analysis of wheat gluten as influenced by N and S nutrition. Plant Soil 327, 225-234.

Zörb, C., Steinfurth, D., Gödde, V., Niehaus, K., and Mühling, K.H. (2012). Metabolite profiling of wheat flag leaf and grains during grain filling phase as affected by sulfur fertilisation. Funct. Plant Biol. 39, 156-166.

Zörb, C., Steinfurth, D., Seling, S., Langenkämper, G., Koehler, P., Wieser, H., Lindhauer, M.G., and Mühling, K.H. (2009). Quantitative protein composition and baking quality of winter wheat as affected by late sulfur fertilization. J. Agric. Food Chem. 57, 3877-3885.

SUPPLEMENTARY INFORMATION

Supplemental Table 1. Dry mass, N and S mass and concentration and protein composition for mature wheat grains grown with four combinations of N and S supply.
Data are means ± 1 SE for *n* = 3 independent replicates. Different letters within a row indicate significant differences (= 5) among N/S treatments following a Tukey test. *P*-value from ANOVAs for the four N/S treatments are reported.

	Treatments				
Variables	NS	NS-	N+S-	N+S-/N+S	*P*-value
Grain dry mass (mg DM grain^{-1})	36.27±0.6a	35.7±1.48a	31.45±2.46a	29.69±1.16a	0.048
Grain total N (mg N grain^{-1})	0.69±0.03b	0.78±0.04ab	0.85±0.02a	0.85±0.03a	0.025
Grain total S (mg S grain^{-1})	0.0478±0.0018a	0.0426±0.0013a	0.0344±0.0004b	0.0483±0.0017a	0.00036
Grain S concentration (g S 100 g^{-1} DM)	0.13±0.005b	0.12±0.004b	0.11±0.008b	0.16±0.001a	0.00044
Grain protein concentration	10.74±0.38b	12.3±0.26b	15.25±0.92a	16.09±0.76a	0.0011
Grain N:S ratio	14.5±0.26c	18.33±0.58b	24.57±0.41a	17.6±0.85b	1.12E-05
Total gliadin (% of total grain N)	16.65±0.59b	21.34±0.46a	23.44±1.26a	24.16±0.43a	0.00023
Total glutenin (% of total grain N)	28.68±1.79a	26.13±1.83ab	21.01±0.22b	27.09±0.92ab	0.020
Gliadin:glutenin ratio	0.58±0.02c	0.82±0.05b	1.1±0.06a	0.89±0.04b	0.00039
ω5-gliadin (% of total grain N)	0.65±0.01c	1.65±0.09b	2.75±0.17a	1.91±0.06b	3.91E-06
ω1,2-gliadin (% of total grain N)	0.8±0.07d	1.59±0.05c	2.8±0.07a	2.22±0.08b	1.77E-07
α/β -gliadin (% of total grain N)	9.33±0.38c	13.37±0.31b	14.43±0.53ab	15.32±0.36a	2.85E-05
γ -gliadin (% of total grain N)	5.86±0.14a	4.73±0.18b	3.1±0.07c	4.69±0.09b	2.92E-05
LMW-GS (% of total grain N)	22.43±1.47a	17.89±1.79a	12.18±0.16b	18.6±0.44a	0.0020
HMW-GS (% of total grain N)	6.25±0.37b	8.24±0.27a	8.84±0.13a	8.48±0.5a	0.0029
HMW-GS:LMW-GS ratio	0.28±0.01c	0.47±0.04b	0.73±0.01a	0.46±0.02b	1.20E-05

Supplemental Table 2. Estimated rate and duration of the accumulation of grain dry mass, total N, total S, and grain storage proteins.
Data were fitted with a 3-parameters logistic function equation:

$$Q = \frac{Q_{max}}{1 + 0.05\exp\left(\frac{-4r(t - t_{95})}{P_{max}}\right)}$$

Where Q is the quantity of dry mass or N per grain, t is the degree-days after anthesis, Q_{max} is the final value of Q approached as $t \rightarrow \infty$, r is the maximum rate of accumulation defined as the derivative of the point of inflexion, and t_{95} is the duration of accumulation defined as the duration, from anthesis, in which 95% of Q_{max} is accumulated.

		Q_{max}		t_{95}		r	
Variables	Treatments	Estimate	SE	Estimate	SE	Estimate	SE
Dry mass	NS	35.1	1.51	774.5	84.4	0.064	0.0099
	NS-	35.4	0.56	822.2	29.4	0.061	0.0030
	N+S-	31.5	0.58	687.5	29.8	0.069	0.0046
	N+S-/N+S	29.2	0.86	631.7	47.1	0.072	0.0089
Nitrogen	NS	0.71	0.037	968.3	111.2	0.0010	0.000125
	NS-	0.80	0.011	884.9	27.2	0.0012	4.81×10^{-5}
	N+S-	0.84	0.015	725.9	28.7	0.0017	9.94×10^{-5}
	N+S-/N+S	0.83	0.020	705.3	40.3	0.0018	1.56×10^{-4}
Sulfur	NS	0.048	0.0036	906.4	179.7	0.000065	1.50×10^{-4}
	NS-	0.043	0.0008	706.2	35.8	0.000086	6.47×10^{-6}
	N+S-	0.036	0.0010	568.6	46.4	0.000093	1.31×10^{-5}
	N+S-/N+S	0.050	0.0011	840.9	44.7	0.000080	5.65×10^{-6}
LMW	NS	0.16	0.0085	828.3	77.6	0.00036	5.56×10^{-5}
	NS-	0.15	0.0045	742.0	36.5	0.00038	3.50×10^{-5}
	N+S-	0.11	0.0073	659.3	73.7	0.00035	7.64×10^{-5}
	N+S-/N+S	0.18	0.0112	776.1	79.1	0.00047	9.92×10^{-5}
HMW	NS	0.045	0.0026	884.6	88.7	0.000089	1.31×10^{-5}
	NS-	0.060	0.0023	760.9	48.1	0.000148	1.69×10^{-5}
	N+S-	0.072	0.0020	578.6	27.7	0.000309	3.65×10^{-5}
	N+S-/N+S	0.071	0.0019	568.7	27.6	0.000315	3.87×10^{-5}
ω5-gliadins	NS	0.005	0.0004	1115.4	129.0	0.000007	9.94×10^{-7}
	NS-	0.015	0.0032	1432.0	336.0	0.000016	2.91×10^{-6}
	N+S-	0.024	0.0022	1111.1	136.8	0.000036	5.95×10^{-6}
	N+S-/N+S	0.014	0.0013	736.8	120.9	0.000037	1.14×10^{-5}
ω1,2-gliadins	NS	0.006	0.0001	967.5	24.1	0.000012	4.92×10^{-7}
	NS-	0.013	0.0005	1023.5	57.5	0.000022	1.79×10^{-6}
	N+S-	0.024	0.0011	817.3	53.2	0.000057	6.81×10^{-6}
	N+S-/N+S	0.018	0.0007	655.9	42.5	0.000060	8.80×10^{-6}
α/β-gliadins	NS	0.065	0.0019	905.5	37.5	0.00014	1.05×10^{-5}
	NS-	0.103	0.0032	869.6	42.1	0.00023	2.10×10^{-5}
	N+S-	0.122	0.0029	739.2	26.8	0.00034	2.42×10^{-5}
	N+S-/N+S	0.119	0.0075	800.9	84.7	0.00028	5.40×10^{-5}
γ-gliadins	NS	0.040	0.0017	940.6	63.5	0.000075	7.16×10^{-6}
	NS-	0.040	0.0032	848.5	116.0	0.000080	1.74×10^{-5}
	N+S-	0.028	0.0013	524.7	52.4	0.000122	2.96×10^{-5}
	N+S-/N+S	0.039	0.0019	764.2	67.4	0.000089	1.34×10^{-5}

Supplemental Table 3. Summary statistics of linear regressions between the rate of accumulation of grain storage protein subunits or classes and their gene expression for developing wheat grains grown in the growth chamber under three combinations of N and S supply.

Slopes and intercepts are based on standard major axis regressions. CI, confidence intervals. Different letters indicate a statistical difference (= 0.05) in the intercept or slope. Data are plotted in Figure 8.

	Intercept ($\times 10^{-3}$ µg N grain^{-1} [^{o}Cday]$^{-1}$)		Slope ($\times 10^{-3}$ µg N grain^{-1} [^{o}Cday]$^{-1}$ [Normalized expression]$^{-1}$)			
Protein	Estimated	95% CI	Estimated	95% CI	r^2	*P*-value
LMW-GS	40.9 [a]	-25.9 to 107.6	8.6 [b]	6.1 to 12.1	0.59	< 0.001
HMW-GS	-6.8 [a]	-29.3 to 15.7	9.4 [b]	7.7 to 11.4	0.87	< 0.001
ω5-gliadins	0.5 [a]	-3.6 to 4.5	190.9 [c]	155.1 to 234.9	0.86	< 0.001
ω1,2-gliadins	-2.5 [a]	-11.0 to 6.02	290.9 [d]	219.7 to 385.1	0.73	< 0.001
α/β-gliadins	-16.6 [a]	-75.4 to 42.2	8.7 [b]	6.4 to 11.8	0.68	< 0.002
γ-gliadins	9.0 [a]	-13.7 to 31.7	1.6 [a]	1.0 to 2.6	0.21	0.062

Supplemental Table 4. Clusters and functional annotations of the 122 differentially affected genes in response to N and S supply.
The probe ID of each gene listed in the heatmap (Figure 5) was provided following its putative functional category and functional annotation. Abbreviation was provided for each gene when available, and these abbreviations were used in the directed multi-level regulator network for node name (Figure 9). Cluster N° indicates the cluster in which a gene is involved in the heatmap (Figure 5), and Module N° indicates the network module in which a gene is involved in the network (Figure 9).

Probe ID	Abbreviation	Cluster N°	Module N°	Functional Category	Functional Annotation
Ta_S12897113	CM16	1	5		Alpha-amylase/trypsin inhibitor CM16 precursor (Chloroform methanol-soluble protein CM16)
Ta_S12902327	CM3	1	5	Cysteine-rich	Alpha-amylase/trypsin inhibitor CM3 precursor (Chloroform methanol-soluble protein CM3)
Ta_S12903063	CM16	1	5	Cysteine-rich	Alpha-amylase/trypsin inhibitor CM16 precursor (Chloroform methanol-soluble protein CM16)
Ta_S12913304	CM16	1		Cysteine-rich	Alpha-amylase/trypsin inhibitor CM16 precursor (Chloroform methanol-soluble proteinCM16)
Ta_S12923041	CMX2	1	5	Cysteine-rich	Trypsin/alpha-amylase inhibitor CMX2 precursor
Ta_S13073077	CM2	1		Cysteine-rich	Alpha-amylase/trypsin inhibitor CM2 precursor
Ta_S13109281	CM**X1**	1	5	Cysteine-rich	Trypsin/alpha-amylase inhibitor CMX1 CMX3 precursor
Ta_S13236140	CM16	1	5	Cysteine-rich	Alpha-amylase/trypsin inhibitor CM16 precursor (Chloroform methanol-soluble protein CM16)
Ta_S13236387	CM16	1	5	Cysteine-rich	Alpha-amylase/trypsin inhibitor CM16 precursor (Chloroform methanol-soluble protein CM16)
Ta_S13243416	CM3	1	5	Cysteine-rich	Alpha-amylase/trypsin inhibitor CM3 precursor (Chloroform methanol-soluble protein CM3)
Ta_S17989754	CM3	1	5	Cysteine-rich	Alpha-amylase/trypsin inhibitor CM3 precursor (Chloroform methanol-soluble protein CM3)
Ta_S50382456	CMx	1	5	Cysteine-rich	CMx; trypsin inhibitor
Ta_S13109604	GST	2		misc	glutathione S-transferase [EC 2.5.1.18]
Ta_S12902004	LTPL29	2	3	Cysteine-rich:proteinase inhibitor	Protease inhibitor/seed storage/LTP family protein
Ta_S12902632	IAAA	2	3	Cysteine-rich:proteinase inhibitor	Alpha-amylase/trypsin inhibitor CMa precursor (Chloroform methanol-soluble protein CMa)
Ta_S13236817	Ta_S13236817	2	3	Cysteine-rich:proteinase inhibitor	WCI proteinase inhibitor
Ta_S13260018	LTPL36	2	3	Cysteine-rich:proteinase inhibitor	Protease inhibitor/seed storage/LTP family protein

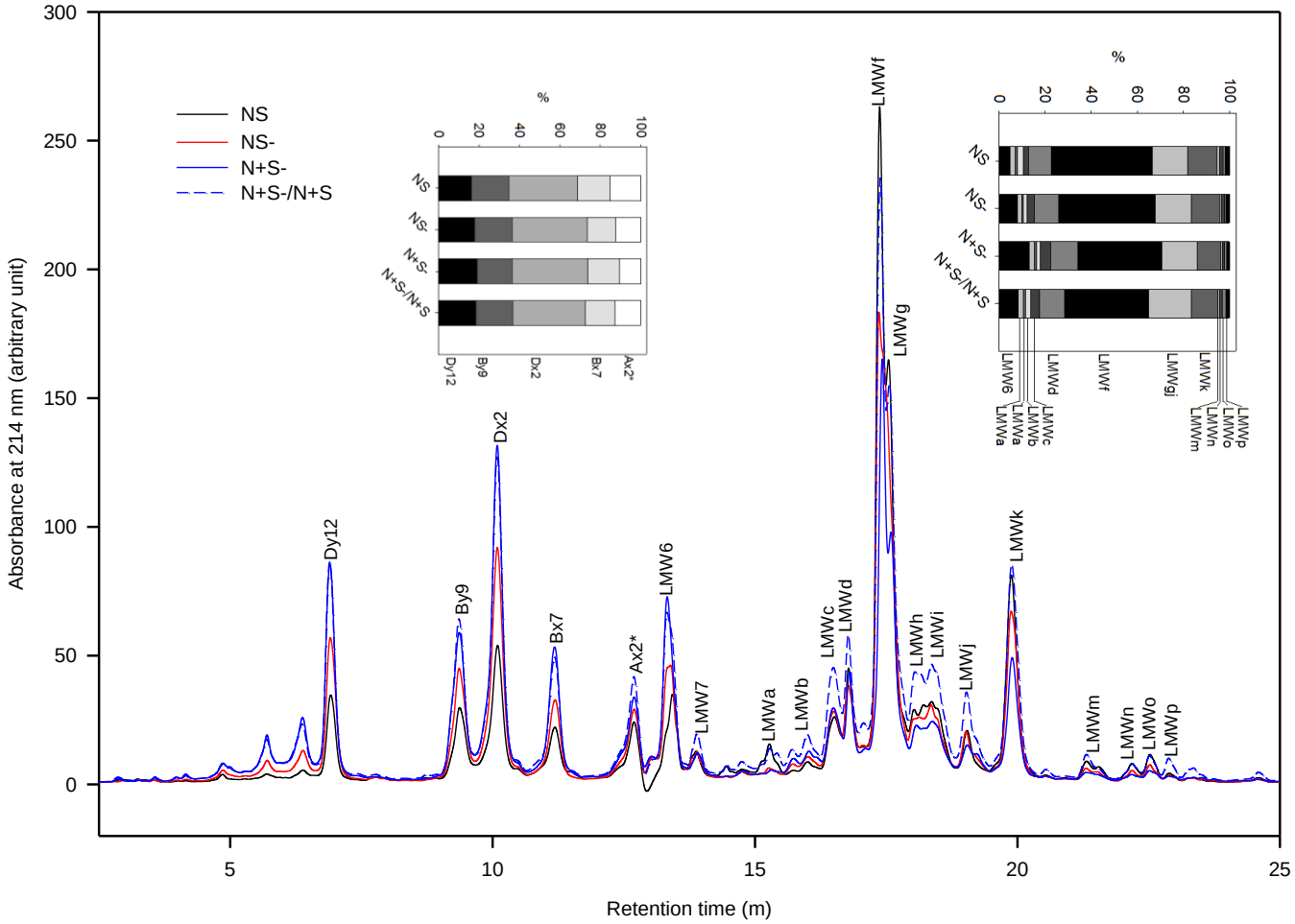
Ta_S32739963	Ta_S32739963	2	3	unknown	unknown
Ta_S32532842	ZmEBE-1	2		misc	ZmEBE-1 protein, within unknown function
Ta_S16254769	Ta_S16254769	2		unknown	unknown
Ta_S12913280	Ta_S12913280	2	3	unknown	unknown
Ta_S13006307	Ta_S13006307	2	3	unknown	unknown
Ta_S12940984	Ta_S12940984	2		Defense response	invertase/pectin methylesterase inhibitor superfamily
Ta_S16225535	Ta_S16225535	2	3	Cysteine-rich	low-molecular-weight cysteine-rich 72
Ta_S12981167	Ta_S12981167	3		unknown	Unknown
Ta_S12903036	Avenin-like	3	7	Storage protein	avenin like seed protein
Ta_S12915182	Avenin-like	3	7	Storage protein	avenin-like seed protein
Ta_S17988333	Avenin-like	3	7	Storage protein	avenin-like protein s1
Ta_S26023317	Avenin-like	3	7	Storage protein	avenin-like protein s1
Ta_S12950024	Ta_S12950024	4		Pectin metabolism	polygalacturonase [EC 3.2.1.15]
Ta_S13135075	Expansin	4		Cell expansion	Expansin
Ta_S13165283	Expansin	4		Cell expansion	Unknown
Ta_S16198352	Ta_S16198352	4	4	unknown	Unknown
Ta_S16246404	Ta_S16246404	4		Protein processing	Calcineurin
Ta_S22390741	Ta_S22390741	4		unknown	Unknown
Ta_S32508418	Ta_S32508418	4	5	unknown	Unknown
Ta_S32585899	Ta_S32585899	4		CHO	Glycosyl Hydrolase
Ta_S37941838	Ta_S37941838	4		unknown	Unknown
Ta_S12913540	Ta_S12913540	5	6	unknown	unknown
Ta_S12913844	IAAE	5	5	Cysteine-rich	Trypsin inhibitor CMeprecursor (Chloroform methanol-soluble protein CMe)
Ta_S12914794	Ta_S12914794	5		Cystein-rich	Glycine-rich protein family, retrotransposon protein
Ta_S12923152	BBTI11	5	6	Cysteine-rich	Bowman-Birk type bran trypsin inhibitor
Ta_S13220528	BBTI12	5	6	Cysteine-rich	Bowman-Birk type bran trypsin inhibitor
Ta_S16058417	DEF	5	6	defense response	Defensin and Defensin-like DEFL family

Ta_S17987091	CHIT	5	6	defense response	Chitinase
Ta_S22390611	Ta_S22390611	5	1	defense response	defense response
Ta_S50383474	AMBP1	5	6	defense response	Antimicrobial peptide MBP-1 family protein
Ta_S12923098	SERPIN	6	1	Storage protein	Serine protease inhibitor (SERPIN) family protein
Ta_S12923099	SERPIN	6	1	Storage protein	Serine protease inhibitor (SERPIN) family protein
Ta_S12923100	SERPIN	6	1	Storage protein	Serine protease inhibitor (SERPIN) family protein
Ta_S12923101	SERPIN	6	1	Storage protein	Serine protease inhibitor (SERPIN) family protein
Ta_S16235281	Ta_S16235281	6	1	Cysteine-rich	protein inhibitor family protein
Ta_S16058430	THI	6	1	stress response and sulfur metabolism	Thiazole biosynthetic enzyme
Ta_S13008421	Dehydrin	6	1	Cystein-rich	Dehydrin family protein
Ta_S17888539	Dehydrin	6	1	Cystein-rich	dehydrin family protein
Ta_S13173393	MIOX	6	1	Oxidative response	myo-inositol oxygenase [EC 1.13.99.1]
Ta_S18008306	MIOX	6	1	Oxidative response	myo-inositol oxygenase [EC 1.13.99.1]
Ta_S26027452	SMO	6	1	Oxidative response	sterol-4alpha-methyl oxidase 1-1
Ta_S32601090	TPP	6		CHO:starch and sucrose	trehalose 6-phosphate phosphatase [EC 3.1.3.12]
Ta_S32707577	globulin-3A	6	1	Storage protein	globulin-3A, seed storage protein
Ta_S13242938	Vicilin-like	6	1	Storage protein	Vicilin-like storage protein
Ta_S13099497	LEA	6	1	LEA	late embryogenesis abundant domain-containing protein
TA101312_4565_NAC	TA101312_4565_NAC	6	1	TF	NAC TF
CA652994_bZIP	CA652994_bZIP	6		TF	ZIP TF
Ta_S16266289	Ptr	7	2	Phosphate translocation	Phosphate/phosphate translocator
Ta_S17985695	Ta_S17985695	7	2	Oxidative stress	Copper chaperone , anti-oxidant 1, heavy metal-associated domain containing protein
Ta_S13209746	Ta_S13209746	7	1	Cation transport	Cation transport regulator-like protein 1
Ta_S17881015	Ta_S17881015	7		unknown	Unknown
Ta_S32642859	Ta_S32642859	7		unknown	Unknown
Ta_S32678734	Ta_S32678734	7	1	unknown	Unknown
Ta_S32554654	Gly-rich	7		Cystein-rich	Glycine-rich protein

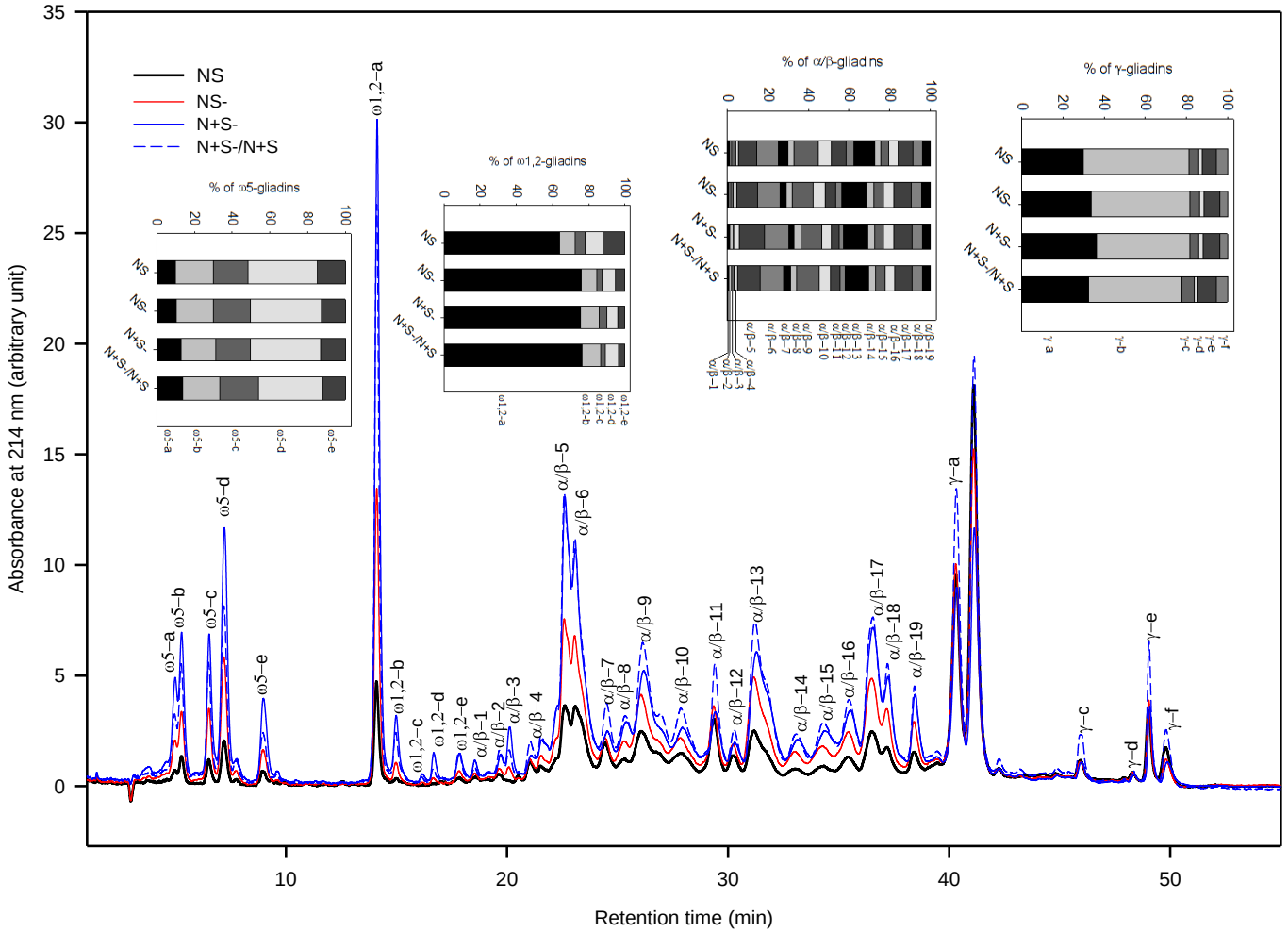
Ta_S22383167	MTD1	7	2	N response	NAD-dependent 5,10-methylenetetrahydrafolate dehydrogenase
Ta_S32586942	AAP	7	2	aa transporter	Amino acid-polyamine transporter
Ta_S13261577	AAP	7		aa transporter	Amino acid transporter
Ta_S22377742	HMT	7	2	aa metabolism:methionine metabolism	homocysteine S-methyltransferase protein [EC 2.1.1.14]
Ta_S17890392	SHMT	7		aa metabolism:serine metabolism	serine hydroxy methyltransferase [EC 2.1.2.1]
Ta_S16058247	APR	7	2	S metabolism	adenosine-5-phosphosulfate reductase [EC 1.8.4.9]
Ta_S50378741	Sultr	7		S metabolism	Sulphate transporter
Ta_S50384131	Sultr	7	2	S metabolism	Sulphate transporter
Ta_S16058102	Sultr4;1	7	2	S metabolism	Sulfate transporter 4;1
Ta_S22373652	Sdi1	7	2	S response	Sulfur deficiency induced1
Ta_S22980914	Sdi2	7	2	S response	Sulfur deficiency induced2
Ta_S26027938	UP9A	7	1	S response	UP-9A , Response to low sulfur
Ta_S17897343	IFR	7	2	Oxidative response:flavonoid synthesis	Isoflavone reductase
Ta_S17986475	Ta_S17986475	7	2	Oxidative response:flavonoid synthesis	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
Ta_S22368319	MSR	7	3	Oxidative response	Methionine sulfoxide reductase
Ta_S26025138	OPR3	7	3	Oxidative response:JA synthesis	12-oxophytodienoate reductase [EC 1.3.1.42]
Ta_S17986414	RNR1	8		DNA replication	ATP binding, Ribonucleoside-diphosphate reductase large chain
Ta_S17988069	MCM6	8	2	DNA replication	DNA replication licensing factor
Ta_S17988603	MCM3	8		DNA replication	DNA replication licensing factor
Ta_S18010943	HMG	8	2	DNA replication:TF	Transcripton factor HMG
Ta_S32518650	RNR1	8		DNA replication	ATP binding, Ribonucleoside-diphosphate reductase large chain
Ta_S32636668	Ta_S32636668	8	2	DNA replication	DNA methylation, DNA replication, DNA replication initiation, biological_process, cell proliferation, histone H3-K9 methylation, regulation of DNA replication, regulation of cell cycle

Ta_S32706356	PRL	8	2	DNA replication	ATP binding, DNA binding, DNA-dependent ATPase nucleoside-triphosphatase nucleotide binding
Ta_S37759259	Ta_S37759259	8	2	DNA replication	DNA methylation, DNA replication, biological_process, cell proliferation, histone H3-K9 methylation
TA60739_4565_HMG	HMG	8	2	DNA repiclation:TF	Transcripton factor HMG
Ta_S12930860	Ta_S12930860	8	2	DNA replication	Small nucleolar ribonucleo protein complex subunit
Ta_S16227636	AGPase	8	2	CHO:starch and sucrose	ADP glucose pyrophosphorylase [2.7.7.27]
Ta_S17893552	RPL18	8	2	Translation	Ribosomal protein L18
Ta_S16221292	Ta_S16221292	9		unknown	Unknown
Ta_S17893255	Ta_S17893255	9	3	Protein trafficking:Golgi organization	saposin-like type B
Ta_S12922866	PDIL	9	2	Protein processing	Disulfide isomerase-like protein
Ta_S12923284	PDIL	9	2	Protein processing	Disulfide isomerase-like protein
Ta_S12923285	PDIL	9	2	Protein processing	Disulfide isomerase-like protein
Ta_S12982481	RPL14	9		Translation	Ribosomal protein L14p/L23e family protein
Ta_S13172511	PFP	9	2	CHO:glycolysis	Pyrophosphate-dependent 6-phosphofructo-1-phosphotransferase [EC2.7.1.90]
Ta_S16631388	RPL3B	9		Translation	ribosomal protein L3
Ta_S17972266	RPL3B	9		Translation	ribosomal protein L3B
Ta_S18009361	PFK	9	2	CHO:glycolysis	Phosphofructokinase [EC 2.7.1.11]
Ta_S24513428	RPL14	9	2	Translation	Ribosomal protein L14
Ta_S32564154	ALT	9	2	aa metabolism	Alanine transaminase [EC:2.6.1.2]
Ta_S37897688	ROC4	9		Protein processing	Cyclophilin-like peptidyl-prolyl cis-trans isomerase
CA656570_AP2_EREBP	CA656570_AP2_EREBP	10		TF	Transcripton factor EREBP
Ta_S12938993	Ta_S12938993	10	3	unknown	Unknown
Ta_S12941371	HCP	10		Cystein-rich	Histidine-containing phosphotransfer protein
Ta_S13046591	PROLM28	10	3	Cysteine-rich	Prolamin precursor

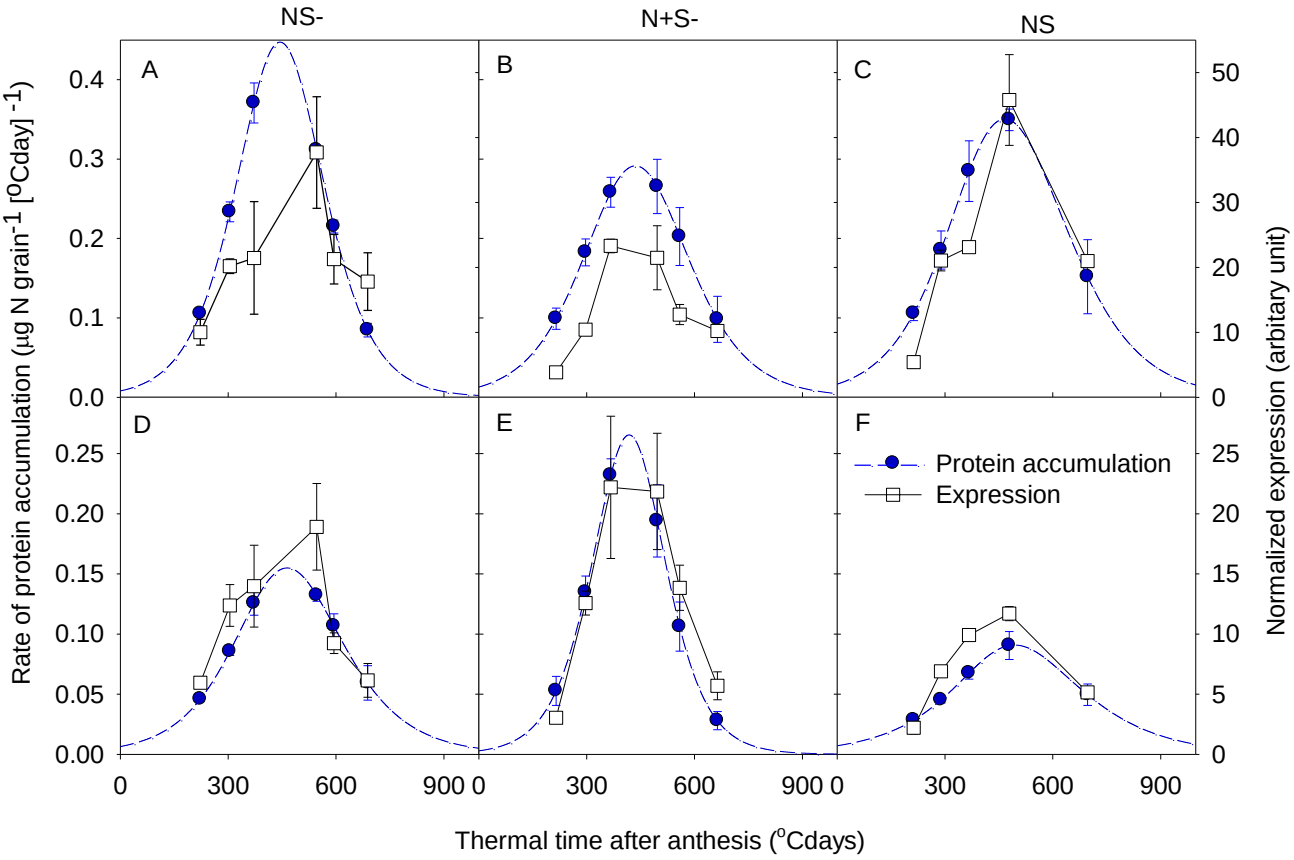
Ta_S16057933	CBL	10		S metabolism	cystathionine beta-lyase [EC 4.4.1.8], in Cysteine and methionine metabolism
Ta_S17985288	CPDW-II	10		Protein processing	Serine carboxypeptidaseII-1 [EC3.4.16.6], Preferential release of a C-terminal arginine or lysine residue
Ta_S32554153	APSK	10	3	S metabolism	adenosine-5-phosphosulfate (APS) kinase [EC:2.7.1.25], catalyzes the phosphorylation of APS to 3'-phospho-APS (PAPS) and competes with APS reductase to partition sulfate between the primary and secondary branches of the sulfur assimilatory pathway
Ta_S32645960	Ta_S32645960	10		unknown	Galactose oxidase/kelch repeat superfamily protein
Ta_S37826290	TPI	10		CHO: glycolysis	Triosephosphate isomerase [EC 5.3.1.1], catalyzes the reversible interconversion of the triose phosphate isomers dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate in the glycolysis pathway.
Ta_S12913225	Ta_S12913225	10		Cysteine-rich	low-molecular-weight cysteine-rich 52



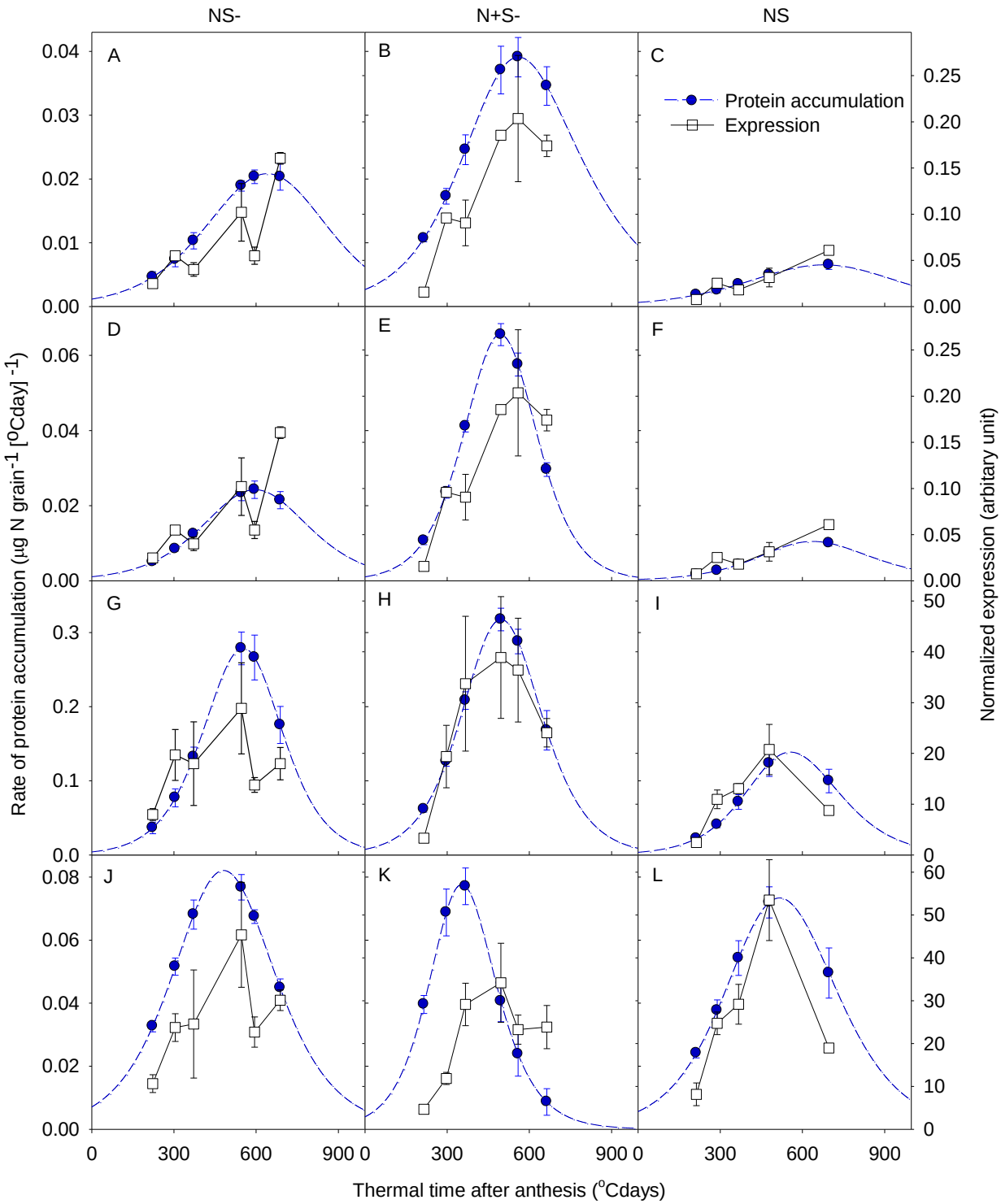
Supplemental Figure 1. *HPLC chromatograms of glutenins and the relative proportion of glutenin fractions (inserted figures) in mature wheat grains grown under four combinations of N and S supply. The four treatments were NS (3 mM N and 2 mM S), NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and N+S-/N+S (shift at mid-grain filling from N+S- to 15 mM N and 2 mM S). The peaks corresponding to high- (Dy12, By9, Dx2, Bx7, and Ax2*) and low-molecular weight (LMW6,7,a,b,c...) glutenin subunits were indicated.*



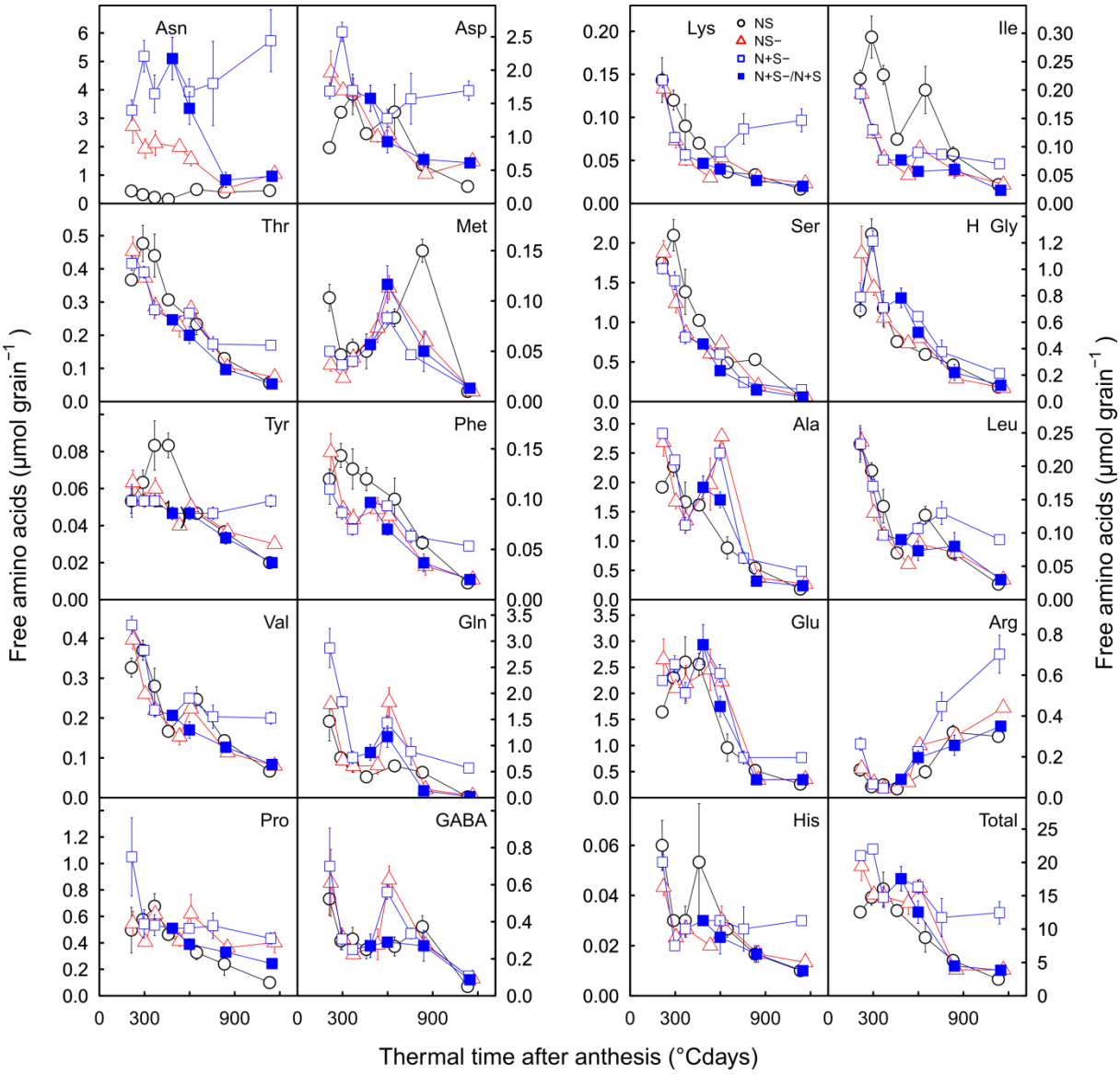
Supplemental Figure 2. *HPLC chromatograms of gliadins and the relative proportion of gliadin fractions (inserted figures) in mature wheat grains grown under four combinations of N and S supply. The four treatments were NS (3 mM N and 2 mM S), NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and N+S-/N+S (shift at mid-grain filling from N+S- to 15 mM N and 2 mM S). The peaks corresponding to ω5-, ω1,2-, α/β-, and γ-gliadins were indicated.*



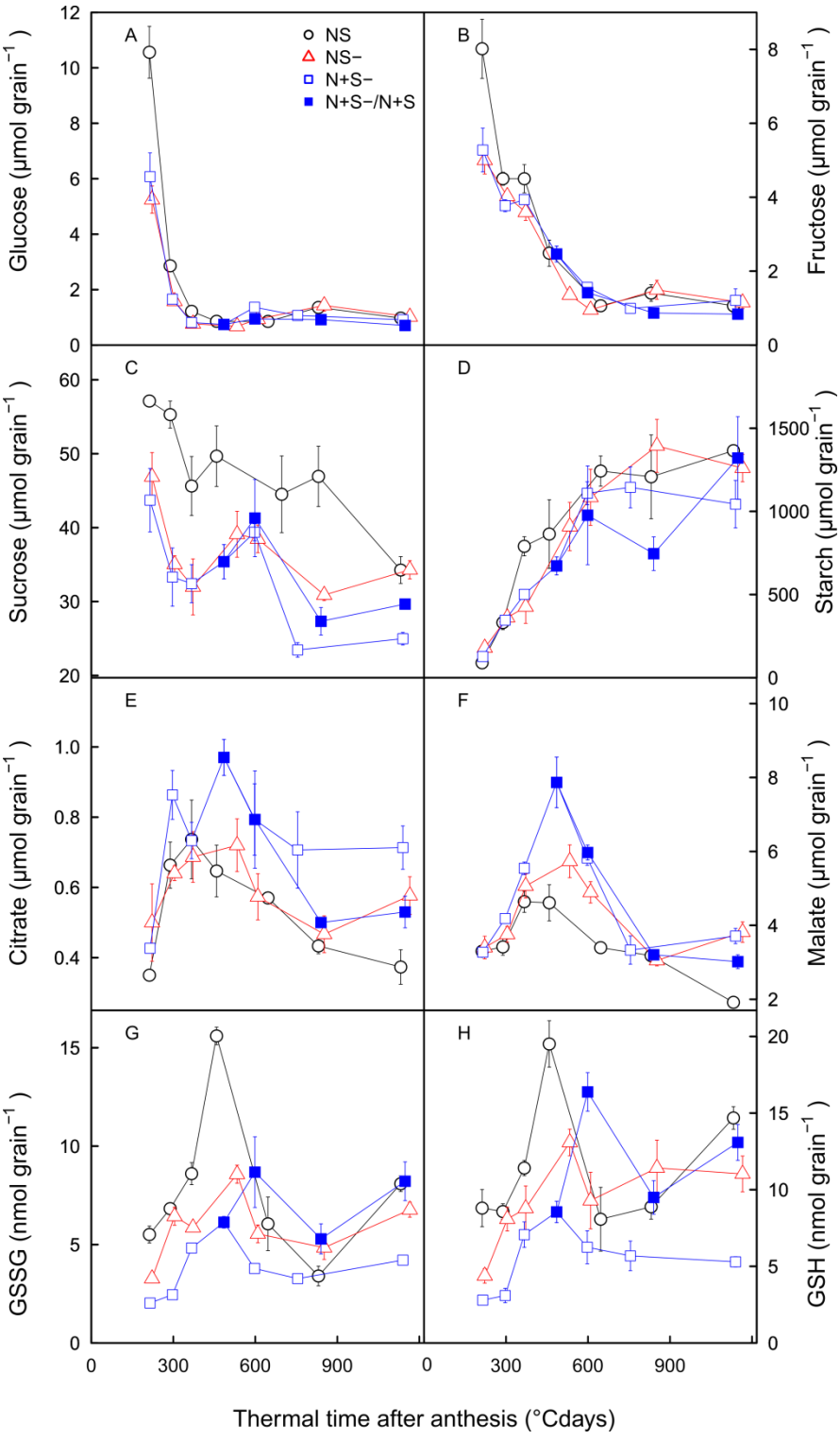
Supplemental Figure 3. *Gene expression and rate of accumulation of LMW-GS (A, B, C) and HMW-GS (D, E, F) versus thermal time after anthesis for wheat grains grown under three combinations of N and S supply. The three treatments were NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and NS (3 mM N and 2 mM S). Data are means ± 1 SE (n = 2 biological replicates for gene expression, and n=3 for the rate of protein accumulation).*



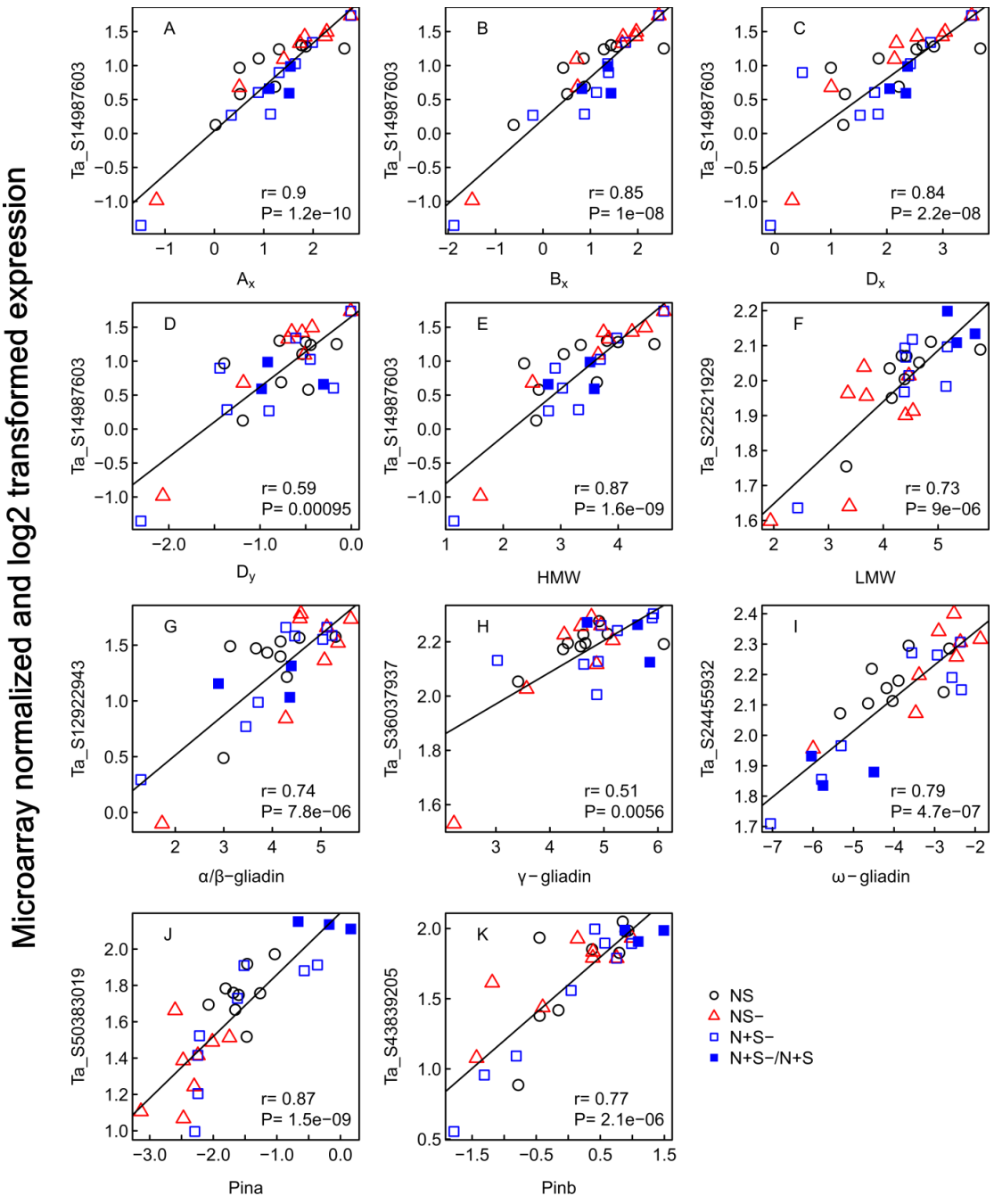
Supplemental Figure 4. *Gene expression and rate of accumulation of gliadin fractions versus thermal time after anthesis for wheat grains grown under three combinations of N and S supply. The three treatments were NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and NS (3 mM N and 2 mM S). Results of ω5-gliadin were presented in the panels A, B, and C, ω1,2-gliadin in D, E, and F; α/ß-gliadin in G, H, and I; γ-gliadin in J, K, and L. Data are means ± 1 SE (n = 2 biological replicates for gene expression, and n=3 for the rate of protein accumulation).*



Supplemental Figure 5. *Mass of free amino acids per grain versus thermal time after anthesis for wheat grains grown under four combinations of N and S supply. The four treatments were NS (3 mM N and 2 mM S), NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and N+S-/N+S (shift at mid-grain filling from N+S- to 15 mM N and 2 mM S). Data are means ± 1 SE for n = 3 biological replicates.*



Supplemental Figure 6. *Mass of primary metabolites versus thermal time after anthesis for wheat grains grown under four combinations of N and S supply. The four treatments were NS (3 mM N and 2 mM S), NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and N+S-/N+S (shift at mid-grain filling from N+S- to 15 mM N and 2 mM S). GSSG and GSH represent glutathione disulfide and glutathione, respectively. Data are means ± 1 SE for n = 3 biological replicates.*



Supplemental Figure 7. *Correlation of gene expressions determined by microarray and qPCR for wheat grains grown under four combinations of N and S supply. The four treatments were NS (3 mM N and 2 mM S), NS- (3 mM N and 0.02 mM S), N+S- (15 mM N and 0.02 mM S), and N+S-/N+S (shift at mid-grain filling from N+S- to 15 mM N and 2 mM S). The Pearson's correlation coefficient r and the P value were also indicated in each figure.*

CONCLUSION PUBLICATION 4

Dans cette partie de l'étude, nous avons constaté que, indépendamment de la phase de développement et des niveaux d'approvisionnement en N et S, il existe une relation allométrique entre la teneur du grain en PR pauvres en S et l'N total du grain. Il existe également une relation allométrique entre la teneur du grain en PR riches en S est et le S total du grain. Ceci suggère que le contrôle de la synthèse des PR repose sur un mécanisme de régulation élaboré et étroitement lié aux teneurs totales du grain en N et S. Il semble que la teneur total en PR du grain en conditions de carence S soit maintenu par un ajustement, promouvant la synthèse des PR pauvres en S au détriment de la synthèse des PR riches en S.

Nous avons tout d'abord montré que les vitesses d'accumulation des différentes fractions de PR dépendent directement du niveau d'expression des gènes correspondant, et ce de manière indépendante des conditions (développementales, ou nutritionnelles). La logique biologique pour des protéines sans turn-over (qui s'accumulent) telles que les PR nous impose de comparer la quantité de transcrits à un instant *t* non pas à la quantité de PR, mais à leur vitesse d'accumulation. Les résultats montrent que la vitesse instantanée d'accumulation des PR est proportionnelle à la quantité de transcrits. La pente de ces relations permet d'estimer l'efficacité de la traduction pour les différentes fractions de PR. Cette efficacité de traduction est comparable pour LMW-GS et HMW-GS, ainsi que pour les α-gliadines, mais est 200 et 300 plus importante pour les ω5-gliadines et les ω1,2-glidaines, respectivement, et 5 fois plus faible pour les γ-gliadines. Ces différences d'efficacité de la traduction pourraient être expliquées par les différences de séquence des ARNm et la disponibilité en certains ARNt (i.e. par l'utilisation des codons), comme cela a été montré pour les PR (zéines) du grain de maïs (Hamilton & Mascarenhas, 1991) ou pour d'autres protéines chez des algues (Campbell & Gowri, 1990).

L'inférence de réseau est ici appliquée dans une approche systémique et nous permet notamment d'observer un lien entre *MYBS3*, un FT impliqué dans le réseau élaboré par Rubio-Somoza *et al.* (2006a, et 2006b) et la vitesse d'accumulation des gliadines α/β. Dans ce même réseau, les FT *FUSCA3* et *MCB1* étaient liés à des protéines riches en cystéine et des inhibiteurs de trypsine/alpha-amylase. *FUSCA3* est un régulateur clé du développement de la graine d'*A. thaliana* (Wang & Perry, 2013) et son orthologue, tout comme *MCB1* font partie

chez l'orge, du réseau élaboré par Rubio-Somoza *et al.* (2006). Ces observations font de ces FT de bons candidats pour des expériences de caractérisation fonctionnelle.

La cinétique de la teneur en méthionine libre n'a pas été altérée par les niveaux d'approvisionnement en N et S. En revanche, deux gènes (*homocystéine S-méthyltransférase* et *cystathionine beta-lyase*) impliqués dans la biosynthèse de la méthionine et de la cystéine sont surexprimés en conditions de carence en S, expliquant l'homéostasie de la teneur en méthionine. Nous suggérons que le métabolisme du grain donne la priorité à l'utilisation de S pour la biosynthèse de méthionine et cystéine afin de maintenir la teneur en PR riches en S, et ce au détriment d'autres molécules contenant du S. Cette hypothèse est étayée par la diminution de glutathion observée en conditions de carence en S. Le réajustement du métabolisme des acides aminés ne se limite pas aux acides aminés riches en S. Nous avons observé que l'asparagine est l'acide aminé dont la biosynthèse est la plus augmentée par la carence en S.

Nos résultats soulignent l'importance des analyses incluant une dimension temporelle afin de mieux comprendre les mécanismes impliqués dans la réponse à la disponibilité en N et S. Nous avons par exemple constaté que les grains ont été capables de maintenir les teneurs en différents acides aminés pendant les stades précoces de développement du grain indépendamment des niveaux d'approvisionnement en N et S et ce jusqu'au milieu de la phase de remplissage du grain. C'est également à ce stade de développement du grain que la plupart des PR ont cessé de s'accumuler dans des conditions de carence en S. Ces résultats évoquent une étape critique dans la réponse du grain à la carence en S, ce qui pourrait aider à optimiser les apports en engrais au champ.

REFERENCES

Campbell, W. H., & Gowri, G. (1990). Codon Usage in Higher Plants , Green Algae , and. *Plant Physiology*, *92*, 1–11.

Hamilton, DA & Mascarenhas, JP (1991). Codon usage tables for zein and non-zein genes: an update. *MNL* **65**:2-3

Rubio-Somoza, I., Martinez, M., Diaz, I., & Carbonero, P. (2006). HvMCB1, a R1MYB transcription factor from barley with antagonistic regulatory functions during seed development and germination. *The Plant Journal*, *45*(1), 17–30. doi:10.1111/j.1365-313X.2005.02596.x

Rubio-Somoza, I., Martinez, M., Abraham, Z., Diaz, I., & Carbonero, P. (2006). Ternary complex formation between HvMYBS3 and other factors involved in transcriptional control in barley seeds. *The Plant Journal*, *47*(2), 269–281. doi:10.1111/j.1365-313X.2006.02777.x

Wang, F., & Perry, S. E. (2013). Identification of direct targets of FUSCA3, a key regulator of Arabidopsis seed development. *Plant Physiology*, *161*(3), 1251–64. doi:10.1104/pp.112.212282

.

DISCUSSION

CONCLUSION

La biologie des systèmes permet aujourd'hui de conceptualiser l'organisation cellulaire (Oltvai et barabasi, 2002), d'en discuter en partant des molécules, en passant par les modules pour finir par l'organisation globale des réseaux de régulation (approche ascendante ou « bottom-up ») ou au contraire, en partant d'un réseau global pour aller jusqu'aux molécules individuelles (approche descendante ou « «top-down »). Quelle que soit l'approche analytique choisie, il faut garder à l'esprit que la structure, la topologie, la robustesse et la fonction du réseau représentant le système sont profondément interdépendantes. La plupart des fonctions cellulaires sont portées par des groupes de molécules organisées en modules fonctionnels. Ces modules ne sont pas isolés les uns des autres et se chevauchent ou communiquent au sein d'un réseau ordonné. La structure organisationnelle du réseau global a été formée au cours de l'évolution et est en grande partie conservée entre les espèces. Le domaine de l'inférence de réseaux et le développement des techniques d'acquisition de données biologiques offrent aujourd'hui la possibilité d'étudier ces réseaux dans leur ensemble. Des résultats récents (Prill *et al*., 201) ont montré que les performances des différentes méthodes varient largement en fonction des espèces et systèmes étudiés. Cela tend à montrer que leur performance dépend plus du contexte de leur implémentation que de leur qualité intrinsèque. Il n'existe aujourd'hui aucune méthode capable d'élucider entièrement et sans erreur le système complexe qu'est la cellule. Les carences et biais des différentes méthodes peuvent être atténués par l'intégration de résultats d 'inférence provenant de plusieurs méthodes complémentaires dans des réseaux « communautaires ».

Les travaux mis en œuvre dans cette thèse rentrent dans le cadre de la biologie des systèmes, qui est surement l'approche théorique la plus adaptée à l'étude de la régulation de la synthèse des PR. L'impact des disponibilités en N et S a été mis en évidence. D'un point de vue méthodologique, ces travaux ont permis de mettre en place des outils adaptés à l'analyse des données. Tout d'abord le développement d'une base de données permet aujourd'hui d'annoter, par homologie, rapidement un grand nombre de gènes de blé. En outre, l'approche d'inférence de réseaux qui a été développée permet de définir un nombre presque infini de sémantiques et d'inférer ainsi des réseaux dans un même cadre global. L'objectif suivant étant donc de définir un ensemble de prédicats répondant aux objectifs biologiques et adaptés aux données dans le contexte d'une étude spécifique.

L'utilisation de ces deux outils a permis d'élargir nos connaissances sur le transcriptome du grain de blé, sur les FT du blé et sur les liens entre les métabolismes en interaction avec la

synthèse des PR. De plus, les outils méthodologiques développés peuvent être mis à profit dans de futures études. Notons également que l'approche d'inférence de réseaux développée au cours de ce travail peut être utilisée dans le cadre d'études très diverses.

Nos travaux ont concrètement permis d'établir un atlas de l'expression des gènes et des FT du blé. Ainsi, la spécificité spatiale de l'expression de certains FT permet d'en faire des candidats sérieux pour la suite des études. Il est admis que les FT sont des protéines d'importance particulière de par leur fonction de régulation de l'expression de leurs gènes cibles. Dans le contexte de nos recherches ces FT sont d'une importance particulière s'ils sont spécifiques du grain de blé et s'ils répondent aux niveaux de disponibilité en N et en S. L'identité des organes de la plante est portée par leur signature transcriptionnelle (Schmid *et al.,* 2005 ; Aceituno *et al.,* 2008 ; Libault *et al.,* 2010) nous avons ici mis en évidence qu'elle était avant tout véhiculée par le transcriptome des FT.

Différentes études chez les céréales ont également mis en avant la reprogrammation transcriptomique qui a lieu entre les grandes phases de développement du grain (Wan *et al*., 2008 ; Drea *et al*., 2005 ; Capron *et al*., 2012). Nous avons ici montré que la reprogrammation qui intervient entre ces phases de développement était également véhiculée par le transcriptome des FT. Nous avons également montré qu'il n'existe pas de bouleversement de cette reprogrammation dans nos gammes d'approvisionnement en N et en S. Nos résultats suggèrent plutôt qu'une délicate régulation des métabolismes du S, des PR et des acides aminés permet de maintenir une homéostasie de la teneur totale du grain en S.

Des études récentes ont montré que l'apport de S au stade de l'émergence de l'épi peut empêcher la carence en S à des stades de développement ultérieurs (Zorb et al, 2009 ; Steinfurth et al, 2012). Dans cette thèse, nous avons montré que l'apport de S à 490°Cj après la floraison (c'est à dire peu de temps avant le stade critique de l'effet de la carence en S que nous avons identifié) peut également atténuer efficacement la carence en S. Ce changement dans l'approvisionnement en S conduit à une récupération rapide du S total, que ce soit au niveau des PR riches en S ou d'autres métabolites importants (glutathion, citrate, et acides aminés libres). De plus, cet effet du S apporté au milieu de la phase d'accumulation s'accompagnait de changements dans l'expression d'un transporteur de S et du métabolisme des acides aminés. Ainsi, les transporteurs de S semblent être de bons candidats pour expliquer la récupération rapide d'un métabolisme du S normal. Il a été mis en évidence que leur expression peut répondre à la disponibilité en S en quelques heures (Smith *et al.*, 1997),

ainsi nos travaux suggèrent que le grain de blé possède un mécanisme efficace de détection du niveau de S et que celui-ci régule finement le transport et le métabolisme du S.

Comme nous pouvions imaginer, la carence en S a un effet important sur le métabolisme du S dans nos études. Trois transporteurs du S et une enzyme de type *adénosine phosphosulfate réductase*, un régulateur central du métabolisme du S, étaient fortement sur-exprimés lors de la carence en S. Cette activation combinée du transport et du métabolisme du S est plutôt bien identifiée dans la littérature (Maruyama-Nakashita *et al.*, 2003 ; Nikiforova *et al.*, 2003). L'activation de ces voies pourrait améliorer l'absorption du S et compenser en partie la carence, faisant de ces gènes de bons candidats pour la suite du travail.

L'effet de la carence en S sur le rapport entre les PR riches et pauvres en S était déjà bien documenté pour le grain de blé mature (Wieser *et al.,* 2004; Zorb *et al.*, 2009), la réponse de la composition en PR à l'approvisionnement en S au cours du développement du grain est en grande partie inconnue. Ici, nous avons constaté qu'indépendamment de la phase de développement et de l'apport en N et en S il existe une relation allométrique entre la teneur en PR pauvres en S et la teneur totale en N du grain. D'un autre côté, il existe une relation allométrique entre la teneur en PR riches en S et la teneur totale en S du grain. Physiologiquement, ceci suggère que le contrôle de la synthèse des PR repose sur un mécanisme de régulation réajustant la composition en PR afin de maintenir la teneur totale en PR du grain en fonction de la disponibilité en N et en S.

Nous avons également constaté des corrélations significatives entre l'expression des gènes codant pour les PR et les vitesses d'accumulation des protéines (à l'exception des γ-gliadines). Ceci suggère que la régulation transcriptionnelle explique au moins en partie, les changements observés dans les vitesses d'accumulation de ces fractions de PR en réponse à la disponibilité en N et en S. La valeur de la pente de la relation pour les ω-gliadines (qui ne contiennent pas de S) était beaucoup plus grande que celle des LMW-GS, HMW-GS, et α/β-gliadines, suggérant une grande différence d'efficacité de traduction qui pourrait être expliquée par la disponibilité en acides aminés soufrés.

Globalement, nous confirmons l'interaction entre les métabolismes du N et du S, et notons que ce sont principalement les changements du rapport N:S qui entrainement les plus grands réarrangements métaboliques, plutôt que leur disponibilité absolue.

PERSPECTIVES

Les variétés de blé cultivées actuellement nécessitent des apports importants d'intrants afin d'atteindre leur potentiel de rendement tout en maintenant une teneur en PR compatible des industries de transformation. Au vu du cout de ces fertilisants (supérieur à 1€/kg), la perte financière est significative, il faut y ajouter la perte d'énergie indispensable à leur production. De même, les retombées environnementales sont importantes et les pays développés s'accordent sur la nécessité d'utiliser tous les leviers possibles afin d'inciter les pays producteurs à diminuer l'utilisation d'intrants. Dans ce contexte, un certain nombre de projets à travers le monde visent à étendre nos connaissances sur les mécanismes moléculaires impliqués dans l'efficacité de l'absorption, de l'assimilation et de l'utilisation du N et du S (Foulkes et al., 2009).

Les résultats du travail de thèse confirment globalement l'intérêt d'utiliser une approche systémique et de prendre en compte l'interaction entre les métabolismes de l'N et du S. Mon travail pourrait être approfondi, notamment grâce aux possibilités qu'offre l'approche d'inférence de réseaux utilisée. En effet un nombre réduit de sémantiques a été utilisé dans ce travail. L'utilisation de nouvelles sémantiques pourrait permettre de mettre en évidence des liens encore non identifiés au sein de nos données.

Des études récentes sur des espèces étroitement liées au blé soulignent le taux relativement élevé de modularité des FT, capables de se lier à un large éventail de promoteurs de gènes (Mitsuda *et al.*, 2010 ; Ou *et al.*, 2011). Ces résultats suggèrent que les FT peuvent fonctionner comme des interrupteurs contrôlant des réseaux de gènes spécifiques de certains organes, certains stades de développement ou certaines conditions. En outre, il semble que le nombre de FT liés au promoteur d'un gène dans certaines conditions soit sensiblement inférieur à ce que l'on pourrait attendre en s'appuyant exclusivement sur nos connaissances des motifs de liaison à l'ADN. Il est donc non seulement important d'étudier les « master-régulateurs » principaux, mais aussi des réseaux plus spécifiques et leur dynamique. Ici, nous avons identifié des FT susceptibles d'être impliquées dans la réponse du grain à la disponibilité en N et S, il serait intéressant de les étudier par des approches plus fonctionnelles et d'identifier à quel niveau des cascades de signalisation ils interviennent.

REFERENCES

Aceituno, F. F., Moseyko, N., Rhee, S. Y., & Gutiérrez, R. a. (2008). The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in Arabidopsis thaliana. BMC Genomics, 9, 438. doi:10.1186/1471-2164-9-438

Capron, D., Mouzeyar, S., Boulaflous, A., Girousse, C., Rustenholz, C., Laugier, C., … Bouzidi, M. F. (2012). Transcriptional profile analysis of E3 ligase and hormone-related genes expressed during wheat grain development. BMC Plant Biology, 12(1), 35. doi:10.1186/1471-2229-12-35

Drea, S., Leader, D. J., Arnold, B. C., Shaw, P., Dolan, L., & Doonan, J. H. (2005). Systematic Spatial Analysis of Gene Expression during Wheat Caryopsis Development. The Plant Cell Online , 17 (8), 2172–2185. doi:10.1105/tpc.105.034058

Foulkes, M. J., Hawkesford, M. J., Barraclough, P. B., Holdsworth, M. J., Kerr, S., Kightley, S., & Shewry, P. R. (2009). Identifying traits to improve the nitrogen economy of wheat: Recent advances and future prospects. Field Crops Research, 114(3), 329–342. doi:10.1016/j.fcr.2009.09.005

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., Franklin, L. D., … Stacey, G. (2010). An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. The Plant Journal : For Cell and Molecular Biology, 63(1), 86–99. doi:10.1111/j.1365-313X.2010.04222.x

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., … Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nat Meth, 9(8), 796–804. Retrieved from http://dx.doi.org/10.1038/nmeth.2016

Maruyama-Nakashita, A., Nakamura, Y., Watanabe-Takahashi, A., Inoue, E., Yamaya, T., & Takahashi, H. (2005). Identification of a novel cis-acting element conferring sulfur deficiency response in Arabidopsis roots. The Plant Journal, 42(3), 305–314. doi:10.1111/j.1365-313X.2005.02363.x

Mitsuda, N., Ikeda, M., Takada, S., Takiguchi, Y., Kondou, Y., Yoshizumi, T., … Ohme-Takagi, M. (2010). Efficient yeast one-/two-hybrid screening using a library composed only of transcription factors in Arabidopsis thaliana. Plant & Cell Physiology, 51(12), 2145–51. doi:10.1093/pcp/pcq161

Nikiforova, V., Freitag, J., Kempa, S., Adamik, M., Hesse, H., & Hoefgen, R. (2003). Transcriptome analysis of sulfur depletion in Arabidopsis thaliana: interlacing of biosynthetic pathways provides response specificity. The Plant Journal, 33(4), 633–650. doi:10.1046/j.1365-313X.2003.01657.x

Oltvai, Z. N., & Barabási, A.-L. (2002). Systems biology. Life's complexity pyramid. Science (New York, N.Y.), 298(5594), 763–4. doi:10.1126/science.1078563

Ou, B., Yin, K.-Q., Liu, S.-N., Yang, Y., Gu, T., Wing Hui, J. M., … Qu, L.-J. (2011). A high-throughput screening system for Arabidopsis transcription factors and its application to Med25-dependent transcriptional regulation. Molecular Plant, 4(3), 546–55. doi:10.1093/mp/ssr002

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., … Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. Nat Genet, 37(5), 501–506. Retrieved from http://dx.doi.org/10.1038/ng1543

Smith, F. W., Hawkesford, M. J., Ealing, P. M., Clarkson, D. T., Vanden Berg, P. J., Belcher, A. R., & Warrilow, A. G. S. (1997). Regulation of expression of a cDNA from barley roots encoding a high affinity sulphate transporter. The Plant Journal, 12(4), 875–884. doi:10.1046/j.1365-313X.1997.12040875.x

Steinfurth, D., Zörb, C., Braukmann, F., & Mühling, K. H. (2012). Time-dependent distribution of sulphur, sulphate and glutathione in wheat tissues and grain as affected by three sulphur fertilization levels and late S fertilization. Journal of Plant Physiology, 169(1), 72–77. doi:http://dx.doi.org/10.1016/j.jplph.2011.08.012

Wan, Y., Poole, R. L., Huttly, A. K., Toscano-Underwood, C., Feeney, K., Welham, S., … Mitchell, R. A. (2008). Transcriptome analysis of grain development in hexaploid wheat. BMC Genomics, 9, 121. doi:10.1186/1471-2164-9-121

Zörb, C., Steinfurth, D., Seling, S., LangenkÄmper, G., Koehler, P., Wieser, H., … Mühling, K. H. (2009). Quantitative Protein Composition and Baking Quality of Winter Wheat as Affected by Late Sulfur Fertilization. Journal of Agricultural and Food Chemistry, 57(9), 3877–3885. doi:10.1021/jf8038988

ANNEXE : PUBLICATION 5

Characterization of non-coding cis-regulatory elements in the promoter region of high-molecular weight glutenin subunit genes for Triticum aestivum L.

Catherine Ravel[1,2], Samuel Fiquet[1,2], Julie Boudet [1,2], Mireille Dardevet[1,2], Jonathan Vincent [1,2,3], Marielle Merlino[1,2], Robin Michard[1,2] and Pierre Martre [1,2]

[1]INRA, UMR1095 Génétique, Diversité et Ecophysiologie des Céréales, 5 chemin de Beaulieu, F-63 039 Clermont-Ferrand, France; [2]Université Blaise Pascal, UMR1095 Diversité et Ecophysiologie des Céréales, F-63 177 Aubière, France; [3]Université Blaise Pascal, UMR6158 CNRS LIMOS Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, F-63 173 Aubière, France.

frontiers in PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 12 November 2014
doi: 10.3389/fpls.2014.00621

Conserved *cis*-regulatory modules in promoters of genes encoding wheat high-molecular-weight glutenin subunits

Catherine Ravel[1,2]*, Samuel Fiquet[1,2], Julie Boudet[1,2], Mireille Dardevet[1,2], Jonathan Vincent[1,2], Marielle Merlino[1,2], Robin Michard[1,2] and Pierre Martre[1,2]

[1] Institut National de la Recherche Agronomique, UMR1095, Genetics, Diversity and Ecophysiology of Cereals, Clermont-Ferrand, France
[2] UMR1095, Genetics, Diversity and Ecophysiology of Cereals, Department of Biology, Blaise Pascal University, Aubière, France

Edited by:
Paolo A. Sabelli, University of Arizona, USA

Reviewed by:
Juan José Ripoll, University of California, San Diego, USA
Nigel G. Halford, Rothamsted Research, UK

***Correspondence:**
Catherine Ravel, Institut National de la Recherche Agronomique, UMR1095, Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu,
F-63 100 Clermont-Ferrand, France
e-mail: catherine.ravel@clermont.inra.fr

The concentration and composition of the gliadin and glutenin seed storage proteins (SSPs) in wheat flour are the most important determinants of its end-use value. In cereals, the synthesis of SSPs is predominantly regulated at the transcriptional level by a complex network involving at least five *cis*-elements in gene promoters. The high-molecular-weight glutenin subunits (HMW-GS) are encoded by two tightly linked genes located on the long arms of group 1 chromosomes. Here, we sequenced and annotated the HMW-GS gene promoters of 22 electrophoretic wheat alleles to identify putative *cis*-regulatory motifs. We focused on 24 motifs known to be involved in SSP gene regulation. Most of them were identified in at least one HMW-GS gene promoter sequence. A common regulatory framework was observed in all the HMW-GS gene promoters, as they shared conserved *cis*-regulatory modules (CCRMs) including all the five motifs known to regulate the transcription of SSP genes. This common regulatory framework comprises a composite box made of the GATA motifs and GCN4-like Motifs (GLMs) and was shown to be functional as the GLMs are able to bind a bZIP transcriptional factor SPA (Storage Protein Activator). In addition to this regulatory framework, each HMW-GS gene promoter had additional motifs organized differently. The promoters of most highly expressed x-type HMW-GS genes contain an additional box predicted to bind R2R3-MYB transcriptional factors. However, the differences in annotation between promoter alleles could not be related to their level of expression. In summary, we identified a common modular organization of HMW-GS gene promoters but the lack of correlation between the *cis*-motifs of each HMW-GS gene promoter and their level of expression suggests that other *cis*-elements or other mechanisms regulate HMW-GS gene expression.

Keywords: *cis*-elements, conserved *cis*-regulatory modules (CCRMs), high-molecular-weight glutenin subunits (HMW-GS), transcriptional regulation, seed storage proteins (SSPs), transcription factors (TFs), wheat (*Triticum aestivum* L)

INTRODUCTION

Wheat is one of the three most economically important crops in the world with maize and rice, with a global annual production of about 700 Mt in 2012 (FAOSTAT; http://faostat.fao.org/). Wheat is a broad term for crops including tetraploid species ($2n = 28$) like durum wheat (*Triticum turgidum* spp. *durum*) and hexaploid species ($2n = 42$) like bread wheat (*T. aestivum* spp. *aestivum*). Wheat is one of the most important sources of carbohydrates and vegetable proteins in human diets as it accounts for about 20% of all calories and proteins consumed. It is mostly transformed before it is consumed, and each type of transformation depends on the unique visco-elastic properties of gluten, a network formed by water and seed storage proteins (SSPs). It is mainly the SSPs that determine the technological quality of wheat flour (for instance, see reviews by Shewry et al., 2002 and Shewry, 2009). Prolamins, the major component of wheat SSPs, comprise monomeric gliadins and polymeric glutenins. The latters have both low- (LMW-GS) and high- (HMW-GS) molecular-weight subunits. Glutenins account for 30–50% of the total SSP content of grain, with HMW-GS alone representing up to 12% of the total. Glutenins strongly influence dough elasticity (Payne et al., 1987; Shewry et al., 2002), with HMW-GS more so than LMW-GS (Branlard and Dardevet, 1985; Gupta and MacRitchie, 1994; He et al., 2005).

As glutenins are so important for technological quality, the genes coding for HMW-GS have been extensively studied. The genome of the hexaploid bread wheat is divided into three subgenomes (called A, B, and D) forming three homoeologous groups. HMW-GS are encoded by the three loci *Glu-A1*, *-B1* and *-D1* located on the long arms of the group 1 chromosomes. As confirmed by the sequencing of these three regions (Gu et al., 2006), each locus consists of two closely linked paralogous genes, *Glu-1-1* and *Glu-1-2*, that encode x-type and y-type HMW-GS, respectively. Thus, bread wheat HMW-GS form a small multigene family of six genes with two orthologous sets of *Glu-1-1* and *Glu-1-2* genes (Allaby et al., 1999). HMW-GS genes are highly

polymorphic (e.g., Payne and Lawrence, 1983). These six genes are not always all expressed. *Glu-A1-2* is silent so from three to five HMW-GS genes are usually expressed in grain. A duplication of *Glu-B1-1* is observed in lines with the overexpressed Bx7 HMW-GS giving an additional expressed gene (Ragupathy et al., 2008). SSPs are specifically expressed in the endosperm and all HMW-GS have similar patterns of expression and represent 60–65% of the total RNA from the endosperm between 10 and 30 days post anthesis (Shewry et al., 2009).

SSP synthesis is primarily controlled both spatially and temporally at the transcriptional level. Transcription factors (TFs) bind specifically to short conserved DNA sequences (5–15 nucleotides) called *cis*-regulatory elements or *cis*-elements, which are usually located in the proximal promoter of genes and characterized by a consensus motif. In barley (*Hordeum vulgare*), the regulatory mechanisms of SSP genes have been extensively studied by transient expression experiments using an hordein promoter (Mena et al., 1998; Vicente-Carbajosa et al., 1998; Oñate et al., 1999; Diaz et al., 2002, 2005; Isabel-La Moneda et al., 2003; Rubio-Somoza et al., 2006a,b; Moreno-Risueno et al., 2008) and have been described as a network of *cis*-elements and their interacting TFs (Rubio-Somoza et al., 2006a). This network is conserved in other cereals as reviewed by Verdier and Thompson (2008) and Xi and Zheng (2011). It consists of five *cis*-elements able to recognize eight TFs belonging to four families (bZIP of the Opaque-2 family, and the B3, DOF, and MYB proteins), which are all reported to be activators of SSP genes. More precisely, the GCN4 like-motif (GLM, 5′-ATGAG/CTCAT-3′) and the prolamin box (P-box, or PB, 5′-TGTAAAG-3′), also called the endosperm motif, constitute the bipartite endosperm box, which plays a key role in activating the expression of prolamin genes as also shown in wheat (Hammond-Kosack et al., 1993). GLM is recognized by bZIP TFs, like BLZ1 and BLZ2 in barley (Vicente-Carbajosa et al., 1998; Oñate et al., 1999) or SPA (Storage Protein Activator) in wheat (Albani et al., 1997), while the P-box is bound by PBF and SAD, both DOF-type TFs (Vicente-Carbajosa et al., 1997; Mena et al., 1998; Diaz et al., 2005). Two additional *cis*-elements, 5′-AACA/TA-3′ and 5′-TATC/GATA-3′ core sequences, are able to bind R2R3-MYB (notably GAMYB) and R1MYB (MCB1 and MYBS3) TFs, respectively (Diaz et al., 2002; Rubio-Somoza et al., 2006a,b). The last *cis*-regulatory sequence is the RY repeat (5′-CATGCATG-3′), which binds FUSCA3, a B3 protein (Bäumlein et al., 1992; Moreno-Risueno et al., 2008). In addition to these DNA-protein interactions, protein-protein interactions consolidate the formation of larger complexes that regulate SSP expression (Rubio-Somoza et al., 2006b).

Wheat promoters of α-gliadin classes (Van Herpen et al., 2008), LMW-GS (Hammond-Kosack et al., 1993; Conlan et al., 1999), and HMW-GS (Norre et al., 2002) have been functionally analyzed. Van Herpen et al. (2008) reported differences in regulatory-elements between promoter sequences of α-gliadin genes from A and B genomes. The LMW-GS promoter studied is characterized by a tandem repeat of two endosperm motifs known as the long endosperm box that is important for controlling endosperm-specific expression (Hammond-Kosack et al., 1993). Thomas and Flavell (1990) and Norre et al. (2002) analyzed extensively the promoters of *Glu-D1* by transient expression assay in tobacco and maize. A 38-bp enhancer element has been identified (Thomas and Flavell, 1990). In addition, the promoter of *Glu-D1-1* contains an atypical endosperm box where the P-box is associated with a G-like box of the ACGT family able to bind bZIP proteins (Norre et al., 2002). Moreover, these authors suggested that the enhancer element may act with the G-like box to increase reporter gene expression.

The exponential growth of genomic sequence databases, and the development of specialized databases of *cis*-acting elements in plants (Higo et al., 1999; Rombauts et al., 1999), coupled with the development of bioinformatics tools to discover specific motifs in DNA or protein sequences (e.g., MEME; Bailey et al., 2006), greatly facilitate the *in silico* analysis of promoters. However, the discovery of *cis*-regulatory elements is hindered by the variability within their sequences, which typically tolerate nucleotide substitutions without a loss of functionality. There are ways of taking this variability into account when predicting the presence of *cis*-regulatory elements (Stormo, 2000). Another aspect to consider is that, in higher eukaryotes, TFs often regulate gene expression by binding DNA in cooperation with other regulatory proteins. As reviewed by Armone and Davidson (1997), separate *cis*-elements of a given promoter often interact with different parts of an overall regulatory complex. This type of organization of *cis*-elements in a region of up to a few 100 bases in the vicinity of the gene being regulated is called a *cis*-regulatory module (CRM), where the relative positions of *cis*-elements and the distances between them are crucial.

Recently, the LMW-GS and HMW-GS gene promoters have been analyzed *in silico* (Juhász et al., 2011; Makai et al., 2013). The *cis*-acting elements present in published sequences of LMW-GS genes, mainly ESTs, were computationally retrieved and differences in the numbers and combinations of specific sequences were highlighted allowing the identification of conserved non-coding sequence regions (CRMs). Models for the transcriptional regulation of LMW-GS genes were then proposed (Juhász et al., 2011). The promoter profiles of HMW-GS genes are highly conserved in the Triticeae family despite differences between paralogous genes (Makai et al., 2013). Here the aim was to understand in more detail the transcriptional regulation of HMW-GS genes through a comparative promoter analysis. The promoters of the main alleles at each HMW-GS gene were analyzed *in silico* for the predicted presence of *cis*-regulatory elements. The organization of these elements within orthologous (homoeologous) and paralogous copies was compared. This work shows the presence of conserved CRMs (CCRMs). In addition, the HMW-GS gene promoters were sequenced in a set of wheat lines to determine whether their sequence variability correlates with the organization of *cis*-elements and hence the expression levels of these genes. A functional analysis of conserved regions consisting of *cis*-motifs potentially able to bind bZIP TFs was carried out by using transient expression and electrophoretic mobility shift assays (EMSA).

MATERIALS AND METHODS

DIVERSITY ANALYSIS

Forty-two lines representative of the genetic diversity (Haseneyer et al., 2008; Ravel et al., 2009) and of the main electrophoretic

alleles of HMW-GS of the INRA worldwide hexaploid wheat (*Triticum aestivum* L.) core collection (Balfourier et al., 2007) were analyzed (**Table 1**). Genomic DNA was extracted from leaves as described in Ravel et al. (2009) and used for PCR amplification of the proximal promoter of HMW-GS genes. Fragments of approximately 700–1100 nucleotides were obtained (Supplementary Table 1) and sequenced. We did not amplify *Glu-A1-2* genomic DNA as it was silent in all the 42 lines. Diversity indices including nucleotide diversity (π), number of segregating sites (θ), number of haplotypes (H), haplotype diversity (Hd), and Tajima's *D*-test of neutral evolution were calculated for each sequence with SNiPlay (Dereeper et al., 2011).

EXPRESSION ANALYSIS

To quantify HMW-GS gene expression, RNA was extracted from developing grains harvested at 400°C days after anthesis from 13 lines representing the main promoter alleles (**Table 1**). Lines were cultivated in the greenhouse as described in Ravel et al. (2009). For each of the four lines 964, 1288, 2135, and 4874, four independent biological replicates were obtained. Two independent biological replicates were used for each of the nine remaining accessions. Quantitative-real-time PCR (qRT-PCR) was performed as described in Ravel et al. (2009) using a LightCycler® 480 II sequence detection system and the LightCycler 480 SYBR Green I Master (Roche) according to the manufacturer's instructions. Primer pairs used for qRT-PCR and their amplification efficiency are given in Supplementary Table 2. The specificity of each primer pairs was confirmed by a single peak in the real-time melting temperature curves for each gene.

Amplification plots and predicted threshold cycle values were obtained with LightCycler 480 SW 1.5 software (Roche). Genes coding for glyceraldehyde 3-phosphate dehydrogenase (GAPDH), elongation factor 1 alpha (eF1α), β-tubulin, and 18S RNA were used as internal controls to normalize expression results (Ravel et al., 2009). The geometric mean of control gene expression was calculated so that HMW-GS gene expression could be quantified and normalized also taking into account the efficiency of each primer pair.

PROMOTER ANNOTATIONS

Twenty motifs known to participate in the regulation of SSP and two light responsive motifs were selected from the PLACE *cis*-motif database, which contains 469 entries (**Table 2**; Higo et al., 1999). We included a light responsive (Abox) and a circadian rhythm-responsive (CIACADIANLELHC) motif because diurnal fluctuations in carbohydrate pools and Opaque 2 (O2) binding activity during seed filling may impact SSP synthesis (Ciceri et al., 1997, 1999; Carman and Bishop, 2004). We also added two additional motifs, 5′-AACNNA-3′ and 5′-TATAWA-3′, which were not in the PLACE database. The first motif is able to bind a MYB protein from rice (*Oriza sativa*) belonging to the GAMYB sub-family (Takaiwa et al., 1996). The second motif is the TATA-variant sequence of SSP genes involved in the formation of a transcription initiation complex (Fauteux and Strömvik, 2009; Bernard et al., 2010).

Both strands of the 1-kb region upstream of the start codon for the six HMW-GS genes from cv. Renan retrieved from public databases (DQ537335.1, DQ537336.1, and DQ537337.1 for *Glu-A1*, *Glu-B1*, and *Glu-D1*, respectively; Gu et al., 2006) and the promoter sequences of the five (i.e., all but *Glu-A1-2*) HMW-GS genes obtained in this study for 42 lines (including cv. Renan) of the INRA worldwide hexaploid wheat core collection were annotated using a custom-made PERL program (named PlantPAD) that extracts the name, sequence and coordinates of the motifs and produces a graphical representation of the query sequence on which the starting position of each *cis*-motif is plotted. Based on the assumption that functional *cis*-motifs are conserved among HMW-GS genes, we used PlantPAD to search for co-occurrence of *cis*-motifs in these genes. To build the consensus, the program considers each motif and its coordinates (the position of its first nucleotide relative to the start codon). Any motif that appears at the same coordinates ($\pm$5 bp) in all the sequences being annotated is considered to be conserved. As insertion-deletion events (indels) within a sequence cause motifs to shift along the gene, the program also recognizes conserved motifs which appear in all the sequences with the same coordinates plus or minus the shift size (the length of indels). The consensus is then plotted and the distances between conserved motifs corresponding to those found in more than 50% of the sequences are analyzed. Such a consensus is designed to highlight the conserved regulatory regions. This approach was used to analyze separately both sets of orthologous genes and produce a consensus plot for each of them. These consensuses were then used to generate an overall consensus annotation of HMW-GS gene promoters.

FUNCTIONAL VALIDATION

Particle bombardment was performed in developing wheat endosperm to validate *cis*-motifs potentially able to bind bZIP TFs. The promoter of *Glu-B1-1* gene (hereafter termed PrBx7) was amplified and cloned using the primers from cv. Renan given in Supplementary Table 1. A total of 747-bp upstream fragment of the start codon was used. In addition, to assess the role of the distal conserved regulatory regions of this promoter, the 597-bp fragment from the start codon (hereafter termed tPrBx7) was synthesized.

All constructs used for transient expression assay were obtained using Gateway technology (Invitrogen). Three entry clones were used (pDONRP4-P1R, pDONR221, and pDONRP2R-P3). pDONRP4-P1R contained the rice actin promoter, PrBx7 or tPrBx7, while pDONR221 and pDONRP2R-P3 contained a reporter gene (either GUS or GFP, respectively) and the 3′-terminator nopaline synthase gene (3′-NOS). Three expression pDESTR4-R3-based vectors (pAct-GFP, pPrBx7-GUS, and ptPrBx7-GUS) were created. A transient promoter activation assay based on co-bombardment with pPrBx7-GUS or ptPrBx7-GUS and pAct-GFP constructs was performed using immature endosperm from cv. Récital collected at 230°C day after anthesis from plant grown in the greenhouse under optimal growth conditions. Seeds were surface-sterilized and endosperms were carefully isolated. Endosperms were cultured on Murashige and Skoog medium supplemented with maltose (100 g L^{-1}) for 2–3 h for subsequent bombardment. Gold particles (0.6 μm in diameter; Bio-Rad) were prepared with 500 ng of a 1:1 molar ratio mixture of pAct-GFP and pPrBx7-GUS or ptPrBx7-GUS.

Table 1 | Country of origin, protein coding alleles, and haplotypes of the promoters of five HMW-GS genes for 42 accessions of the INRA worldwide hexaploid wheat core collection.

Name[a]	Country of origin[b]	*Glu-A1-1*		*Glu-B1-1*		*Glu-B1-2*		*Glu-D1-1*		*Glu-D1-2*	
		Protein[c]	Promoter[d]	Protein[c]	Promoter[d]	Protein[c]	Promoter[d]	Protein[c]	Promoter[d]	Protein[c]	Promoter[d]
A4 (748)	AFG	1	h1	7	h1	8	h1	3	h1	12	h1
Aifeng NO 4 (822)	CHN	1	h1	7	h1	HZ	h1	HZ	h2	HZ	h1
ARCHE (964)	FRA	null	h2	6	h2	8	h2	2	h2	12	h1
AURORE (1110)	AUS	2*	h3	7	h1	9	h1	2	h2	12	h1
BALKAN (1192)	YUG	2*	h3	7	h1	9	h1	5	h3	10	h2
BARBU DU FINISTERE (1323)	FRA	null	h2	20	h1	20	h3	2	h2	12	h1
BELLIEI 590 (1288)	HUN	2*	h3	7	h1	9	h1	5	h3	10	h2
CHINESE SPRING (2135)	CHN	null	h2	7	h1	8	h1	2	h2	12	h1
CHORTANDINKA (2153)	RUS	null	h1	HZ	h1	HZ	h1	HZ	h3	HZ	h2
CHYAMTANG (2171)	NPL	null	h2	7	h1	8	h1	2	h2	12	h1
COPPADRA (2330)	TUR	2*	h3	7	h1	8	h1	3	h2	12	h1
COTIPORA (2353)	BRA	2*	h3	N	h1	N	h3	2	h2	12	h1
COURTOT (2358)	FRA	2*	h3	7	h1	8	h1	2	h2	12	h1
DI7202-103 (2526)	FRA	1	h1	7	h1	8	h1	5	h3	10	h2
GLENLEA (3358)	CAN	2*	h3	7OE	h1	8	h1	5	h3	10	h2
GODOLLOI 15 (3366)	HUN	null	h2	N	h1	N	h4	5	h3	10	h2
JO3045 (3942)	FIN	2*	h3	7	h1	9	ND	2	h2	12	h1
M708//G25/N163 (4482)	ISR	2*	h4	HZ	h1	HZ	h1	2	h2	12	h1
MARS DE SUEDE ROUGE BARBU (4645)	FRA	null	h2	7	h1	8	h1	2	h2	12	h1
MISKAAGANI (4874)	LBN	2*	h3	N	h1	N	h1	2	h2	12	h1
MOCHO DE ESPIGA BRANCA (4901)	PRT	2*	h3	13	h3	16	h1	2	h2	12	h1
N46 (5088)	ISR	null	h2	7	h1	8	h1	2	h2	12	h1
NABU EPI BLANC (5102)	NPL	null	h2	7	h1	8	h1	2	h2	12	ND
NANKING 25 (5116)	CHN	null	h2	7	h1	8	h1	2	h2	12	h1
NEPAL 84 (5166)	NPL	null	h2	7	h1	8	h1	2	h2	12	h1
NP120 (5308)	IND	null	h2	17	h1	18	h1	2	h2	12	h1
NYU BAY (5399)	JPN	null	h2	7	h1	8	h4	2	h2	12	h1
OPAL (5486)	DEU	1	h1	7	h1	9	h1	5	h3	10	h2
PITIC 62 (5748)	MEX	1	h1	7	h1	8	h1	2	h2	12	h1
RECITAL (6027)	FRA	2*	h3	6	h2	8	h2	5	h3	10	h2
RENAN (6086)	FRA	2*	h3	7	h1	8	h1	5	h3	10	h2
SEU SEUN 27 (6529)	KOR	null	h2	7	h4	8	h1	4	h2	12	h1
RALET (8048)	FRA	null	h5	20	h1	20	h3	2	h2	12	h1
ZANDA (8058)	BEL	1	h2	20	h1	20	h3	2	h2	12	h1
HOPEA (9048)	FIN	1	N	6	h2	8	h2	2	h2	12	h1
FRUH-WEIZEN (13310)	DEU	null	h2	22	h5	22	h5	5	h3	10	h2
ORNICAR (13471)	FRA	2*	h3	6	h2	8	h2	5	h3	10	h2
TALDOR (13476)	FRA	2*	h3	7	h1	8	h1	4	h2	12	h1
APACHE (13481)	FRA	null	h2	7	h1	8	h1	3	h2	12	h1
OPATA 85 (13811)	MEX	2*	h3	7	h1	9	h1	5	h3	10	h2
SYNTHETIQUE-W7984 (13812)	MEX	null	h6	7	h1	8	h1	N	h4	N	ND
BLE DE REDON BLANC 1/2 LACHE 1 1 (15658)	FRA	1	h1	13	h3	16	h1	2	h5	12	h1

Accessions used for expression studies are shown in bold.

[a] *Accession no. in the INRA Triticeae Genetic Resources Collection (http://www6.clermont.inra.fr/umr1095) is given in brackets.*

[b] *Country names are given as three-letter ISO codes (http://www.unc.edu/~rowlett/units/codes/country.htm).*

[c] *Protein coding allele for the x- or y-type HMW-GS identified by SDS-PAGE. HZ, heterozygous.*

[d] *Haplotype of the promoter for HMW-GS genes. ND, no data.*

Table 2 | Characteristics of *cis*-motifs from PLACE database and bibliographic references used to annotate the promoters of HMW-GS genes.

Names	Name used in PLACE	Sequence	Binding transcription factor[a]	References
DOF				
DOF core	DOFCOREZM	AAAG	**DOF** SAD, *PBF*, BPBF WPBF	Hammond-Kosack et al., 1993; Yanagisawa and Schmidt, 1999
Pbox1	PROLAMINBOXOSGLUB1	TGCAAAG		Wu et al., 2000; Isabel-La Moneda et al., 2003
Pbox2	300CORE	TGTAAAG		Thomas and Flavell, 1990
Pbox3	−300ELEMENT	TGHAAARK		Marzábal et al., 1998
bZIP				
GLM 1	GCN4OSGLUB1	TGAGTCA	**bZIP** *O2*	Albani et al., 1997; Vicente-Carbajosa et al., 1998; Wu et al., 2000
GLM 2	−300MOTIFZMZEIN	RTGAGTCAT	BLZ1, BLZ2	Thomas and Flavell, 1990; Oñate et al., 1999
GLM 3	GLMHVCHORD	RTGASTCAT	SPA	Norre et al., 2002
ACGT core	ACGTATERD1	ACGT		
G-box motif 1	ABREATCONSENSUS	YACGTGGC		Kang et al., 2002; Choi et al., 2000
G-box motif 2	ABRELATERDI	ACGTG		
CAAT	CAATBOX1	CAAT	**bZIP**[b]	Shirsat et al., 1989
RY-REPEAT				
RY_core	RYREPEATLEGUMINBOX	CATGCAY	**AB3/VP1** FUSCA3	Fujiwara and Beachy, 1994; Moreno-Risueno et al., 2008; Van Herpen et al., 2008
AACA MYB				
AACA motif 1	AACACOREOSGLUB1	AACAAAC	**R2R3-MYB** GaMYB GaMYB	Takaiwa et al., 1996; Suzuki et al., 1998; Wu et al., 2000; Diaz et al., 2002
AACA motif 2	ANAERO1CONSENSUS	AAACAAA		
MYB1AT	MYB1AT	WAACCA		
AACA motif 3	Not referred in PLACE	AACNNA		
GATA MYB				
GATA box 1	GATABOX	GATA	**R1MYB** MCB1 MYBS3	Rubio-Somoza et al., 2006a
GATA box 2	MYBST1	GGATA	**MYB**	Baranowskij et al., 1994
OTHERS				
E box	EBOXBNNAPA	CANNTG	**bHLH**	Chaudhary and Skinner, 1999
CCAAT	CCAATBOX1	CCAAT	**HAP** *CBF*	Albani and Robert, 1995
ESP	ESPASGL01	ACATGTCATCATGT	Not identified	Vickers et al., 2006
TATA-variant	Not refered in PLACE	TATAWA	TATA-box-binding proteins	Fauteux and Strömvik, 2009
CIACADIANLELHC	CIACADIANLELHC	CAANNNNATC		Piechulla et al., 1998
ABox	PALBOXAPC	CCGTCC		

[a] *Transcription factor families are indicated in bold followed by the name of corresponding transcription factor in maize (italics), barley, wheat (underlined), or other species (italics and underlined).*

[b] *Interaction not functionally validated.*

Bombardments were conducted at a distance of 6 cm from the stopping plate using a biolistic helium gun device (PDS-1000, Bio-Rad) with a pressure of 6.21 MPa. Following bombardment, endosperms were incubated for 2 days in the dark at 24°C in a Murashige and Skoog medium supplemented with 3% (w/v) sucrose and 0.15 mM of each of the 20 proteinogenic amino acids. For GUS expression, endosperms were stained with 5-bromo-4-chloro-3-indolyl glucuronide according to Jefferson et al. (1987). Endosperms were observed using a MZ16 F stereomicroscope equipped with a DFC300 FX digital camera (Leica Microsystems) and GUS and GFP activities were determined by counting the number of blue and green cells, respectively. Expression results were normalized by dividing the number GUS foci by the number of GFP foci. For each construct, 10 independent bombardments of eight endosperms each were performed. The pAct-GFP construct was used to determine the efficiency of bombardment as proposed by Eini et al. (2013).

The DNA-binding activity of *cis*-motifs with SPA was studies by EMSA. The SPA protein was expressed in *E. coli* (BL21 AI strain) by cloning *Spa* cDNA into the pDEST17 plasmid vector

(Invitrogen) producing pHis-SPA. *Spa* expression was induced with 0.2% (w/v) arabinose for 3 h. Proteins extracts were obtained after re-suspension of the induced cells in a 10 mM Tris buffer (pH 8) containing 6 M urea and 100 mM NaH_2PO_4 (10 mL g^{-1} pellet). Purification of the recombinant protein was achieved by loading protein extracts onto a Ni^{2+}-NTA resin and bound proteins were eluted in a 10 mM Tris buffer (pH 4.5) containing 6 M urea and 100 mM NaH_2PO_4. The eluate was dialyzed against a 10 mM Tris buffer (pH 8.3) containing 2 M urea, 100 mM NaH_2PO_4, 100 mM KCl, 0.02% Tween-20, 10% glycerol, and 0.5 mM phenylmethylsulfonyl fluoride (PMSF) for 36 h to renaturate the recombinant protein and then against a 10 mM Tris buffer (pH 7.5) containing 50 mM KCl, 1 mM dithiothreitol, 0.02% Tween™ 20, 10% glycerol, and 0.5 mM PMSF for 16 h. The dialysate was then concentrated with an Amicon 10 kDa filter (Millipore).

DNA oligonucleotides able to bind bZIP TFs (GLM and G-box) used in EMSA are described in Supplementary Table 3. Each single-strand oligonucleotide was labeled using the Biotin 3′ End DNA Labeling Kit (Pierce) following the manufacturer's instructions and hybridized for 30 min at the annealing temperature of the probes. The labeled dsDNA probe (20 fmol) was incubated with 560 ng to 4 µg of recombinant His-SPA protein in 20 µL of a binding buffer containing 10 mM Tris (pH 7.5), 2 mM dithiothreitol, 100 mM KCl, 10% glycerol, 0.05% nonyl phenoxypolyethoxylethanol, 2 mM ethylenediaminetetraacetic acid, 100 ng μL^{-1} poly(dI.dC), 250 ng μL^{-1} fish sperm DNA, 0.5 mM PMSF for 30 min at room temperature. DNA-protein complexes were analyzed by non-denaturing 6% polyacrylamide gel electrophoresis in a 45 mM Tris, 45 mM Borate, and 1 mM ethylenediaminetetraacetic acid buffer (pH 8.3). After separation (100 V, 1 h at 4°C), gels were electroblotted to nylon membranes using the same buffer (380 mA, 45 min at 4°C). The biotin end-labeled DNA was detected using streptavidin, horseradish peroxidase conjugate following the manufacturer's instructions (LightShift Chemiluminescent EMSA kit, Pierce).

STATISTICAL ANALYSES

All statistical analyses were done using R 3.0 software (R Core Team, 2013). The normality of and homogeneity of variances of expression data were tested by the Shapiro–Wilk and Bartlett's tests, respectively. Depending on the results of previous analyses, expression data were submitted to non-parametric or parametric variance analysis with the Kruskal–Wallis or the general linear model procedure. Multiple comparison tests between groups after Kruskal–Wallis tests were done with the Kruskalmc function while the Student–Newman–Keuls test was used to compare means after the general linear model procedure. The Kruskal–Wallis and Student–Newman–Keuls tests used were those available in the R "agricolae" (version 1.1-8) package (De Mendiburu, 2014), all other tests were done using the R "Stats" (version 2.15.3) package. All the data were used in a first analysis based on a model with one factor (gene). In a second step, analyses were carried out gene per gene to study the promoter haplotype factor.

To analyze the differences in expression of HMW-GS genes and haplotypes one-way ANOVA were performed. First, an ANOVA with the gene as the main factor was carried out. The four lines with the null allele at *Glu-A1-1* and the line with protein allele 7 overexpressed (7OE) at *Glu-B1-1* were excluded from this analysis to avoid bias. Secondly, ANOVAs with the promoter haplotype as the main factor were performed for each gene (including the null allele at *Glu-A1-1*).

Differences in normalized expression from transient expression assays were analyzed using *t*-test. All statistically significant differences were judge at 5%.

RESULTS

THE VARIABILITY OF THE PROMOTER IS NOT SYSTEMATICALLY CONNECTED WITH PHENOTYPIC VARIABILITY

The variability in the nucleotide sequence of the promoters of the five HMW-GS genes was extensively studied by sequencing a set of 42 lines representative of the diversity present in the INRA worldwide hexaploid wheat core collection. The following results deal with the noncoding DNA region upstream of the start codon given that for HMW-GS genes the transcription start site (TSS) is about 60 bases upstream of the start codon for translation. In some cases, the hybridization sites of reverse primers were downstream of the start codons, so the sizes of the upstream fragments studied ranged from 467 to 1138 bp. A total of 36 single-base changes, 2 single-base insertion-deletions (indels) and 1 larger indel were identified in an average of 3858 bp promoter sequence per line (**Table 3**, Supplementary Table 4). These specific regions have an average of one polymorphism every 100 bases. The number of polymorphisms varied between promoters. *Glu-B1-2* promoter has one polymorphism every 58 bp, threefold more frequently than the *Glu-D1-2* promoter, which has one polymorphism every 145 bp. One large deletion of 54 bp spanning from 291 to 344 upstream of the start codon in the *Glu-B1-1* promoter was observed in two lines (accession nos. 4901 and 15658). Thus, nucleotide diversity estimated by the mean pairwise difference (π) varied from one promoter to another, ranging from 1.5×10^{-3} for *Glu-D1-1* to 3.0×10^{-3} for *Glu-B1-1*. Except for *Glu-D1-2*, we observed that the nucleotide diversity (π) and the number of segregating sites (θ) are about equal in values as confirmed by the non-significant Tajima's D statistic (**Table 3**). This suggests that there has been no particular pattern of selection in these regions.

The polymorphisms are linked by a high level of linkage disequilibrium (data not shown). Therefore, for all loci, most of the lines clustered into two main haplotypes with the remaining haplotypes being generally represented by single lines. Notably, the number of haplotypes found for each promoter fits the number of protein coding alleles for *Glu-D1-2* only (**Table 1**, **Figure 1**). For *Glu-B1-1*, we observed more protein coding alleles than promoter haplotypes. For the three other *Glu1* genes, we observed more promoter haplotypes than protein coding alleles. Each electrophoretic allele, except for *Glu-B1-2* alleles, tends to have a more-frequent promoter haplotype (**Figure 1**).

THE VARIABILITY OF THE HMW-GS GENE PROMOTER IS OFTEN CONNECTED WITH THE LEVEL OF GENE TRANSCRIPTION

To assess whether the gene transcrition level is influenced by the promoter haplotype of each HMW-GS gene, HMW-GS

Table 3 | Number of electrophoretic alleles revealed by SDS-PAGE, haplotype and diversity statistics for the promoters of five HMW-GS genes from 42 accessions of the INRA worldwide hexaploid wheat core collection.

Genes	No. of electrophoretic alleles[a]	Promoter length (bp)	No. of polymorphic sites/No. of indels[b]	No. of haplotypes[c]	Haplotype diversity, Hd	Nucleotide diversity, π	No. of segregating sites, θ	Tajima's statistic
Glu-A1-1	3	878	7 (4)/0	6 (3)	0.71	1.51×10^{-3}	1.60×10^{-3}	NS[d]
Glu-B1-1	7	747	10 (3)/1 (0) *54*	5 (2)	0.34	2.67×10^{-3}	3.16×10^{-3}	NS
Glu-B1-2	6	465	7 (2)/1 (0) *1*	5 (1)	0.45	3.03×10^{-3}	3.03×10^{-3}	NS
Glu-D1-1	4	667	5 (3)/0	5 (2)	0.49	1.48×10^{-3}	1.52×10^{-3}	NS
Glu-D1-2	2	1163	7 (0)/1 (0) *1*	2 (0)	0.43	2.93×10^{-3}	1.39×10^{-3}	$P < 0.05$

[a] *Protein coding allele for the x- or y-type HMW-GS identified by SDS-PAGE.*

[b] *The number of singletons (i.e., a polymorphism found in a single line) is given in brackets; the size of indels is indicated in italics.*

[c] *The number of haplotypes including a single line is indicated in brackets.*

[d] *NS, not significant.*

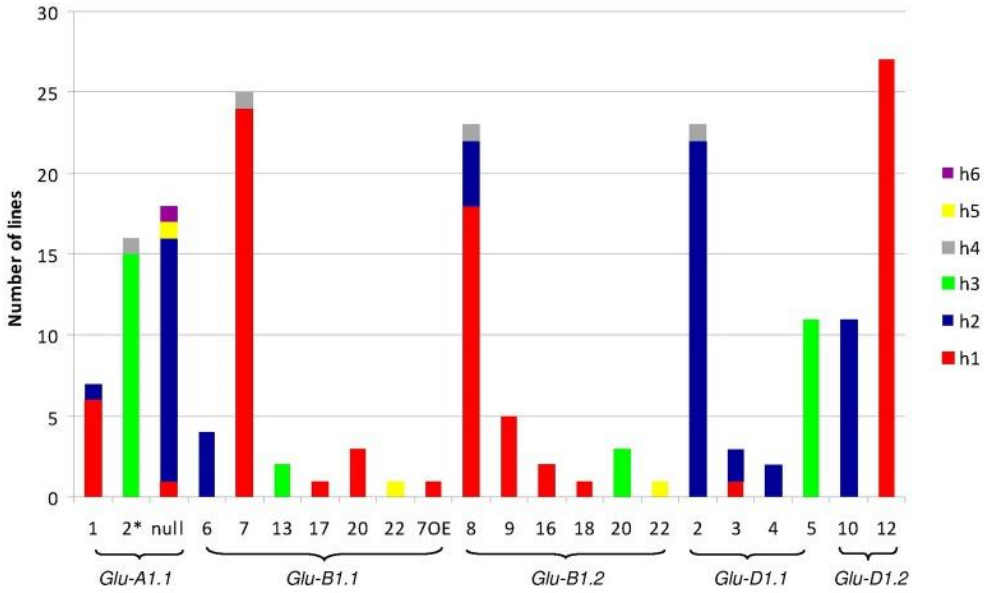


FIGURE 1 | Number of lines of each haplotype of HMW-GS gene promoter for all electrophoretic forms of HMW-GS present in the set of 42 lines studied. The promoter haplotypes are named h1 to h6. The same color is used for the same haplotype number of a given gene promoter. Although the color is identical for all h1 haplotypes, their sequences differ (e.g., the sequences of haplotype h1 at *Glu-A1-1* and *-B1-1* are different).

transcripts were quantified at 400°C days after anthesis for 13 lines by qRT-PCR (**Table 1**). The five HMW-GS genes had different levels of transcription ($P = 2 \times 10^{-16}$). On average, *Glu-B1-1* and *Glu-D1-1* showed a higher level of transcription than the remaining genes, while *Glu-D1-2* was expressed at lower levels (**Table 4**). The two x-type HMW-GS genes were expressed up to 10-fold higher than the genes coding the y-type. The transcription of *Glu-A1-1* was intermediate.

Among the four accessions with the null allele at *Glu-A1-1*, three harbor the h2 promoter haplotype and one the h5 haplotype. These two haplotypes differ by only one single nucleotide polymorphism (SNP) and their transcription was close to zero (**Table 5**). The transcription of the two other promoter haplotypes for *Glu-A1-1* were not different ($P = 0.95$). One line (accession no. 8058) harbors the h2 haplotype but has the protein allele 1 and had a transcription close to that of the h1 and h3 promoter haplotypes. For *Glu-B1-1*, once the line with the Bx7OE protein allele was discarded, the promoter haplotype effect was significant ($P = 0.014$). The transcription for the h1, h3, and h4 haplotypes was similar and, on average, 2.6-fold higher than that for haplotype h2 (**Table 5**), which only includes the Bx6 protein allele (**Figure 1**). The line with the Bx7OE protein allele has the h1 promoter haplotype, as most of the BX7 protein alleles, but it expressed *Glu-B1-1* at a level (195.33 ± 29.25, $n = 2$) twice that of Bx7 lines. For *Glu-B1-2*, the haplotype effect was significant ($P = 0.023$) and transcription from h1 was higher than from h3 (**Table 5**). For this gene, the promoter haplotypes were not linked with separate protein alleles (**Figure 1**). The RNA expression of the *Glu-D1-1* and *Glu-D1-2* alleles was not influenced by their promoter haplotypes (data not shown).

Table 4 | Comparison of the transcription levels of HMW-GS genes at 400°C days after anthesis for 13 lines of INRA worldwide hexaploid wheat core collection.

Genes	No. of lines[a]	RNA expression levels[b]
Glu-A1-1	9 (22)	56.70 ± 3.99 (B)
Glu-B1-1	12 (32)	83.64 ± 7.65 (A)
Glu-B1-2	13 (34)	17.14 ± 0.97 (C)
Glu-D1-1	13 (34)	83.69 ± 5.02 (A)
Glu-D1-2	13 (34)	8.063 ± 0.42 (D)

Data are means ± 1 SE.

[a] The number of data points is indicated in brackets.

[b] Different letters in brackets indicate a significant difference (α = 5%) calculated according to a Kruskal–Wallis non-parametric test followed by the Kruskal multiple comparisons test.

Table 5 | Multiple comparison of the mean levels of RNA expression from promoter alleles of HMW-GS genes at 400°C days after anthesis.

Genes	Promoter haplotype	No. of lines[a]	RNA expression levels[b]
Glu-A1-1	h1	1 (2)	58.25 ± 1.83 (B)
	h2 (null)[d]	3 (10)	1.40 ± 0.18 (A)
	h2 (1)[d]	1 (2)	57.38 ± 7.70 (B)
	h3	6 (16)	58.56 ± 5.19 (B)
	h5	1 (2)	1.41 ± 0.32 (A)
Glu-B1-1[c]	h1	7 (20)	102.27 ± 9.08 (A)
	h2	3 (8)	35.36 ± 3.43 (B)
	h3	1 (2)	87.23 ± 7.35 (A)
	h4	1 (2)	86.94 ± 11.50 (A)
Glu-B1-2	h1	8 (22)	18.96 ± 1.24 (A)
	h2	3 (8)	14.74 ± 1.28 (AB)
	h3	2 (4)	11.89 ± 1.45 (B)

Data are means ± 1 SE.

[a] The number of data points is indicated in brackets.

[b] Different letters in brackets indicate a significant difference (α = 5%) calculated according to a Kruskal–Wallis non-parametric test followed by the Kruskal multiple comparisons test.

[c] The line accession no. 3358 with the 7OE allele was discarded.

[d] For Glu-A1-1 haplotype 2, results for the null and 1 protein alleles (indicated in brackets) were treated as two different haplotypes in the ANOVA.

These results highlight different RNA expression levels for different HMW-GS genes and, for three HMW-GS genes, the effects of the promoter haplotype. Thus, differences in the regulation of these genes might stem from the organization of the *cis*-motifs in their promoters.

COMMON *cis*-MOTIFS ORGANIZATION OF HMW-GS GENE PROMOTERS

To analyze the organization of *cis*-motifs in HMW-GS gene promoter, we first searched for similar patterns in the 1-kb promoter region of the six HMW-GS genes of cv. Renan, as HMW-GS genes have similar expression patterns during development and in response to environmental factors. We then compared the consensus organization of *cis*-motifs found for cv. Renan with that found for the haplotypes of each gene to relate differences in *cis*-motifs organization with differences in gene expression.

In all six HMW-GS gene promoters of cv. Renan we found all the 24 *cis*-motifs we annotated but the Pbox2 and ESP motifs. Most of these motifs were annotated several times and a total of 44 (for *Glu-B1-2*) to 54 (for *Glu-D1-2*) *cis*-motifs per gene were annotated. All the *cis*-motifs able to bind all TFs known to regulate the expression of SSP genes were present, but the typical bipartite endosperm box was not found. The number of *cis*-motifs found was over-estimated as the sequences of a few motifs (**Table 2**) were nested within some others. Most of the nested *cis*-motifs bind TFs of the same family (**Table 2**). Therefore, we took into account only the longest motif where nested motifs were predicted, which reduced the number of *cis*-motifs per gene by 15–24%. Motifs able to bind MYB TFs (GAMYB, MCB1, MYBS3) were predominant, with 9–14 *cis*-motifs per gene, followed by motifs able to bind bZIP TFs, with 9–13 *cis*-motifs per gene, and DOF TFs (PBF, SAD), with 4–8 *cis*-motifs per gene. The CAAT *cis*-motif accounted for about two-thirds of the total number of *cis*-motifs able to bind bZIP TFs (**Table 6**).

The organization of orthologous promoters from cv. Renan showed few differences (**Figures 2A,B**) on the plus strand. For x-type HMW-GS genes, the organization was well conserved between 0 and −400 (nucleotide position relative to the start site). The TATA-box was at −90. A few differences were detected like an AACA motif at −144 in *Glu-A1-1* and *-D1-1*, which was absent in the orthologous B sequence. Between −400 and −1000, the organization was also well conserved but a 55-bp insertion in the *Glu-B1-1* promoter shifted the *cis*-motif upstream (i.e., at a more negative nucleotide position) of the insertion. Interestingly, we discovered a composite box named the GLM-GATA box. This box includes two repeated units, each of them made of a GATA motif and a GLM separated by a third GGATA motif. The relative positions of the constitutive *cis*-motifs in this box were conserved among the three orthologous sequences of cv. Renan (**Figure 2**). An ACGT motif was present a few bases upstream of this box in the B and D sequences. About 50 and 200 nucleotides upstream of this box, a DOF core motif (AAG) and an AACA motif (able to bind R2R3-MYB TFs), respectively, were detected in all the homoeologs. Downstream of this box, we found an AACA motif able to bind R2R3-MYB and the RY repeat.

Similar observations were made for the y-type sequences (**Figures 2B,C**). *Cis*-motifs organization presented many similarities between positions 0 and −400, although the promoter of *Glu-B1-2* includes some additional motifs at about position −150. In addition, the entire composite GLM-GATA box was lacking in the promoters of *Glu-A1-2* and *-B1-2*, the latter containing only a single copy of the GLM. None of these three sequences included the ACGT motif near the GLM-GATA found in the x-type HMW-GS gene promoters. We observed a composite motif at position −400, which was conserved in these three homoeologous sequences, composed of a G-box and three consecutive MYB motifs (two GATA and one AACA motifs). At about position −400, a deletion shortened the distances between the motifs at −400 and the adjacent ones on the *Glu-A1-2* promoter causing

Table 6 | Number of motifs in the upstream 1000-bp region of the six HMW-GS genes from the hexaploid wheat cv. Renan.

Motif sequence	Binding transcription factor	*Glu-A1-1*	*Glu-B1-1*	*Glu-D1-1*	*Glu-A1-2*	*Glu-B1-2*	*Glu-D1-2*
DOF							
AAAG	DOF	6	4	4	7	4	5
TGCAAAG	DOF		1	1	1		1
TGHAAARK	DOF	1		2			1
bZIP							
ACGT[a]	bZIP		1	1	2	3	2
ACGTG[a]	bZIP	1					
RTGAGTCAT[b]	bZIP				1	1	
TGAGTCA[b]	bZIP	2	2	2			2
YACGTGGC[b]	bZIP		1	1	1	1	1
CAAT	bZIP[c]	7	9	7	6	6	7
AACA MYB							
AAACAAA	R2R3-MYB	2	1	2	1	1	1
WAACCA	GAMYB			1	2	2	1
AACNNA	R2R3-MYB	1	1	1	1	1	1
GATA MYB							
GGATA	MYB/R1MYB	8	7	9	7	8	7
RY REPEAT							
CATGCA	AB3/VP1	2	2	2	3	1	2
OTHERS							
CANNTG	bHLH	4	4	4	7	5	4
CCAAT	HAP	2	3	2	2	2	2
TATAWA	TATA-box-Binding Proteins	1	1	1	1	1	1
CAANNNNATC		1	2	2	2	2	2
CCGTCC		1	1		1		1
Total		39	40	42	45	38	41

[a] *Related to GLM.*

[b] *Related to G-box motifs.*

[c] *Interaction not functionally validated.*

a deletion of a few motifs. For the three y-type homoeologous genes, an RY repeat and an AACA motif (binding R2R3-MYB) were located between position −400 and the GLM-GATA box.

The overall consensus generated from all HMW-GS genes of cv. Renan (**Figure 2C**) consisted of 21 motifs including motifs able to bind all the TFs known to regulate SSP synthesis so far. They were organized into five CCRMs. CCRMs were numbered from 1 to 5 from the start codon and composed of two to five *cis*-elements. As expected, CCRM1, a few nucleotides upstream of the TSS, was composed of the TATA-box variant and the CAAT motif. CCRM2 included a G-box-like motif and a CAAT motif, nested into an E-box (CANNTG), while CCRM3 clustered two GATA boxes. CCRM4 was the most interesting module. It included the incomplete GLM-GATA box, an AACA motif and the RY repeat. The GLM-GATA box was incomplete because of a missing GLM in the cv. Renan allele at *Glu-B1-2*. The fifth module, CCRM5, has a DOF motif and a CAAT box nested into an E-box and is located between positions −900 and −1000 in all promoters. A few bases downstream of CCRM5, E-boxes and circadian motifs were conserved. No typical bipartite endosperm box was detected. On the minus strand, we noted an over-representation of the DOF core AAAG motif (data not shown).

For each HMW-GS gene, except *Glu-B1-1*, the annotation of haplotypes was almost identical (**Figures 3**, **4**). Three groups were observed for *Glu-B1-1*. Haplotypes h2 and h5 have identical annotations, but compared to the other haplotypes, they contain an additional RY repeat at position −160. The second group contains h1 and h4, which are distinct from h3 because of an indel. Distances between motifs upstream and downstream of position −400 are therefore shorter in h3 than in the other haplotypes. In addition, a bZIP motif present in the insertion is deleted in h3. The haplotype h3 of *Glu-D1-1* promoter differs from other haplotypes as it has two additional bZIP motifs, one being a G-box.

The relative position of the GLM-box was conserved in all haplotypes of the three orthologous sequences of the x-type HMW-GS genes (**Figure 3**) and the y-type *Glu-D1-2* gene (**Figure 4**). For *Glu-B1-2*, the region sequenced in this study did not cover the GLM-GATA box (**Figure 4**), but the analysis of *Glu-B1-2* promoter sequences of cv. Chinese Spring (KC20630) and Xiaoyan

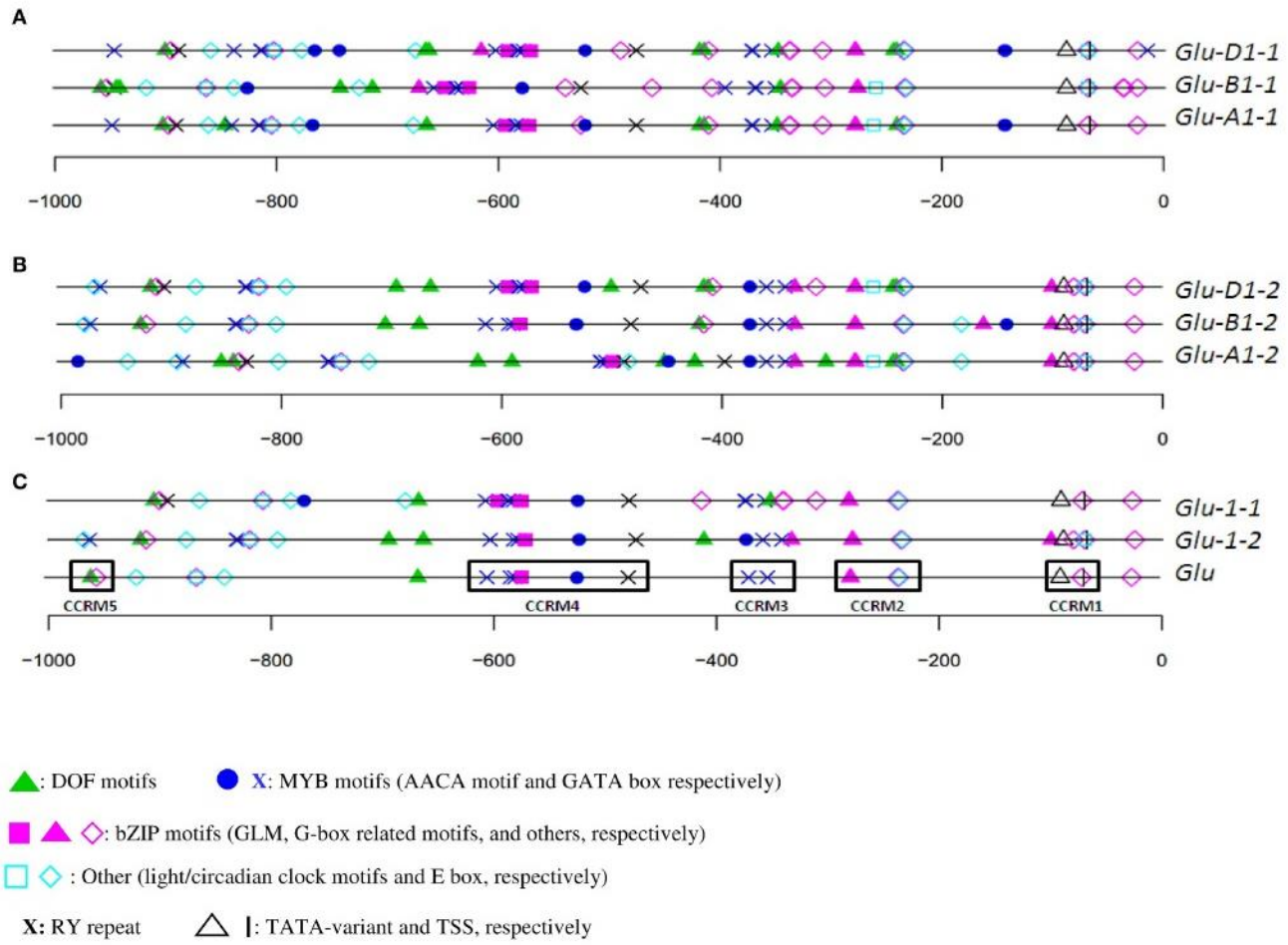


FIGURE 2 | *In silico* annotation of HMW-GS promoters of cv. Renan. Positions are indicated relatively to the start site. Sequences were obtained from public databases. **(A)** x-type *Glu-A1-1*, *-B1-1*, and *-D1-1* homoeologs; **(B)** y-types *Glu-A1-2*, *-B1-2*, and *-D1-2* homoeologs; and **(C)** consensus annotations of all orthologous set of sequences (*Glu-1-1* and *Glu-1-2* for the x- and y-type HMW-GS promoters, respectively), and of paralogous sequences (*Glu*) with its five conserved *cis*-regulatory modules CCRM1 to CCRM5.

54 (EU137874), available in public databases, shows that, in these cases, the relative position of the GLM-box is also conserved in this gene (data not shown).

THE GLM-GATA BOX IS INVOLVED IN THE REGULATION OF *Glu-B1-1* EXPRESSION

To investigate the involvement of the GLM-GATA box in the regulation of HMW-GS gene expression, we analyzed the effect of the 5′ deletion from positions −747 to −597 (fragment carrying the GLM-GATA box) by transient expression experiment (**Figures 5**, **6**). The deletion of the GLM-GATA box reduced normalized GUS expression by 59%.

To verify the potential binding activity of the two GLMs (GLM1 and GLM2 at positions −647 and −626, respectively) present in the GLM-GATA box of the *Glu-B1-1* gene promoter, we performed EMSAs with synthetic oligonucleotides and a recombinant SPA protein expressed as a His fusion in *E. coli* (**Figure 7**). We also determined the *in vitro* binding of SPA to the G-box motif, which was previously shown to bind bZIP proteins (Norre et al., 2002). As shown in **Figure 7A**, arabinose treatment induced expression of a protein of 50–75 kDa that was not present in uninduced cell extracts. The apparent size of the recombinant protein determined by SDS-PAGE was larger than the expected 48 kDa molecular mass of the His-tagged SPA fusion protein. A similar apparent increase in size on SDS gels was already reported by Albani et al. (1997) in their study of SPA. The recombinant His-SPA protein was purified to near homogeneity and used for binding assays. A DNA-protein complex was clearly observed with the GLM2 motif, while the shifted band detected for the GLM1 and the G-box was considerably fainter (**Figure 7B**). No shifted band was observed when incubation was carried out with the mutated probes (*glm1*, *glm2*, and *G-box*). DNA-binding affinity of the recombinant protein seems to be greater with the GLM2 probe than the other probes tested.

DISCUSSION

Here we characterized and annotated wheat HMW-GS gene promoters. The expression of these genes in developing grain was

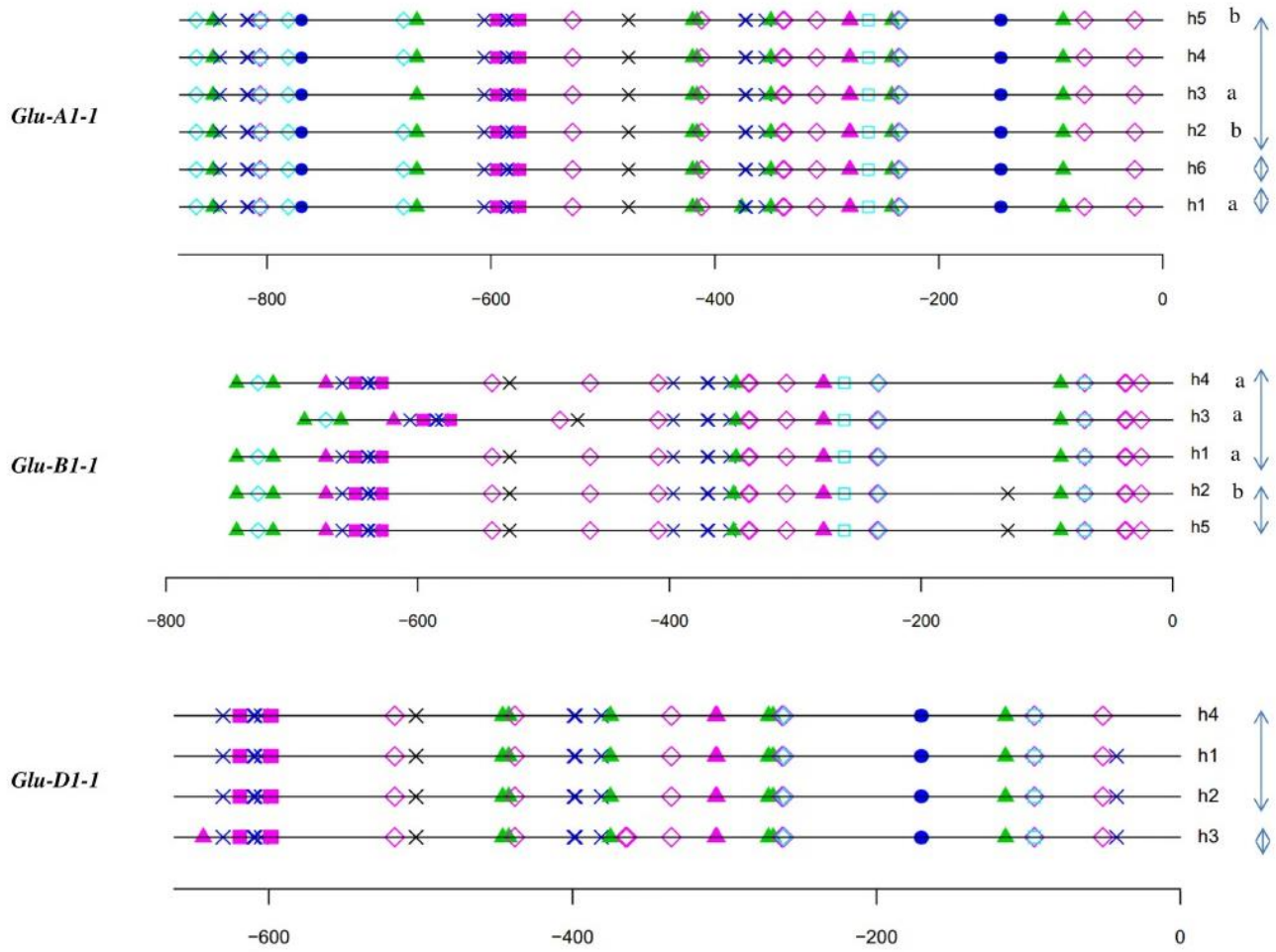


FIGURE 3 | *In silico* annotation of the x-type HMW-GS gene promoter haplotypes. Positions are indicated relative to the transcription start site. Sequences were obtained from a set of 42 lines representative of the genetic diversity of the INRA worldwide hexaploid wheat core collection. For each gene, the haplotype of the promoter is indicated by the letter h followed by the number of the haplotype. Letters a and b indicate the significantly different groups for the mean of expression for haplotypes studied by qRT-PCR. Clusters of haplotypes differing by one polymorphism are shown with gray arrows on the right. See the key to **Figure 2** for descriptions of *cis*-motif symbols.

quantified by qRT-PCR and the correlations between the variability in expression and the variability in predicted *cis*-element motifs of the corresponding promoter were also analyzed. We considered regions of 467–1138 bp upstream of the start codon. In *Arabidopsis thaliana*, based on the density of polymorphisms in gene upstream regions, functional promoters require 250–500 nucleotides upstream of the TSS (Korkuć et al., 2014). Under the assumption that promoter length is conserved, the lengths of the regions surveyed here provide a reasonable coverage of functional SSP gene promoters in wheat. Moreover, we analyzed the role of the GLM-GATA box of the *Glu-B1-1* gene promoter by transient expression assay and evaluated the functionality of the *cis*-motifs reported to bind bZIP TFs.

VARIABILITY OF HMW-GS PROMOTER HAPLOTYPES CANNOT BE USED DIRECTLY TO SCREEN FOR ELECTROPHORETIC ALLELES

In *A. thaliana*, the nucleotide variability in promoters varies depending on the function of their downstream gene (Korkuć et al., 2014). It is higher for genes involved in adaptive processes and transcriptional regulation than for genes involved in housekeeping functions. In wheat, the diversity of promoters is not widely documented so far. The range of nucleotide diversity observed for HMW-GS promoters, approximately one polymorphism every 100 bases, is comparable to that reported for the SPA promoter (Ravel et al., 2009), but is higher than the overall level of polymorphism of one SNP every 212 nucleotides reported for promoters of other genes (Ravel et al., 2006). Although upstream gene regions are somewhat constrained as they are involved in gene regulation, they are reported to show higher variability than coding regions. Constraints most likely apply to *cis*-regulatory elements (Korkuć et al., 2014). As they affect short regions, mutations could occur with little or no incidence, whereas the entire coding sequence has to withstand greater constraints. In addition, the modular organization of *cis*-elements, together with their redundancy, may buffer the effects of mutations (reviewed by Purugganan, 2000). These reasons probably explain why the

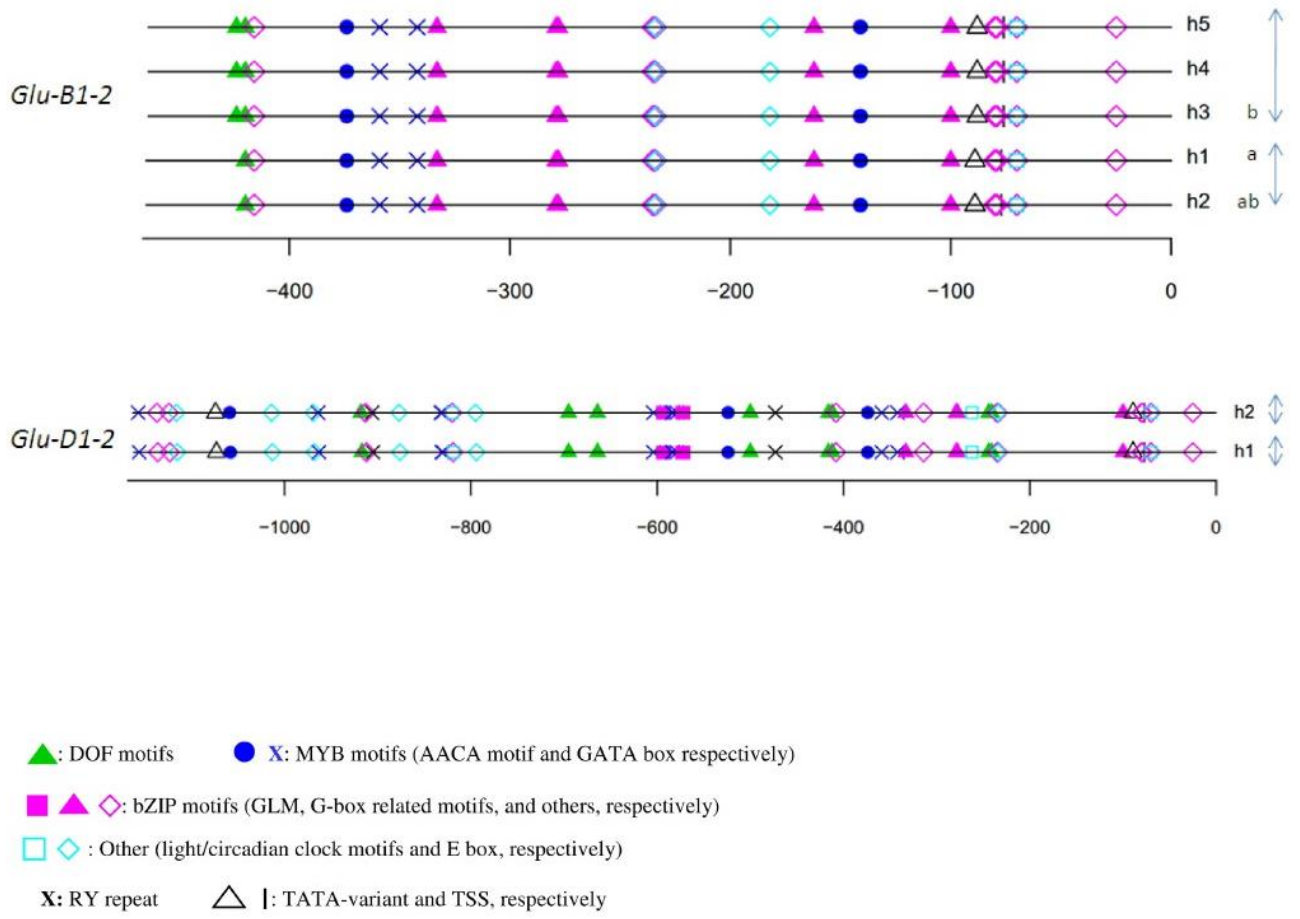


FIGURE 4 | *In silico* annotation of the γ-type HMW-GS gene promoter haplotypes. Positions are indicated relative to the transcription start site. Sequences were obtained from a set of 42 lines representative of the genetic diversity of the INRA worldwide hexaploid wheat core collection. For each gene, the haplotype of the promoter is indicated by the letter h followed by the number of the haplotype. Letters a and b indicate the significantly different groups for the mean of expression for haplotypes studied by qRT-PCR. Clusters of haplotypes differing by one polymorphism are shown with gray arrows on the right.

diversity is higher in promoter regions than in coding sequences. As usually reported (e.g., Chao et al., 2009), the level of diversity was the lowest in HMW-GS sequences from the D genome with 1 polymorphism every 145 base for *Glu-D1-2*, whereas the highest level of diversity was observed for HMW-GS promoters from the B genome with, on average, one polymorphism every 60 bases.

SDS-PAGE is still routinely used for characterization of HMW-GS alleles. Developing diagnostic SNPs to identify electrophoretic forms of HMW-GS from any part of young plants would be a valuable tool to support breeding for improved flour quality. However, there are up to four haplotypes promoter sequences per electrophoretic allele or only one haplotype for several alleles. Anderson et al. (1998) already reported two different alleles for the Bx7 promoter. The promoter haplotypes perfectly match the protein alleles only for *Glu-D1-2*. Currently, the identification of a set of SNPs from the other HMW-GS promoter sequences as a shortcut to distinguish between different protein forms is not possible, so the search for diagnostic SNPs needs to continue.

A MINIMAL FRAMEWORK FOR THE TRANSCRIPTIONAL REGULATION OF HMW-GS GENES IS REVEALED

We screened for *cis*-elements known to regulate SSP synthesis among all the HMW-GS gene promoters of cv. Renan. By annotating these promoters we found that they had a few regulatory elements in common, mostly organized into five CCRMs. Since HMW-GS genes show similar patterns of spatial and temporal expression, these common *cis*-elements might be involved in their global regulation and consequently may provide a minimal regulatory framework needed for the developmental and environmental (i.e., in response to nitrogen supply) regulation of HMW-GS gene expression. Like the long endosperm box described in some LMW-GS gene promoters, which consists of two repeats of the endosperm box (Albani et al., 1997; Juhász et al., 2011), the GLM-GATA box described here for the first time is also formed by two motifs (GATA and GLM) repeated twice in most of the promoters of HMW-GS. Our results demonstrate that the GATA-GLM box has an activator effect. Its two GLMs were able to bind SPA and were thus functional *cis*-motifs. GATA

and GLM motifs are reported to bind R1MYB and bZIP TFs. Modules able to bind MYB and bZIP proteins belong to the seven best-known combinations of *cis*-motifs and are also very well represented in *A. thaliana* and poplar promoters (Ding et al., 2012). However, these modules generally bind R2R3-MYB TFs and thus include AACA rather than GATA motifs.

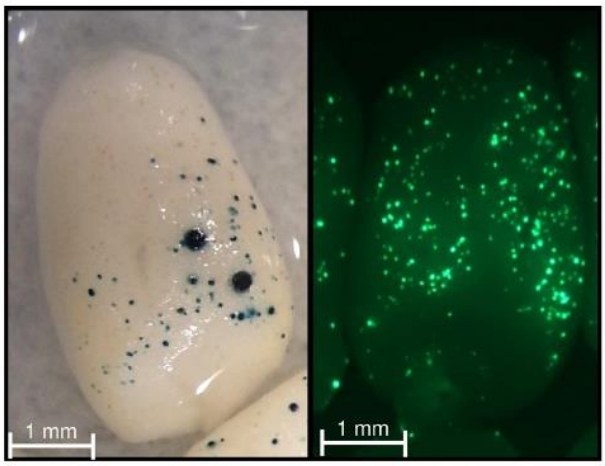


FIGURE 5 | GUS and GFP activities in wheat immature endosperm. Immature endosperm was co-bombarded with the pPrBx7-GUS and pAct-GFP constructs. Note the blue (bottom panel) and green (top panel) foci across the dorsal surface.

This GLM-GATA box is included in a CCRM with an AACA motif and a RY repeat. Notably, this conserved module is able to bind all the *cis*-motifs reported to regulate SSP synthesis. The minimal regulatory framework contains no P-box like those responsible for endosperm-specific expression of LMW-GS genes. However, several motifs have been reported to be involved in endosperm-specific expression like the CAAT, AACA and ESP motifs (Shirsat et al., 1989; Takaiwa et al., 1996; Vickers et al., 2006). The minimal regulatory framework also contains CAAT motifs. Possibly the G-box acts like the GLM in rice, which has been demonstrated to be an essential element conferring endosperm-specific expression, while P-box and AACA motifs are involved in quantitative regulation (Wu et al., 2000). In addition, the HMW-GS framework contains motifs involved in circadian rhythms. The E-box, which is able to bind bHLH and other TFs, has been reported to be involved in circadian transcriptional rhythms (Seitz et al., 2010), although exactly the same E-box sequence (5′-CATCTG-3′) was not found in the HMW-GS promoters.

Previous reports demonstrated that the 277 bp immediately upstream of the TSS are sufficient for temporal and tissue-specific regulation (Halford et al., 1989; Norre et al., 2002). There is also strong evidence indicating that mutations in this region are responsible for the silencing of *Glu-A1-2* (Halford et al., 1989). However, we did not find any mutation that could alter *cis*-motifs known to be involved in SSP gene regulation. In addition, the mutations specific to *Glu-A1-2* promoter did not create or alter any of the *cis*-motifs of the PLACE database. This suggests that

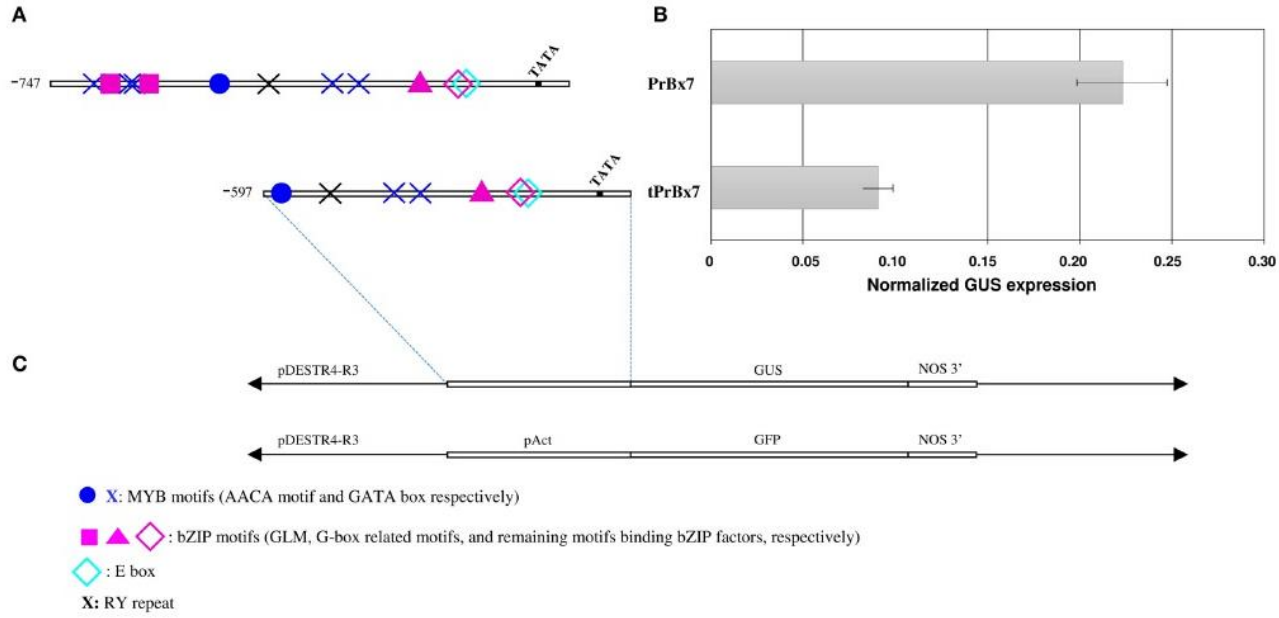


FIGURE 6 | Activity of *Glu-B1-1* gene promoter from cv. Renan (Bx7) in immature wheat endosperms using a transient expression assay. (A) Schematic representation of the constructs used. The TATA box and nucleotide positions relative to the start codon and corresponding to deleted region are indicated. Putative *cis*-regulatory elements, E-box (−259), G-box (−277), GATA box (−658, −638, −633, −368, and −350), RY motif (−525), AACA motif (−233), GLM1 and GLM2 (−647 and 626, respectively) are shown. **(B)** Normalized GUS expression of the corresponding promoters in transiently transformed endosperms. Data are the mean ± 1 SE. for $n = 10$ independent bombardments. **(C)** Schematic representation of the GUS constructs used for transformation.

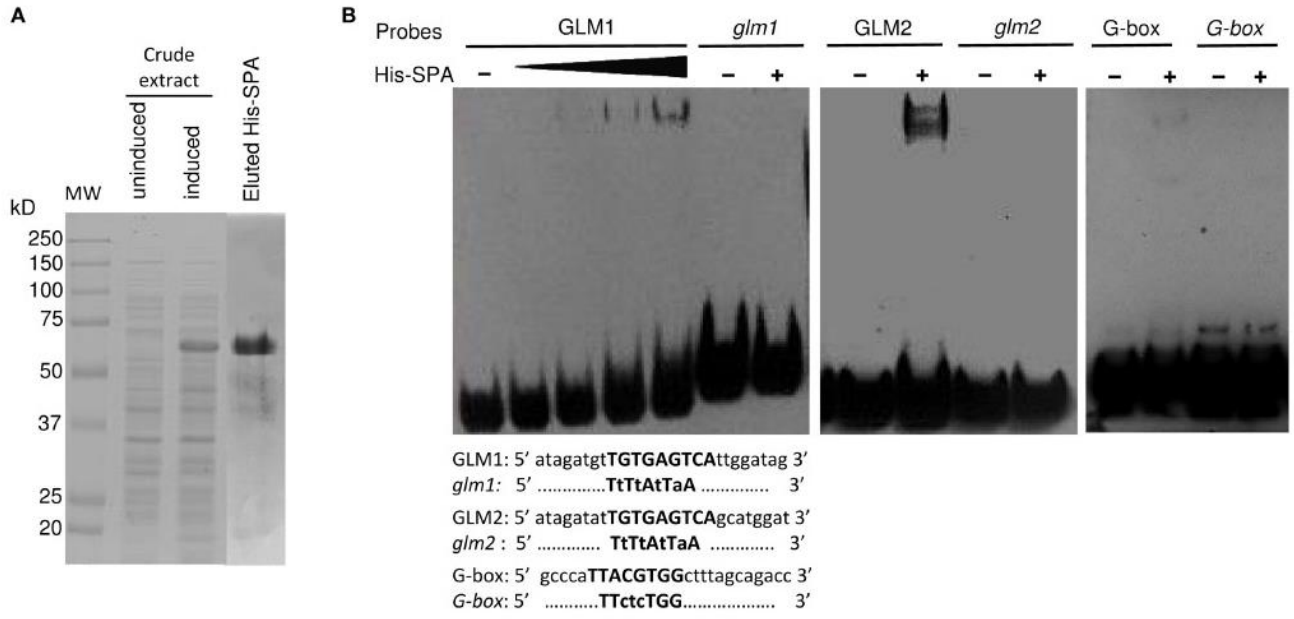


FIGURE 7 | Binding of recombinant SPA protein with the probes derived from the *Glu-B1-1* gene promoter. (A) Expression and purification of recombinant His-SPA protein. Crude extracts from uninduced and induced bacteria harboring the pHis-SPA expression vector and the eluted protein were resolved on an SDS-polyacrylamide gel. The molecular mass markers are indicated at left in kilodaltons. **(B)** EMSA of the recombinant SPA protein with the 25-bp biotin-labeled, GLM1 (−647), GLM2 (−626), and G-box (−227) probes derived from the *Glu-B1-1* gene promoter and their mutated versions *glm1*, *glm2*, and *G-box*. The sequences of the oligonucleotides used as probes are shown with the GLM1, GLM2, and G-box in bold; identical residues are represented by dots, and mutated bases are shown in lowercase.

this region may contain *cis*-motifs not yet known or that the mutations encountered in *Glu-A1-2* promoter may alter the affinity of *cis*-motifs identified for their respective TF. More precisely, this fragment contains CCRM1 and CCRM2. The latter includes the G-box found in the *Glu-D1-1* promoter and described by Norre et al. (2002) as being necessary and sufficient for expression. This box has been demonstrated to bind bZIP factors (Norre et al., 2002). CCRM2 also includes the 5′ part of the enhancer element found by Thomas and Flavell (1990), which confirms its important role. Thus, both functional validation and *in silico* analysis confirm the key role of this G-box in regulating the expression of HMW-GS genes. However, the level of expression of HMW-GS genes can be increased by adding more extensive flanking DNA (Anderson et al., 1998; Lamacchia et al., 2001), suggesting the presence of additional more distal *cis*-regulatory elements to the ones we found. This is in agreement with our results, which show a higher level of activity when the promoter of *Glu-B1-1* contained the distal GATA-GLM box. In addition, the DNA-binding affinity of SPA with one of the two GLMs of the GATA-GLM box was higher than that observed with the G-Box, suggesting a stronger role of this motif.

DIFFERENCES IN EXPRESSION ARE ONLY PARTIALLY EXPLAINED BY ANNOTATED *cis*-ELEMENTS

Our annotation strategy revealed differences at several levels: between paralogous HMW-GS genes, between orthologous HMW-GS genes and between haplotypes of a given HMW-GS gene. To investigate whether different annotated motifs induce quantitative differences in expression, we measured the level of expression from several HMW-GS promoter haplotypes. The expression of x-type gene transcripts was significantly greater than that of y-type transcripts with *Glu-B1-1* and *-D1-1* transcripts being the most expressed, *Glu-A1-1* intermediate and the two remaining genes the least abundant. This result is partially supported by GeneChip® hybridization experiments, which showed that *Glu-B1-1* is the most highly expressed HMW-GS gene in cv. Hereward (Shewry et al., 2009). However, comparing these two sources of results is not straightforward as HMW-GS probe sets cross-hybridize making it difficult to quantify the level of gene expression precisely, and only one wheat line was tested. Comparison of the consensus *cis*-motif framework of *Glu-1-1* with that of *Glu-1-2* showed several differences, which would be expected to impact their expression. Particularly, all *Glu-1-1* promoters contain an additional motif able to bind GAMYB upstream of the GLM-GATA box. Moreover, in the two most highly expressed genes, a G-box-related motif and a CAAT motif were located a few bases upstream of the GLM-GATA box and the RY repeat motif, respectively. This may enhance the activator effect of CCRM4, which contains two additional motifs.

Our results also demonstrate significant differences in the expression levels in relation to the haplotypes of the promoters for *Glu-A1-1*, *-B1-1*, and *-B1-2*. For *Glu-A1-1*, the transcription from haplotypes h2 and h5 was severely reduced for the null allele. This is in agreement with previous data on SSP synthesis in developing grains of cv. Hereward, which also has a null allele (Shewry et al., 2009). A C/T change in the coding sequence of this null allele creates a premature stop codon that could explain why this

gene is inactive (De Bustos et al., 2000). However, this does not explain the low levels of expression of these haplotypes as the qRT-PCR primers used to detect transcripts in this analysis are located upstream of this mutation. The very low transcription level of this null allele may be due to sequence polymorphism in the promoter as it has been demonstrated for the null *Glu-A1-2* allele (Halford et al., 1989). There were no obvious differences in our annotation of haplotypes of the *Glu-A1-1* promoter that could explain the large differences in expression observed. This is unlike the case of *Glu-A1-2*, which is silent and shows a particular *cis*-motif organization upstream of position −370 when compared with other y-type HMW-GS genes. However, a 277-bp fragment immediately upstream of the *GluA1-2* TSS was not able to generate any transcriptional activity (Halford et al., 1989). The organization of this fragment is quite similar to that of other expressed y-type promoters, so it is difficult to hypothesize how the gene is silenced. As expected, *Glu-B1-1* in Glenlea (line accession no. 3358) strongly expresses the Bx7 subunit transcript. This over-expression is explained by a 10.3-kb duplication including a second copy of *Glu-B1-1* (Ragupathy et al., 2008). Again, our annotation of the promoter alone does not show obvious differences that could explain the different levels of expression. In agreement with the results of Halford et al. (1989), the deletion found in the h3 haplotype does not impact the level of expression, which confirms that it plays no role in transcriptional regulation.

These results suggest that other mechanisms are able to modulate HMW-GS gene expression, such as *cis*-elements located further upstream of the region studied here. This would agree with results of Wang et al. (2013), who described the presence of key regulatory sequences in the distal sequence of *Glu-B1-1*, especially a Py-rich stretch at about position -2000. This sequence has been reported to cause a high level of expression in tomato (Daraselia et al., 1996). Methylation of DNA may also be involved in HMW-GS expression regulation, as shown for hordein genes in barley (Sorensen et al., 1996; Radchuk et al., 2005), even though no CpG islands were detected in the wheat promoter regions studied here using the PlantPAN search engine (Chang et al., 2008).

In conclusion, this work reveals a minimal regulatory framework shared by all the wheat HMW-GS gene promoters. The *cis*-elements organization is conserved, including all the motifs known to be involved in the regulation of SSP genes. The conservation of this regulatory framework strongly suggests that it is involved in the regulation of this gene family. The bipartite endosperm box was not found but a CCRM with the GATA-GLM box with an RY repeat and an AACA motif is present in all the promoters. The CCRMs, which occur at similar relative positions in all the promoters of this small family, presumably have a common evolutionary origin, suggesting that they may be functional. However, validating their functional roles requires further experiments. The "*in silico* footprint" described here will help to select motifs for functional validation, as shown here by transient expression assays of *Glu-B1-1* promoter. Our annotations do not directly account for differences in expression among promoter haplotypes, suggesting that other mechanisms may be involved in regulating HMW-GS gene expression.

ACKNOWLEDGMENTS

The authors would like to thank Rachel Carol from Emendo Bioscience Ltd. for English corrections. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under the grant agreement n° FP7-613556.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00621/abstract

REFERENCES

Albani, D., Hammond-Kosack, M. C., Smith, C., Conlan, S., Colot, V., Holdsworth, M., et al. (1997). The wheat transcriptional activator SPA: a seed-specific bZIP protein that recognizes the GCN4-like motif in the bifactorial endosperm box of prolamin genes. *Plant Cell* 9, 171–184. doi: 10.1105/tpc.9.2.171

Albani, D., and Robert, L. S. (1995). Cloning and characterization of a *Brassica napus* gene encoding a homologue of the B subunit of a heteromeric CCAAT-binding factor. *Gene* 167, 209–213. doi: 10.1016/0378-1119(95)00680-X

Allaby, R. G., Banerjee, M., and Brown, T. A. (1999). Evolution of the high molecular weight glutenin loci of the A, B, D, and G genomes of wheat. *Genome* 42, 296–307. doi: 10.1139/g98-114

Anderson, O. D., Abraham-Pierce, F. A., and Tam, A. (1998). Conservation in wheat high-molecular-weight glutenin gene promoter sequences: comparisons among loci and among alleles of the *GLU-B1-1* locus. *Theor. Appl. Genet.* 96, 568–576. doi: 10.1007/s001220050775

Armone, M. I., and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124, 1851–1864.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analysing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi: 10.1093/nar/gkl198

Balfourier, F., Roussel, V., Strelchenko, P., Exbrayat-Vinson, F., Sourdille, P., Boutet, G., et al. (2007). A worldwide bread wheat core collection restricted to a full 384 deep well storage plate. *Theor. Appl. Genet.* 114, 1265–1275. doi: 10.1007/s00122-007-0517-1

Baranowskij, N., Frohberg, C., Prat, S., and Willmitzer, L. (1994). A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator. *EMBO J.* 13, 5383–5392.

Bäumlein, H., Nagy, I., Villarroel, R., Inze, D., and Wobus, U. (1992). *Cis*-analysis of a seed protein gene promoter: the conservation RY repeat CATGCATG within the legumin box is essential for tissue-specific expression of a legumin gene. *Plant J.* 2, 233–239.

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

Branlard, G., and Dardevet, M. (1985). Diversity of grain protein and bread wheat quality. II. Correlation between high molecular weight subunits of glutenin and flour quality characteristics. *J. Cereal Sci.* 3, 345–354. doi: 10.1016/S0733-5210(85)80007-2

Carman, J. G., and Bishop, D. L. (2004). Diurnal O2 and carbohydrate levels in wheat kernels during embryony. *J. Plant Physiol.* 161, 1003–1010. doi: 10.1016/j.jplph.2004.01.003

Chang, W. C., Lee, T. Y., Huang, H. D., Huang, H. Y., and Pan, R. L. (2008). PlantPAN: Plant Promoter Analysis Navigator, for identifying combinatorial *cis*-regulatory elements with distance constraint in plant gene group. *BMC Genomics* 9:561. doi: 10.1186/1471-2164-9-561

Chao, S., Zhang, W., Akunov, E., Sherman, J., Ma, Y., Luo, M. C., et al. (2009). Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. *Mol. Breed.* 23, 23–33. doi: 10.1007/s11032-008-9210-6

Chaudhary, J., and Skinner, M. K. (1999). Basic helix-loop-helix proteins can act at the E-box within the serum response element of the c-fos promoter to influence hormone-induced promoter activation in Sertoli cells. *Mol. Endocrinol.* 13, 774–786. doi: 10.1210/mend.13.5.0271

Choi, H., Hong, J., Ha, J., Kang, J., and Kim, S. Y. (2000). ABFs, a family of ABA-responsive element binding factors. *J Biol. Chem.* 275, 1723–1730. doi: 10.1074/jbc.275.3.1723

Ciceri, P., Gianazza, E., Lazzari, B., Lippoli, G., Genga, A., Hoschek, G., et al. (1997). Phosphorylation of Opaque2 changes diurnally and impacts Its DNA binding activity. *Plant Cell* 9, 97–108. doi: 10.1105/tpc.9.1.97

Ciceri, P., Locatelli, F., Genga, A., Viotti, A., and Schmidt, R. J. (1999). The activity of the maize Opaque2 transcriptional activator is regulated diurnally. *Plant Physiol.* 121, 1321–1327. doi: 10.1104/pp.121.4.1321

Conlan, R. S., Hammond-Kosack, M., and Bevan, M. (1999). Transcription activation mediated by the bZIP factor SPA on the endosperm box is modulated by ESBF-1 *in vitro*. *Plant J.* 19, 173–181. doi: 10.1046/j.1365-313X.1999.00522.x

Daraselia, N. D., Tarchevskaya, S., and Narita, J. O. (1996). The promoter for tomato 3-hydroxy-3-methylglutaryl coenzyme A reductase gene 2 has unusual regulatory elements that direct high-level expression. *Plant Physiol.* 112, 727–733. doi: 10.1104/pp.112.2.727

De Bustos, A., Rubio, P., and Jouve, N. (2000). Molecular characterization of the inactive allele of the gene *Glu-A1* and the development of a set of AS-PCR markers for HMW glutenins of wheat. *Theor. Appl. Genet.* 100, 1085–1094. doi: 10.1007/s001220051390

De Mendiburu, F. (2014). *Agricolae: Statistical Procedures for Agricultural Research.* R package version 1.1-8. Available online at: http://CRAN.R-project.org/package=agricolae

Dereeper, A., Nicolas, S., Lecunff, L., Bacilieri, R., Doligez, A., Peros, J. P., et al. (2011). SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics* 12:134. doi: 10.1186/1471-2105-12-134

Diaz, I., Martinez, M., Isabel-La Moneda, I., Rubio-Somoza, I., and Carbonero, P. (2005). The DOF protein, SAD, interacts with GAMYB in plant nuclei and activates transcription of endosperm-specific genes during barley seed development. *Plant J.* 42, 652–662. doi: 10.1111/j.1365-313X.2005.02402.x

Diaz, I., Vicente-Carbajosa, J., Abraham, Z., Martinez, M., Isabel-La Moneda, I., and Carbonero, P. (2002). The GAMYB protein from barley interacts with the DOF transcription factor BPBF and activates endosperm-specific genes during seed development. *Plant J.* 29, 453–464. doi: 10.1046/j.0960-7412.2001.01230.x

Ding, J., Hu, H., and Li, X. (2012). Thousands of *cis*-regulatory sequence combinations are shared by Arabidopsis and poplar. *Plant Physiol.* 158, 145–155. doi: 10.1104/pp.111.186080

Eini, O., Yang, N., Pyvovarenko, T., Pillman, K., Bazanova, N., Tikhomirov, N., et al. (2013). Complex regulation by Apetala2 domain-containing transcription factors revealed through analysis of the stress-responsive TdCor410b promoter from durum wheat. *PloS ONE* 8:e58713. doi: 10.1371/journal.pone.0058713

Fauteux, F., and Strömvik, M. (2009). Seed storage protein gene promoters contain conserved DNA motifs in Brassicaceae, Fabaceae and Poaceae. *BMC Plant Biol.* 9:126. doi: 10.1186/1471-2229-9-126

Fujiwara, T., and Beachy, R. N. (1994). Tissue-specific and temporal regulation of a beta-conglycinin gene: roles of the RY repeat and other *cis*-acting elements. *Plant Mol. Biol.* 24, 261–272. doi: 10.1007/BF00020166

Gu, Y. Q., Salse, J., Coleman-Derr, D., Dupin, A., Crossman, C., Lazo, G. R., et al. (2006). Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics* 174, 1493–1504. doi: 10.1534/genetics.106.060756

Gupta, R. B., and MacRitchie, F. (1994). Allelic variation at glutenin subunit and gliadin loci, *Glu-1*, *Glu-3*, and *Gli-1* of common wheats II. Biochemical basis of the allelic effects on dough properties. *J. Cereal Sci.* 19, 19–29. doi: 10.1006/jcrs.1994.1004

Halford, N. G., Forde, J., Shewry, P. R., and Kreiss, M. (1989). Functional analysis of the upstream regions of a silent and an expressed member of a family of wheat seed protein genes in transgenic tobacco. *Plant Sci.* 62, 207–216. doi: 10.1016/0168-9452(89)90083-6

Hammond-Kosack, M. C., Holdsworth, J. M., and Bevan, W. M. (1993). *In vivo* footprinting of a low molecular weight glutenin gene (*LMWG-1Dl*) in wheat endosperm. *EMBO J.* 12, 545–554.

Haseneyer, G., Ravel, C., Dardevet, M., Balfourier, F., Sourdille, S., Charmet, G., et al. (2008). High level of conservation between genes coding for the GAMYB transcription factor in barley (*Hordeum vulgare* L.) and bread wheat (*Triticum aestivum* L.) collections. *Theor. Appl. Genet.* 117, 321–331. doi: 10.1007/s00122-008-0777-4

He, Z. H., Liu, L., Xia, X. C., Liu, J. J., and Peña, R. J. (2005). Composition of HMW and LMW glutenin subunits and their effects on dough properties, pan bread, and noodle quality of Chinese bread wheats. *Cereal Chem.* 82, 345–350.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant *cis*-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.* 27, 297–300. doi: 10.1093/nar/27.1.297

Isabel-La Moneda, I., Diaz, I., Martinez, M., Mena, M., and Carbonero, P. (2003). SAD: a new DOF protein from barley that activates transcription of a cathepsin B-like thiol protease gene in the aleurone of germinating seeds. *Plant J.* 33, 329–340. doi: 10.1046/j.1365-313X.2003.01628.x

Jefferson, R. A., Kavanagh, T. A., and Bevan, M. W. (1987). GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J.* 6, 3901–3907.

Juhász, A., Makai, S., Sebestyén, E., Tamás, L., and Balázs, E. (2011). Role of conserved non-coding regulatory elements in LMW glutenin gene expression. *PLoS ONE* 6:e29501. doi: 10.1371/journal.pone.0029501

Kang, J. Y., Choi, H. I., Im, M. Y., and Kim, S. Y. (2002). Arabidopsis basic leucine zipper proteins that mediated stress-responsive abscisic acid signaling. *Plant Cell.* 14, 343–357. doi: 10.1105/tpc.010362

Korkuć, P., Schippers, J. H. M., and Walther, D. (2014). Characterization and identification of *cis*-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. *Plant Physiol.* 164, 181–200. doi: 10.1104/pp.113.229716

Lamacchia, C., Shewry, P. R., Di Fonzo, N., Forsyth, J. L., Harris, N., Lazzeri, P. A., et al. (2001). Endosperm-specific activity of a storage protein gene promoter in transgenic wheat seed. *J. Exp. Bot.* 52, 243–250. doi: 10.1093/jexbot/52.355.243

Makai, S., Tamás, L., and Juhász, A. (2013). "Evolutionary differences in the transcriptional regulation of HMW glutenin genes in Triticeae," in *International Wheat Genetics Symposium* (Yokohama), 146.

Marzábal, P., Busk, P. K., Ludevid, M. D., and Torrent, M. (1998). The bifactorial endosperm box of alpha-zein gene: characterisation and function of the Pb3 and GZM *cis*-acting elements. *Plant J.* 16, 41–52. doi: 10.1046/j.1365-313x.1998.00272.x

Mena, M., Vicente-Carbajosa, J., Schmidt, R. J., and Carbonero, P. (1998). An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. *Plant J.* 16, 53–62. doi: 10.1046/j.1365-313x.1998.00275.x

Moreno-Risueno, M. A., Gonzalez, N., Diaz, I., Parcy, F., Carbonero, P., and Vicente-Carbajosa, J. (2008). FUSCA3 from barley unveils a common transcriptional regulation of seed-specific genes between cereals and Arabidopsis. *Plant J.* 53, 882–894. doi: 10.1111/j.1365-313X.2007.03382.x

Norre, F., Peyrot, C., Garcia, C., Rancé, I., Drevet, J., Theisen, M., et al. (2002). Powerful effect of an atypical bifactorial endosperm box from wheat *HMWG-Dx5* promoter in maize endosperm. *Plant Mol. Biol.* 50, 699–712. doi: 10.1023/A:1019953914467

Oñate, L., Vicente-Carbajosa, J., Lara, P., Diaz, I., and Carbonero, P. (1999). Barley BLZ2: a seed-specific bZIP protein that interacts with BLZ1 *in vivo* and activates transcription from the GCN4-like motif of B-hordein promoters in barley endosperm. *J. Biol. Chem.* 274, 9175–9182.

Payne, P. I., and Lawrence, G. J. (1983). Catalogue of alleles for the complex gene loci *GluA1*, *GluB1* and *GluD1* which code for the high-molecular-weight subunits of glutenin in hexaploid wheat. *Cereal Res. Comm.* 11, 29–35.

Payne, P. I., Nightingale, M. A., Krattiger, A. F., and Holt, L. M. (1987). The relationship between HMW glutenin subunit composition and the breadmaking quality of British grown wheat varieties. *J. Sci. Food Agric.* 40, 51–65. doi: 10.1002/jsfa.2740400108

Piechulla, B., Merforth, N., and Rudolph, B. (1998). Identification of tomato *Lhc* promoter region necessary for circadian expression. *Plant Mol. Biol.* 38, 655–662. doi: 10.1023/A:1006094015513

Purugganan, M. D. (2000). The molecular population genetics of regulatory genes. *Mol. Ecol.* 9, 1451–1461. doi: 10.1046/j.1365-294x.2000.01016.x

Radchuk, V. V., Sreenivasul, N., Radchuk, R. I., Wobus, U., and Weschke, W. (2005). The methylation cycle and its possible functions in barley endosperm development. *Plant Mol. Biol.* 59, 289–307. doi: 10.1007/s11103-005-8881-1

Ravel, C., Martre, P., Romeuf, I., Dardevet, M., El-Malki, R., Bordes, J., et al. (2009). Nucleotide polymorphism in the wheat transcriptional activator Spa influences

its pattern of expression and has pleiotropic effects on grain protein composition, dough viscoelasticity, and grain hardness. *Plant Physiol.* 151, 2133–2144. doi: 10.1104/pp.109.146076

Ravel, C., Praud, S., Murigneux, A., Canaguier, A., Sapet, F., Samson, D., et al. (2006). Single-Nucleotide Polymorphisms (SNPs) frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). *Genome* 49, 1131–1139. doi: 10.1139/g06-067

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rombauts, S., Dehais, P., Van Montagu, M., and Rouze, P. (1999). PlantCARE, a plant *cis*-acting regulatory element database. *Nucleic Acids Res.* 27, 295–296. doi: 10.1093/nar/27.1.295

Rubio-Somoza, I., Martinez, M., Abraham, Z., Diaz, I., and Carbonero, P. (2006a). Ternary complex formation between HvMYBS3 and other factors involved in transcriptional control in barley seeds. *Plant J* 47, 269–281. doi: 10.1111/j.1365-313X.2006.02777.x

Rubio-Somoza, I., Martinez, M., Diaz, I., and Carbonero, P. (2006b). HvMCB1, a R1MYB transcription factor from barley with antagonistic regulatory functions during seed development and germination. *Plant J.* 45, 17–30. doi: 10.1111/j.1365-313X.2005.02596.x

Ragupathy, R., Naeem, H. A., Reimer, E., Lukow, O. M., Sapirstein, H. D., and Cloutier, S. (2008). Evolutionary origin of the segmental duplication encompassing the wheat *GLU-B1* locus encoding the overexpressed Bx7 (Bx7OE) high molecular weight glutenin subunit. *Theor. Appl. Genet.* 116, 283–296. doi: 10.1007/s00122-007-0666-2

Seitz, S. B., Weisheit, W., and Mittag, M. (2010). Multiple roles and interaction factors of an E-Box element in *Clamydomonas reinhardtii*. *Plant Physiol.* 152, 2243–2257. doi: 10.1104/pp.109.149195

Shewry, P. R. (2009). Wheat. *J. Exp. Bot.* 60, 1537–1553. doi: 10.1093/jxb/erp058

Shewry, P. R., Halford, N. G., Belton, P. S., and Tatham, A. S. (2002). The structure and properties of gluten: an elastic protein from wheat grain. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 357, 133–142. doi: 10.1098/rstb.2001.1024

Shewry, P. R., Underwood, C., Wan, Y., Lovegrove, A., Bhandari, D., Toole, G., et al. (2009). Storage product synthesis and accumulation in developing grains of wheat. *J. Cereal. Sci.* 50, 106–112. doi: 10.1016/j.jcs.2009.03.009

Shirsat, A., Wilford, N., Croy, R., and Boulter, D. (1989). Sequences responsible for the tissue specific promoter activity of a pea legumin gene in tobacco. *Mol. Gen. Genet.* 215, 326–331. doi: 10.1007/BF00339737

Sorensen, M. B., Muller, M., Skerritt, J., and Simpson, D. (1996). Hordein promoter methylation and transcriptional activity in wild-type and mutant barley endosperm. *Mol. Gen. Genet.* 250, 750–760. doi: 10.1007/BF02172987

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16, 16–23. doi: 10.1093/bioinformatics/16.1.16

Suzuki, A., Wu, C. Y., Washida, H., and Takaiwa, F. (1998). Rice MYB protein OSMYB5 specifically binds to the AACA motif conserved among promoters of storage protein glutelin. *Plant Cell Physiol.* 39, 555–559. doi: 10.1093/oxfordjournals.pcp.a029404

Takaiwa, F., Yamanouchi, U., Yoshihara, T., Washida, H., Tanabe, F., Kato, A., et al. (1996). Characterization of common *cis*-regulatory elements responsible for the endosperm-specific expression of members of the rice glutelin multigene family. *Plant Mol. Biol.* 30, 1207–1221. doi: 10.1007/BF000 19553

Thomas, M. S., and Flavell, R. B. (1990). Identification of an enhancer element for the endosperm-specific expression of high molecular weight glutenin. *Plant Cell* 2, 1171–1180. doi: 10.1105/tpc.2.12.1171

Van Herpen, T. W., Riley, M., Sparks, C., Jones, H. D., Gritsch, C., Dekking, E. H., et al. (2008). Detailed analysis of the expression of an alpha-gliadin promoter and the deposition of alpha-gliadin protein during wheat grain development. *Ann. Bot.* 102, 331 –342. doi: 10.1093/aob/mcn114

Verdier, J., and Thompson, R. D. (2008). Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant Cell Physiol.* 49, 1263–1271. doi: 10.1093/pcp/pcn116

Vicente-Carbajosa, J., Moose, S. P., Parsons, R. L., and Schmidt, R. J. (1997). A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7685–7690. doi: 10.1073/pnas.94.14.7685

Vicente-Carbajosa, J., Oñate, L., Lara, P., Diaz, I., and Carbonero, P. (1998). Barley BLZ1: a bZIP transcriptional activator that interacts with endosperm-specific gene promoters. *Plant J.* 13, 629–640.

Vickers, C. E., Xue, G., and Gresshoff, P. M. (2006). A novel *cis*-acting element, ESP, contributes to high-level endosperm-specific expression in an oat globulin promoter. *Plant Mol. Biol.* 62, 195–214. doi: 10.1007/s11103-006-9014-1

Wang, K., Zhang, X., Zhao, Y., Chen, F., and Xia, G. (2013). Structure, variation and expression analysis of glutenin gene promoters from *Triticum aestivum* cultivar Chinese Spring shows the distal region of promoter 1Bx7 is key regulatory sequence. *Gene* 527, 484–490. doi: 10.1016/j.gene.2013.06.068

Wu, C., Washida, H., Onodera, Y., Harada, K., and Takaiwa, F. (2000). Quantitative nature of the Prolamin-box, ACGT and AACA motifs in a rice glutelin gene promoter: minimal cis-element requirements for endosperm-specific gene expression. *Plant J.* 23, 415–421. doi: 10.1046/j.1365-313x.2000.00797.x

Xi, D. M., and Zheng, C. C. (2011). Transcriptional regulation of seed storage protein genes in Arabidopsis and cereals. *Seed Sci. Res.* 21, 247–254. doi: 10.1017/S0960258511000237

Yanagisawa, S., and Schmidt, R. J. (1999). Diversity and similarity among recognition sequences of Dof transcription factors. *Plant J.* 15, 209–214. doi: 10.1046/j.1365-313X.1999.00363.x

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.

Received: 22 May 2014; accepted: 21 October 2014; published online: 12 November 2014.

*Citation: Ravel C, Fiquet S, Boudet J, Dardevet M, Vincent J, Merlino M, Michard R and Martre P (2014) Conserved cis-regulatory modules in promoters of genes encoding wheat high-molecular-weight glutenin subunits. Front. Plant Sci. **5**:621. doi: 10.3389/fpls.2014.00621*

This article was submitted to Plant Evolution and Development, a section of the journal Frontiers in Plant Science.

Copyright © 2014 Ravel, Fiquet, Boudet, Dardevet, Vincent, Merlino, Michard and Martre. 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.

RESUME

Titre : Inférence et analyse des réseaux de régulation de la synthèse des protéines de réserve du grain de blé tendre (*Triticum aestivum* L.) en réponse à l'approvisionnement en azote et en soufre.

La teneur et la composition en protéines de réserve du grain de blé tendre (*Triticum aestivum* L.) sont les principaux déterminants de sa valeur d'usage et de sa qualité nutritionnelle. La composition en protéines de réserve du grain est déterminée par la teneur en assimilâts azotés et soufrés par grain via des lois d'échelle qui pourraient être les propriétés émergentes de réseaux de régulation. Plusieurs facteurs de transcription intervenant dans cette régulation ont été mis en évidence, mais les voies et mécanismes impliqués sont encore très peu connus. Le constat est identique en ce qui concerne l'impact de la nutrition azotée et soufrée sur ce réseau de régulation. Le développement des outils de génomique fonctionnelle et de bioinformatique permet aujourd'hui d'aborder ces régulations de manière globale via une approche systémique mettant en relation plusieurs niveaux de régulation. L'objectif du travail présenté est d'explorer les réseaux de régulation –omiques impliqués dans le contrôle de l'accumulation des protéines de réserve dans le grain de blé tendre et leur réponse à l'approvisionnement en azote et en soufre. Une approche d'inférence de réseaux basée sur la découverte de règles a été étendue, implémentée sous la forme d'une plateforme web. L'utilisation de cette plateforme a permis de définir des sémantiques multiples afin d'inférer dans un cadre global, des règles possédant différentes significations biologiques. Des facteurs de transcription spécifiques de certains organes et certaines phases de développement ont été mis en évidence et un intérêt particulier a été apporté à leur position dans les réseaux de règles inférés, notamment en relation avec les protéines de réserve. Les travaux initiés dans cette thèse ouvrent un champ d'investigation innovant pour l'identification de nouvelles cibles de sélection variétale pour l'amélioration de la valeur technologique et de la qualité nutritionnelle du blé. Ils devraient ainsi permettre de mieux maîtriser la composition en protéines de réserve et ainsi produire des blés adaptés à des utilisations ciblées ou carencé en certaines fractions protéiques impliquées dans des phénomènes d'allergénicité et d'intolérance du gluten, ce dans un contexte d'agriculture durable et plus économe en intrants.

Mots clés : Blé tendre, *Triticum aestivum* (L.), grain, protéines de réserve, –omique, inférence de réseaux, azote, soufre.

ABSTRACT

Title: Inference and analysis of regulatory networks involved in wheat (*Triticum aestivum* L.) grain storage protein synthesis and their response to nitrogen and sulfur supply.

Grain storage protein content and composition are the main determinants of bread wheat (*Triticum aestivum* L.) end-use value. Scaling laws governing grain protein composition according to grain nitrogen and sulfur content could be the outcome of a finely tuned regulation network. Although it was demonstrated that the main regulation of grain storage proteins accumulation occurs at the transcriptomic level in cereals, knowledge of the underlying molecular mechanisms is elusive. Moreover, the effects of nitrogen and sulfur on these mechanisms are unknown. The issue of skyrocketing data generation in research projects is addressed by developing high-throughput bioinformatics approaches. Extracting knowledge on from such massive amounts of data is therefore an important challenge. The work presented herein aims at elucidating regulatory networks involved in grain storage protein synthesis and their response to nitrogen and sulfur supply using a rule discovery approach. This approach was extended, implemented in the form of a web-oriented platform dedicated to the inference and analysis of regulatory networks from qualitative and quantitative –omics data. This platform allowed us to define different semantics in a comprehensive framework; each semantic having its own biological meaning, thus providing us with global informative networks. Spatiotemporal specificity of transcription factors expression was observed and particular attention was paid to their relationship with grain storage proteins in the inferred networks. The work initiated here opens up a field of innovative investigation to identify new targets for plant breeding and for an improved end-use value and nutritional quality of wheat in the context of inputs limitation. Further analyses should enhance the understanding of the control of grain protein composition and allow providing wheat adapted to specific uses or deficient in protein fractions responsible for gluten allergenicity and intolerance.

Keywords: Wheat, *Triticum aestivum* (L.), grain, storage proteins, –omics, network inference, nitrogen, sulfur.