Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :
Gwenaëlle Marchand

Le vendredi 6 juin 2014

Titre :

Gene regulatory networks involved in drought stress responses : identification, genetic control and variability in cultivated sunflower, Helianthus annuus and its relatives.

ED SEVAB : Agrosystèmes, écosystèmes et environnement

Unité de recherche :

Laboratoire des Interactions Plantes Micro-organismes (LIPM)

Directeur(s) de Thèse :

Patrick Vincourt et Nicolas Langlade

Rapporteurs :

François Tardieu (INRA-LEPSE)
Marc Lepetit (CNRS-ILBMP)

Autre(s) membre(s) du jury :

John Burke (UGA)
Joël Piquemal (Syngenta Seeds)
Matthieu Reymond (INRA_IJPB)
Matthieu Arlat (UPS-LIPM)

Acknowledgements

All the work presented in this manuscript was co-funded by Syngenta Seeds SA and the Région Midi Pyrénées. This PhD work was also part of the OLEOSOL project funded by French public funds for competitiveness clusters (FUI), the European Regional Development Fund (ERDF), the Government of the Région Midi-Pyrénées, the Departmental Board of Aveyron (France), and the Cities Cluster of Rodez (France). This work was also supported by the French Laboratory of Excellence project "TULIP" (ANR-10-LABX-41; ANR-11-IDEX-0002-02).

All this PhD project benefit from the technical support of the Platform GeT PlaGe from INRA Toulouse and from the help of the common services of the LIPM

More specifically, we deeply thank the Auzeville-Tolosane INRA Experimental Unit, Soltis and Syngenta Seed companies for the field trial implementations described in the chapter II and that allowed the biomarker construction. This work benefited greatly of the help of Loïck Aymard in the greenhouse, field and laboratory parts of the project and Julien Berceron, Valentin Boniface, Kenneth Gamas and Richard Bonnefoy for soil sampling in the field

We also thank Syngenta Seed Company for the field trial implementation described in the chapter III and that permitted the study of the sunflower core collection. We also thank Biogemma and Syngenta Seed companies for the genotyping data described and used in this same chapter III in order to perform the genome-wide association study. A big thank you to the people that helped us to collect leaf samples: David Rengel, Didier Varès, Nicolas Langlade, Elena Garcia Navarro, Eléna Cadic, Stéphane Muños, Marie Thouly, and Samuel Vincourt.

Finally, the work, described in chapter IV related to the systems biology approach and GRN inference, benefited from the support of the Genoscope project AP09/10.

En plus de ces remerciements concernant les financements et les supports techniques qui ont rendu possible la réalisation de ces différentes parties de mon travail de thèse, je souhaiterais remercier plus personnellement certaines personnes qui m'ont permis de mener à bien ce travail.

Je tiens tout d'abord à remercier mes directeurs de thèse, Patrick Vincourt et Nicolas Langlade pour m'avoir encadrée durant ces trois années. Leurs remarques, leurs conseils ainsi que les discussions à propos des résultats des travaux que nous avons menés ensemble et sur la génétique et l'amélioration du tournesol en général, m'ont permis non seulement de finaliser cette thèse mais

aussi d'acquérir (j'espère) la curiosité et l'ouverture d'esprit nécessaire aux chercheurs. Merci également à eux pour leur relecture attentive de ce manuscrit.

Un grand merci également à Baptiste Mayjonade qui a réalisé la quasi-totalité des manips de broyage d'échantillons, d'extraction d'ARN et d'expression de gènes sur lesquelles reposent la majorité des résultats présentés dans ce manuscrit. Je tiens également à le remercier pour son aide lors du phénotypage au champ et pour sa bonne humeur tout au long de ces différents projets.

Un énorme merci aussi à Didier Varès, Marie-Claude Boniface et Nicolas Blanchet pour leur aide précieuse pour la mise en place des expérimentations (serre, champs, chambre de culture) et pour tout leur travail, leur aide et leurs conseils lors des campagnes de phénotypage. Merci également à eux trois, ainsi qu'à Lolita Lorenzon pour leur soutien et encouragements durant ces années de thèse et pour m'avoir toujours remonté le moral au bon moment.

Je tiens à remercier aussi Stéphane Muños et Brigitte Mangin pour leur aide et leur implication dans le projet de génétique d'association. Merci aussi à eux pour leurs conseils avisés en génétique et statistiques.

Merci également à toutes les personnes, outre celles déjà citées, qui au fil de ces années ont partagées mon bureau dans la bonne humeur (et m'ont nourri en chocolat) : Nicolas Pouilly, Amandine Bordat, David Rengel, Yannick Lippi, Elena Cadic, Falah As-Sadi, Quentin Gascuel et Johan Louarn (pour son schéma de tournesol!). Merci aussi à tous les autres membres de l'équipe tournesol qui ont participé à la bonne ambiance général de cette équipe et avec qui j'ai pu partager durant ces trois années.

Je tiens également à remercier tous les membres de mon comité de thèse que je n'ai pas encore pu citer précédemment (Marie Coque, Joël Piquemal, Bénédicte Quilot) pour leur conseils lors de ces étapes clés de ma thèse. En particulier, un grand merci à Matthieu Vignes pour son travail sur l'inférence de réseau et à Philippe Debaeke et Pierre Casadebaig pour leur aide lors du travail sur le biomarqueur. Un merci un peu spécial à Marie Coque qui a été la première à me faire découvrir la génétique du tournesol lorsqu'elle m'a encadrée pendant mon stage de fin d'étude d'ingénieur agro et qui m'a, en tout premier lieu, encouragée à poursuivre en thèse.

Enfin, un grand merci à ma famille qui m'a encouragée durant ces années de thèse. Un merci tout particulier à Pierre Gossart, qui a vécu par procuration cette thèse (y compris les levés à 3h du matin pour les mesures de potentiel hydrique !) ainsi qu'à Matéi Chihaia qui a patiemment relu ce manuscrit pour y faire la chasse aux fautes d'anglais.

Table of Contents

List of Figures and Tables

Abbreviations List

ABA: abscisic acid

ABF: ABRE-binding factor

ABRE: ABA-responsive element

ACC: Ethylene

ANCOVA: Analysis of Covariance

ANOVA: Analysis of variance

ASPL: Average shortest path length

BLAST: Basic Local Alignment Search Tool

BLUP: Best Linear Unbiased Predictors

Brass: Brassinol

CCA: Canonical correlation analysis

CDPK: Calcium dependent protein kinases

CETIOM: Centre technique Interprofessionnel des Oleagineux

CID: Carbon isotopic discrimination

CMS: cytoplasmic male sterility

CNRS: Centre National de Recherche Scientifique

Ct: Threshold Cycle

DNA: Deoxyribonucleic acid

E: Transpiration

EN: Normalized transpiration

eQTL: Expression QTL

ERF: ethylene-responsive element binding factor

EST: Expressed Sequence Tag

ET: Evapotranspiration

ET_0: reference evapotranspiration

ET_M: Maximal crop evapotranspiration

FAO: Food and Agriculture Organization

FtotSW: Fraction of Total Soil Water

FTSW: Fraction of Transpirable Soil Water

GA3: Gibberellic acid

GGM: Gaussian graphical model

GO: Gene ontology

GRN: gene regulatory network

GWAS: Genome wide association study

GxE: Genotype x environment

IAA: Auxine

I**NRA**: Institut National de Recherche Agronomique

IPCC: Intergovernmental Panel on Climate Change

ITW: Integrated Transpired Water

LAI: Leaf Area Index

LD: Linkage disequilibrium

LEA: Late Embryogenic Abundant

LG: Linkage group

MAF: Minor allele frequency

MAGIC: Multiparent Advenced Generation Inter-Cross

MAPK: Mitogen activated protein kinase

MeJA: Methyl Jasmonate

NAM: Nested association mapping

OP: Osmotic potential

ORF: Open Reading Frame

PCA: Principal component analysis

q-RT-PCR: Quantitative Real Time Polymerase Chain Reaction

QTL: Quantitative Trait Locus

RF: Random Forest

RIL: Recombinant Inbred Lines

RMSE: Root Mean Square Error

RNA: Ribonucleic acid

ROS: Reactive oxidative species

RWC: Relative water content

SA: Salicylic acid

SLA: Specific Leaf Area

SNP: Single Nucleotide Polymorphism

Stri : Strigolactone

SWC: Soil Water Content

TAIR: The Arabidopsis Information Resource

TLA: Total Leaf Area

TTSW: Total Transpirable Soil Water

WD: Water deprived

WSB: Water Status Biomarker

WUE: Water Use Efficiency

Ψ_{PD} : Pre-dawn leaf water potential

Chapter I: Introduction

I.1. Drought issues and challenges of agronomy in a context of global warming

In 2013, the Food and Agriculture Organization (FAO) estimated that 842 million people (around one in eight people in the world) suffer from chronic hunger (report 40 of the Committee on world Food Security, 2013). Increase of demography has for consequences that food demand and therefore risks of hunger will also rise. Moreover, climate change will also affect at least two dimensions of the food security i.e availability and stability. Indeed, global climate change not only involves temperature increase and precipitation diminution, which lead to changes in land suitability and crop yield but also increases in the frequency and severity of extreme events such as droughts (Schmidhuber & Tubiello, 2007). Drought is one of the most important constraints to plant productivity (Farooq *et al.*, 2009). Hence, increasing population pressures and climate change is likely to emphasize the effects of drought (Somerville & Briscoe, 2001).

Depending on the climatic scenario studied by the Intergovernmental Panel on Climate Change (IPCC), the number of people at risk of hunger in the world in 2080 would increase by 5 to 26% due to climate change (Schmidhuber & Tubiello, 2007).

Moriondo *et al.*, (2010) projected a scenario of European agriculture in +2°C (above pre-industrial levels) world in order to estimate potential effects of climate change and variability on crop production in this region. With this scenario, in the area of the Mediterranean basin, the summer period is projected to exhibit a rainfall decrease up to 35% (Figure I.1.A) and an increase of higher temperatures implying more frequent drought stress events. These changes in average climate and climate variability would affect yields according to crop type and geographical areas. Some northern regions are expected to benefit from this average increase of temperature; however, southern zones could largely suffer of the impact of climate change. For example, sunflower crop in the Mediterranean basin, in the period 2071-2100, is expected to have a yield reduced by 13% in average with respect to the baseline 1961-1990. Depending on the scenario for future climate defined by the IPCC (A2: medium-high greenhouse gases emission and B2: low-medium greenhouse gases emission) this loss of yield could rise to 35% (Figure I.1.B) due in particular to higher drought stress frequency at anthesis (Moriondo *et al.*, 2011).

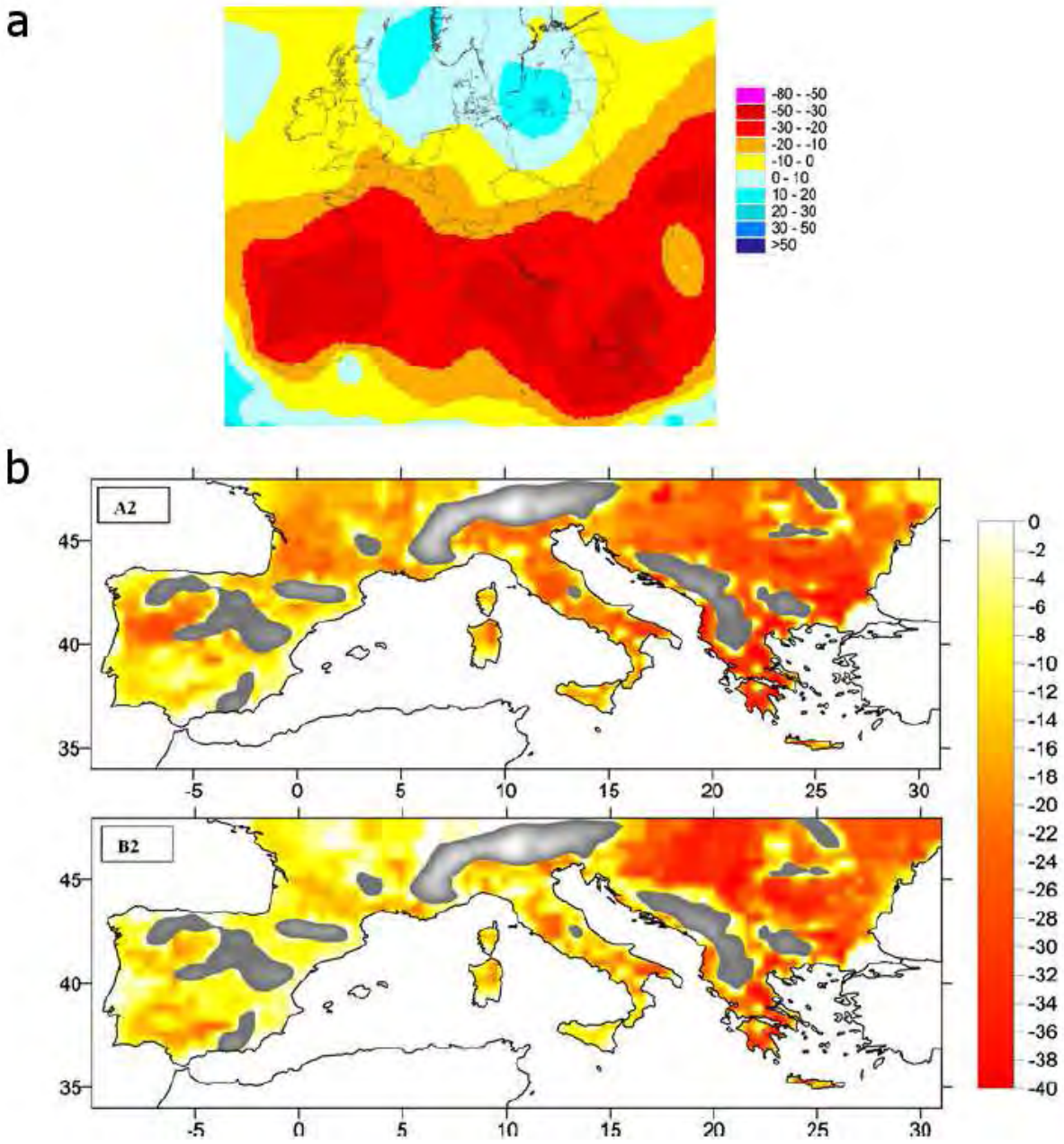


Figure I.1: Projections of rainfall and sunflower yield for the Mediterranean Basin.

(a) Projection of rainfall changes in a scenario of European agriculture in +2°C (above pre-industrial levels) world (Moriondo *et al.*, 2010). (b) Projection of the average change in sunflower yield in A2 and B2 scenarios (2071-2100) with respect to the baseline period (1961-1990). Grey areas are considered not suitable for cultivation (Moriondo *et al.*, 2011).

Therefore, many efforts have been made to improve crop productivity under water-limiting conditions. The negative impacts of drought stress on yield have been reduced thanks to, firstly better crop management and secondly breeding activities. However, there is still a large gap between yields in optimal and in water stress conditions (Cattivelli *et al.*, 2008). For instance for sunflower crops in France, potential of production due to genetic gain increases regularly since 1970 at the rate of 1.3% per year (Vear *et al.*, 2003). However, while the potential sunflower yield can reach 40 qx/ha, the actual average yield in 2012 was 23 qx/ha in France and only 15 qx/ha in Europe (FAOstat, 2014). In this context, selection of drought tolerant varieties remains an important goal for breeders and is of strategic importance to minimize hunger risks for the future. In order to achieve this goal, a better understanding of plant drought stress responses is necessary at physiological, molecular and genetic levels.

I.2. Drought stress in plant physiology

Soil water is used by plants during their development to transport nutrients and to produce biomass through the mechanism of photosynthesis. However, plants lack the capacity to perform photosynthesis without water losses. Therefore, depending on the species, the variety and the environmental conditions, their water use efficiency (WUE), i.e the ratio of CO_2 assimilation or biomass accumulation to water losses, varies. For example, sunflower and soybean have been shown to have a WUE of 54 and 30kg.ha^{-1}.cm^{-1} respectively (Anderson *et al.*, 2003). Plants lose water through the phenomenon of the evapotranspiration (ET). This last one, for crops, takes into account water evaporation at soil and leaf surfaces and also transpiration of free water in plant tissues through stomata (Figure I.2). Evapotranspiration is dependent on climatic and environmental conditions such as the evaporative demand of the atmosphere as well as on plant characteristics. The part of the evapotranspiration due to weather conditions (radiation, air temperature, humidity and wind speed) is called reference evapotranspiration ET_0 and can be calculated with the Peinman-Monteith equation (Monteith, 1965). Evapotranspiration of a crop can be estimated from ET_0 to which cultural and stress coefficients are applied in order to take into account plant characteristics and crop management that influence evapotranspiration (Figure I.2). Indeed, water demand of a crop will vary depending on the crop species, variety, and phenological stage. Maximal crop evapotranspiration (ET_M) evaluates these plant characteristics and refers to the evaporating demand of a crop that grows in large fields under optimum soil water, excellent management, and environmental conditions, and that achieves full production under the given climatic conditions (FAO, Irrigation and drainage paper 56). Moreover, plant physiology and crop management factors

such as soil salinity, fertilizers application, penetrability of soil horizon, diseases control or soil water content affect the crop development and therefore the real evapotranspiration (ET).

Therefore, adopting an eco-physiological point of view, water status can be defined as the difference between the soil water available for the plant and water losses due to evapotranspiration (Tardieu & Tuberosa, 2010). Drought stress is perceived by the plant when an important water deficit occurs (i.e if the water losses are more important than the soil water availability). Most importantly, this definition of drought stress is not only dependent on the environmental conditions such as precipitation frequency, evaporative demand or amount of available soil water but also on plant characteristics.

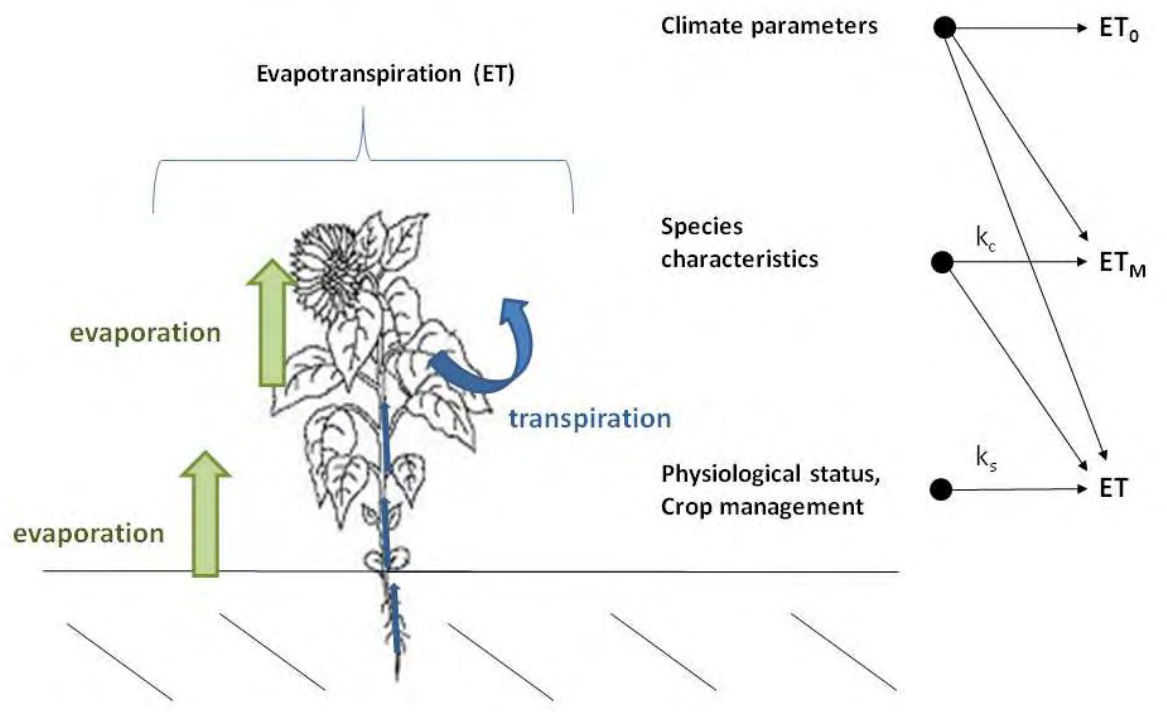


Figure I.2: Principal components of the evapotranspiration (adapted from FAO Irrigation and drainage paper 56)

ET_0: reference evapotranspiration; ET_M: maximal evapotranspiration; ET: real evapotranspiration; k_c: cultural coefficient; k_s: stress coefficient;

I.3. Plant traits affected by drought stress

Drought stress affects the plant at different levels: morphological, physiological, and molecular. Moreover, it can occur at all the phenological stages of the crop and have a final impact on crop yield. In the following sections, we will broach the different plant traits affected by drought.

I.3.1. Whole plant scale

At the whole plant scale, drought stress affects different traits according to the phenological stage of the crop. At the early stage, drought affects seed germination. Later, during the vegetative stages, it reduces plant growth, increases leaf wilting and senescence. All these phenomena lead to a reduce leaf area. During reproductive stages, grain initiation and filling can be impacted. Moreover, germination rate, grains initiation and filling are directly correlated to the crop yield (Figure I.3). That is why many drought-induced yield reductions have been reported in several crop species even though it depends upon the severity and duration of the stress period (Farooq *et al.*, 2009). For example, in sunflower a drought stress during the reproductive stage can lead to a yield reduction of 60% (Farooq *et al.*, 2009).

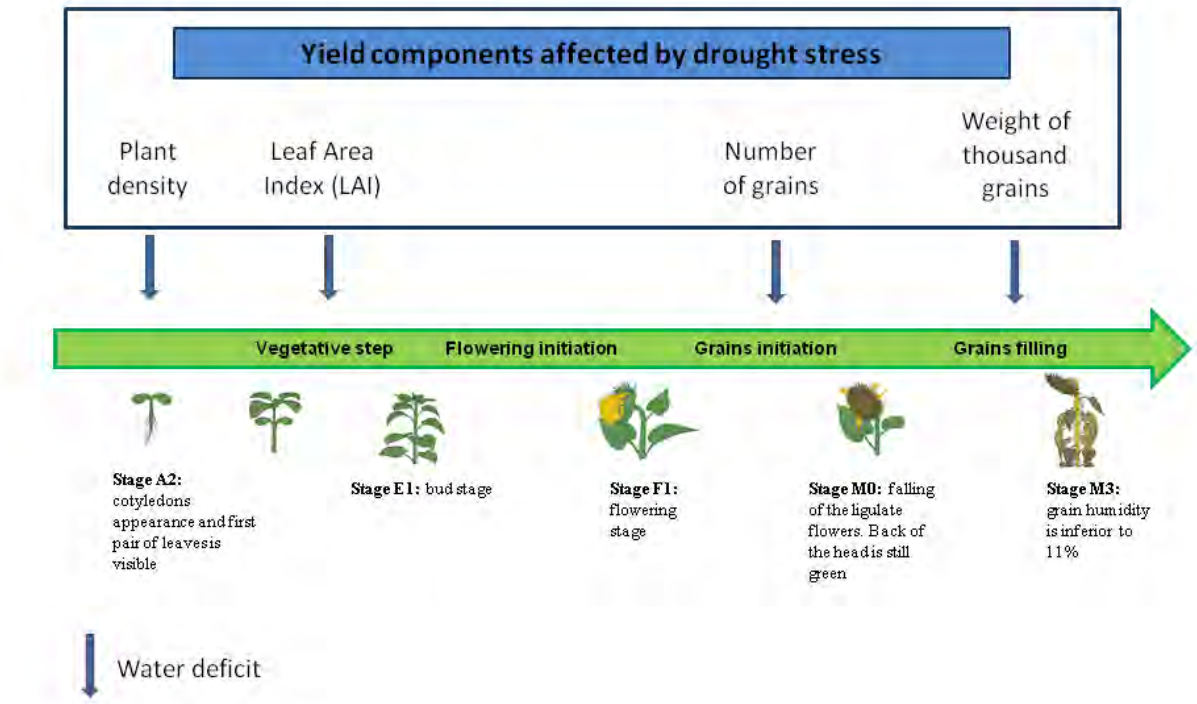


Figure I.3: Whole-plant level traits affected by water deficit and leading to yield losses under water deficit (adapted from CETIOM).

I.3.2. Tissue or cellular scales

Underlying morphological traits or whole plant level traits affected by drought, there are several physiological processes observable at tissue or cellular scales.

During drought stress, changes appear in plant water relationships. They are consequences of modifications of relative water content, leaf water potential, stomatal resistance, transpiration rate or canopy temperature (Farooq *et al.*, 2009). Therefore the tissue and cellular water statuses in

leaves are affected during drought. It can lead to a loss of turgor and therefore to a diminution of the growth and to an increase of leaves wilting described at the whole plant level (Taiz & Zeiger, 2006). Another point is that decreasing water availability under drought generally results in limited total nutrient uptake. Therefore nutrient tissue concentrations diminish. Even though plant species and genotypes may vary in their responses to mineral uptake under water stress, in general, it induces an increase in nitrate, a decline in phosphate and no clear effects on potassium (Garg *et al.*, 2004). Moreover, the nutrient utilization efficiency is also lower under drought stress (Farooq *et al.*, 2009).

During drought stress, the photosynthesis is affected too. Several factors are in cause in this loss of photosynthetic activity: decrease in leaf expansion and premature leaf senescence, stomatal oscillations (Mansfield *et al.*, 1990) and decline in photosynthetic enzymes activity (Bota *et al.*, 2004) (Loreto *et al.*, 1995).

Respiration is also increased during a drought event. One of the consequences to this respiration rate augmentation is the imbalance in the utilization of carbon resources (Farooq *et al.*, 2009).

All together, limitation in nutrient utilization efficiency, decrease in photosynthetic activity and augmentation of the respiration rate can lead to a decrease in biomass production and changes in assimilate partitioning. For example, drought stress frequently enhances allocation of dry matter to the roots (Leport *et al.*, 2006). This phenomenon can lead to the limitation in grain initiation and filling observed at the whole plant level (Asch *et al.*, 2005).

Finally, as other abiotic stresses, drought leads to the production of reactive oxygen species. They can cause oxidative damage and prevent the normal functioning of the cells (Foyer & Fletcher, 2001).

I.4. Drought stress resistance mechanisms

To cope with drought stress, plants develop mechanisms and defense strategies to prevent water deficit and maintain their ability to grow, flower, and produce seeds that are commonly the main valuable production in crops. However, economic yield can be dramatically affected by a deficit in water supply conditions (Chaves & Oliveira, 2004).

A first strategy to deal with water deficit is referred as the *escape strategy*. Drought escape occurs when phenological development is successfully altered to match with the periods of soil moisture availability (Araus *et al.*, 2002). Therefore tolerance to cold during the early stages, flowering time and length of life cycle appear as key traits that can lead to drought escape. However, yield is generally correlated with the length of crop duration under favorable growing conditions

(Turner *et al.*, 2001). This is why, a strategy as the *escape* can also lead to a reduction of the potential yield.

A second strategy to cope with water stress is the *avoidance*. Plants that follow this strategy tend to maintain high tissue water potential through different mechanisms. At one end of the water flux, there is the increase of the water uptake through an extensive and more efficient root system. At the other end of the flux, there are a small leaf area and the control of the stomatal conductance to reduce the transpiration rate. Both of these mechanisms help avoid water losses

Finally, plants can develop a range of mechanisms involving *phenotypic plasticity* to adapt to water deficit: the phenotype of the plant will be modified and then increase its ability to maintain yield crop under drought stress. In contrast to the first two strategies described above that aim at preventing the water deficit events (intensity, duration and frequency), this last strategy corresponds to a strategy of adaptation. The strategy of *phenotypic plasticity* includes both morphological and physiological mechanisms. For example morphological mechanisms, under drought stress, lead some plants to reduce their leaf area by leaf shedding (Farooq *et al.*, 2009) in order to limit plant transpiration. Figure I.4 shows morphological and phenological mechanisms that lead to drought tolerance.

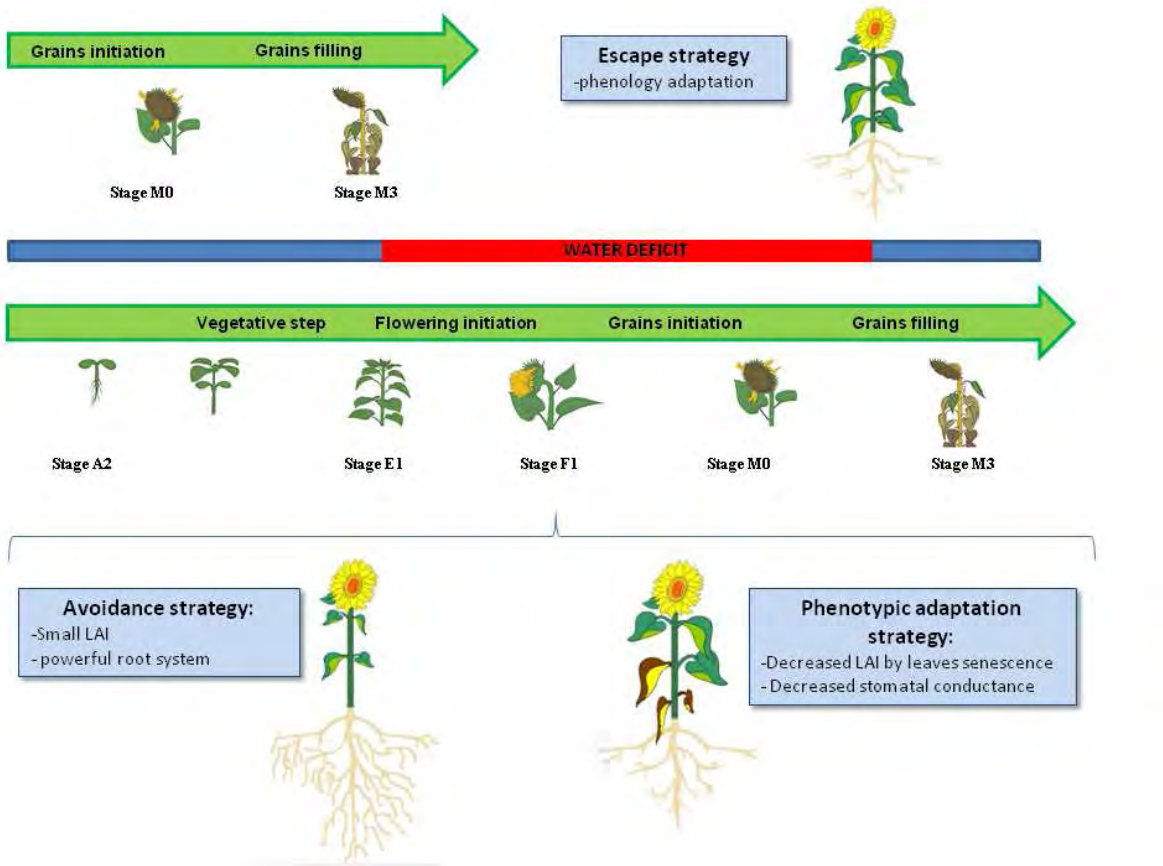


Figure I.4: Morphological and phenological responses to drought stress involved in the escape, avoidance, and phenotypic plasticity strategies (Adapted from CETIOM). LAI: Leaf Area Index.

At the physiological level, several mechanisms occur to produce *phenotype plasticity*. Osmotic adjustment is one of these mechanisms that may confer tolerance against cell drought injuries by maintaining high cell water potential through active accumulation of solutes in the cytoplasm (Turner *et al.*, 2001). It delays damage due to dehydration through the maintenance of cell turgor and physiological cell processes (Taiz & Zeiger, 2006). In a similar way, antioxidant defenses help prevent drought damages in the cell. The reactive oxygen species in plants are removed by a large range of antioxidant enzymes, lipid-soluble and water soluble scavenging molecules (Hasegawa *et al.*, 2000). This includes β-carotenes, ascorbic acid, α-tocopherol, reduced glutathione for the non-enzymatic antioxidant molecules and peroxide dismutase, peroxidase, catalase and glutathione reductase for the enzymatic compartment. Finally, it is generally accepted that the maintenance of cell membrane integrity and stability under a water stress is a major component of drought tolerance at the cellular level (Bajji *et al.*, 2002). Although the causes of membrane disruption are not well understood, it increases the chances of protein denaturation and membrane fusion due to molecular interactions (Farooq *et al.*, 2009). A range of compounds have been identified that can prevent the effects of membrane disruption as for example, proline, glutamate, glycinebetaine, polyols, trehalose and oligosaccharides (Hoekstra *et al.*, 2001). Figure I.5 illustrates physiological mechanisms leading to drought tolerance through a *phenotypic plasticity* at the cell level.

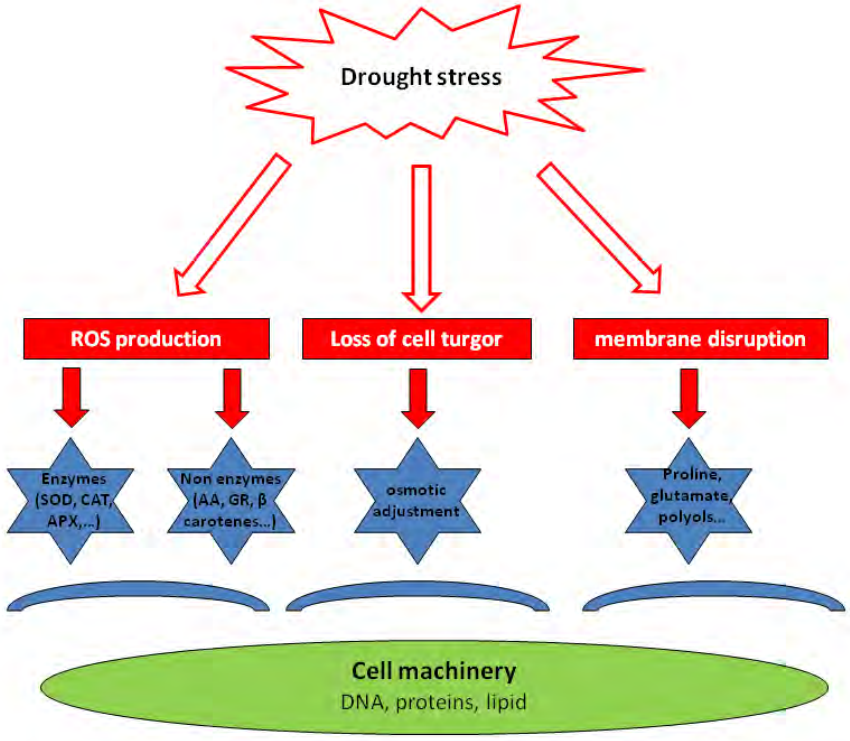


Figure I.5: Physiological mechanisms for drought tolerance. SOD: superoxide dismutase; CAT: catalase, APX: ascorbate peroxydase; AA: ascorbic acid; GR: gluthatione reductase (adapted from Farooq *et al.*, 2009).

To conclude about the strategy of *phenotypic plasticity*, plants can tolerate drought stress by developing physiological and morphological mechanisms that aimed at the conservation of cell and tissue water by reducing water losses by transpiration reduction, osmotic adjustment, scavenging of reactive species, and keeping the cell membrane stabilized (Farooq *et al.*, 2009). However, these physiological mechanisms as well as the stomatal aperture are under a complex genetic and molecular control that needs to be deciphered in order to clearly understand how drought tolerant species and genotypes accommodate to water limiting conditions.

I.5. Molecular mechanisms and regulatory networks for drought stress responses

During a drought stress event, changes in the expression of various genes take place. The transcriptional regulations of these genes in response to drought induce molecular mechanism involved in drought tolerance. However, drought tolerance is a complex phenomenon involving the combined regulation and action of many genes (Cattivelli *et al.*, 2008).

In this section, we will first describe genes which, according to the literature, encode cell effectors proteins involved in generic molecular mechanisms that allow drought tolerance. In this case and throughout this work, we use the term “effectors proteins” as opposed to “receptors and signal transducer proteins”. These proteins are largely involved in cell protection mechanisms. Then we will discuss a second type of genes which are differentially expressed during drought event and contribute to signal transduction and encodes for regulatory proteins.

I.5.1. Molecular mechanisms for cell protection during drought stress

Aquaporins are membrane proteins that can be involved in generic molecular mechanisms for drought tolerance. They facilitate and regulate passive exchange of water across membranes. Although their role in plant drought stress tolerance has not been clearly understood yet, it is generally admitted that they can regulate the hydraulic conductivity of membranes and increase water permeability (Maurel & Chrispeels, 2001). They probably play a role in soil water uptake by the roots as they are abundantly expressed in this tissue.

Synthesis of stress proteins is also a generic molecular response to drought stress. Among these stress proteins, the heat shock proteins that belong to a larger group of molecules called chaperones is of particular importance. Many heat shock proteins have been found to be induced by different abiotic stresses and in particular by drought (Coca *et al.*, 1994). They play a role in stabilizing the structure of other proteins to maintain their activities in adverse biophysical conditions. Late Embryogenic Abundant (LEA) proteins are another important group of stress

proteins. Also known as a class of dehydrins, they are accumulated in the tissues during the desiccation stage of the seed but also in vegetative tissues during periods of water deficit (Farooq *et al.*, 2009). Their particular structure with the highly conserved Lysin-rich domain allows them to be involved in hydrophobic interactions and in macromolecule stabilization (Nylander *et al.*, 2001). They also play a major role in the concentrations of ions that are accumulated during desiccation events in the cell (Gorantla *et al.*, 2007).

All these molecular drought responses are triggered by complex signaling and regulatory pathways that involve the interaction of various genes between them.

I.5.2. Regulatory networks of drought stress responses

During drought stress, various genes have been shown to be induced or repressed at the transcriptomic level using different tools such as microarray and RNA sequencing analysis on *Arabidopsis thaliana* and rice (Shinozaki & Yamaguchi-Shinozaki, 2007; Hirayama & Shinozaki, 2010; Todaka *et al.*, 2012). Transcriptional changes could be induced by drought stress, in particular, via the action of several phytohormones. An important and well-described plant hormone for drought stress is the abscisic acid (ABA). Its production is triggered by water deficit which in turn causes stomatal closure and regulates expression of drought responsive genes. Indeed, many genes involved in ABA biosynthesis were shown to be up-regulated during dehydration in *Arabidopsis* (*AtNCED3*, *AAO3*, *AtABA3* and *AtZEP*). Over-expression of *AtNCED3* in transgenic plants improved drought tolerance while, on the contrary the knockout mutants for this gene showed drought sensitive phenotypes. This suggests that *NCED3*, in particular, plays an important role in ABA accumulation during a drought event (Endo *et al.*, 2008). Accumulation of ABA in guard cells of mature leaves not only induces stomatal closure, as already mentioned, but also plays a role in stomata initiation in young leaves (Chater *et al.*, 2014) and help prevent water losses.

ABA role and signal transduction during drought stress has been studied thanks to mutants, in particular in the plant model *Arabidopsis*. Mutants for the genes *abi1* and *abi2* (ABA insensitive 1 and 2) present a wilty phenotype, that let think that *ABI1* and *ABI2* genes have important roles in ABA-dependent signal transduction pathways during a water deficit (Shinozaki & YamaguchiShinozaki, 1997; Shinozaki *et al.*, 2003). They encode type 2C protein Ser/Thr phosphatases (PP2C). This suggests that a phosphorylation/ dephosphorylation cascade might be involved in ABA signal transduction.

ABA-induced stomatal closure, as a model of plant cell responses to a water stress, has been largely studied. This phenomenon is due to a multiple chain of cellular events involving second messengers. To sum up, ABA is perceived by receptors in the guard cells. ABA-perception induces Ca^{2+} cytosolic concentration and pH increase. It causes first K^{+} and anion efflux and then water efflux.

Finally, water efflux induces guard cell volume reduction and stomatal closure (Zhu *et al.*, 2012). This is why Ca^{2+} is considered to be likely the most important second messenger in the water-stress responses in plant cells. Then, ABA signal transduction of a drought event is mediated by second messengers and various phosphorylation events in the vegetative tissues.

A fraction of drought responsive genes have been shown to be induced by the application of exogenous ABA whereas another group of genes were not affected. This demonstrates the existence of both ABA-independent and ABA-dependent regulatory pathways to regulate drought-inducible genes. Moreover, Shinozaki and Yamaguchi-Shinozaki (1997) hypothesized that four pathways play a role in the activation of stress inducible-genes: two ABA-dependent and two ABA-independent.

The first ABA-dependent pathway gathers together genes which contain ABA-Responsive Elements (ABRE) in their promoter regions. Therefore the ABRE functions as a cis-acting regulation. cDNAs for ABRE-binding (AREB) proteins, also called ABRE-binding factors (ABF) have been isolated (Choi *et al.*, 2000) and show a basic region adjacent to a Leu-zipper motif (bZIP). The ABRE motif is PyACGTGGC. The specificity of the bZIP protein binding to ABRE is due to nucleotides around the core motif ACGT. However, for ABA-responsive transcription, a single copy of ABRE is not sufficient (Yamaguchi-Shinozaki & Shinozaki, 2006). For example, ABA-responsive expression of the *Arabidopsis* gene *RD29B* in vegetative tissue requires two ABRE sequences (Uno *et al.*, 2000). Activation of *AREB/ABF* genes by ABA is not completely understood. However, phosphorylation/dephosphorylation events seem to play a key role in this ABA signaling pathway. In *Arabidopsis*, five of the nine type-2 SNF1-related protein kinases (SnRK2) are activated by ABA (Boudsocq & Lauriere, 2005). These ABA-activated SnRK2 protein kinases were shown to phosphorylate the conserved regions of AREB/ABF and therefore possibly activate them in *Arabidopsis* (Furihata *et al.*, 2006). Similar observations were made for the rice (Kobayashi *et al.*, 2005).

There are other types of ABA-dependent expressive genes involved in response to drought which could be grouped in a second ABA-dependent pathway: genes induced by MYB and MYC factors. For example, *RD22* expression is induced by ABA and is not activated through ABRE cis-acting regulation. A MYC and a MYB transcription factors (AtMYC2/RD22BP1 and AtMYB2) bind MYC and MYB recognition sites in the *RD22* promoter. These recognition sites act as cis-acting elements and cooperatively activate expression of *RD22* under drought stress (Abe *et al.*, 2003). Various other cis-acting elements have been found in drought and ABA-responsive genes. *Arabidopsis RD26* encodes a NAC protein. Microarray analysis showed that ABA and stress inducible genes were up-regulated in *RD26*-overexpressing plants and on the contrary down-regulated in *RD26*-repressed plants indicating that the NAC recognition sites may function as a cis-regulatory factor in ABA-dependent gene expression under drought stress conditions (Fujita *et al.*, 2005). Another example is the *Arabidopsis* gene *AtERF7* which binds to a cis-acting element of ABA-drought responsive genes and acts as a

repressor of their expressions. Indeed, over-expression of *AtERF7* in transgenic plants decreased drought tolerance by a reduction of ABA responses in guard cells (Yamaguchi-Shinozaki & Shinozaki, 2006).

Some genes are differentially expressed during drought events in ABA-deficient (aba) or ABA-insensitive (abi) *Arabidopsis* mutants, indicating that they are regulated by ABA-independent pathways (Shinozaki & YamaguchiShinozaki, 1997). The first ABA independent pathway gathers together genes with cis-acting elements in their promoters called DRE (TACCGACAT). DRE-binding proteins, called DREB1 and DREB2, contain the conserved DNA-binding domain found in the ERF (ethylene-responsive element binding factor) and AP2 proteins cells (Yamaguchi-Shinozaki & Shinozaki, 2006). The DREB1 proteins are mainly involved in cold stress responses, whereas DREB2 proteins are more involved in responses to drought stress. *DREB2A* has been shown to be a major transcription factor that functions in particular under water deficit stress. However, simple over-expression of *DREB2A* in transgenic plants, does not improve drought tolerance. Activation mechanisms of *DREB2A* are not clearly understood and could involve phosphorylation processes (Liu *et al.*, 1998). Finally there is a class of drought inducible genes that show no differential expression under cold or ABA application. It suggests that these genes belong to a second ABA-independent pathway. These genes are called *ERD1* (early response to dehydration1). Promoter analysis of *ERD1* allows identifying two novel cis-acting elements: a MYC like sequence and *a RPS1* site 1-like sequence that are involved in induction of *ERD1* during cell dehydration (Simpson *et al.*, 2003). cDNAs encoding MYC-like and a *RPS1* site 1-like sequences binding proteins are NAC sequences and transcription factor with zinc-finger homeodomain (ZFHD) respectively. They are both necessary to activate expression of *ERD1* (Yamaguchi-Shinozaki & Shinozaki, 2006).

The ABA-dependent and ABA-independent pathways present cross-talks between them. For example, the well studied drought differentially expressed gene *RD29A* has a promoter sequence with both ABRE and DRE cis-acting elements. This gene is governed by both ABA-dependent and ABA-independent regulations. It was confirmed by its induction by ABA and drought in non-transformed plants and by its induction by drought only in *aba* or *abi* mutants.

Figure I.6 schematizes the four regulatory pathways for drought responsive genes presented above.

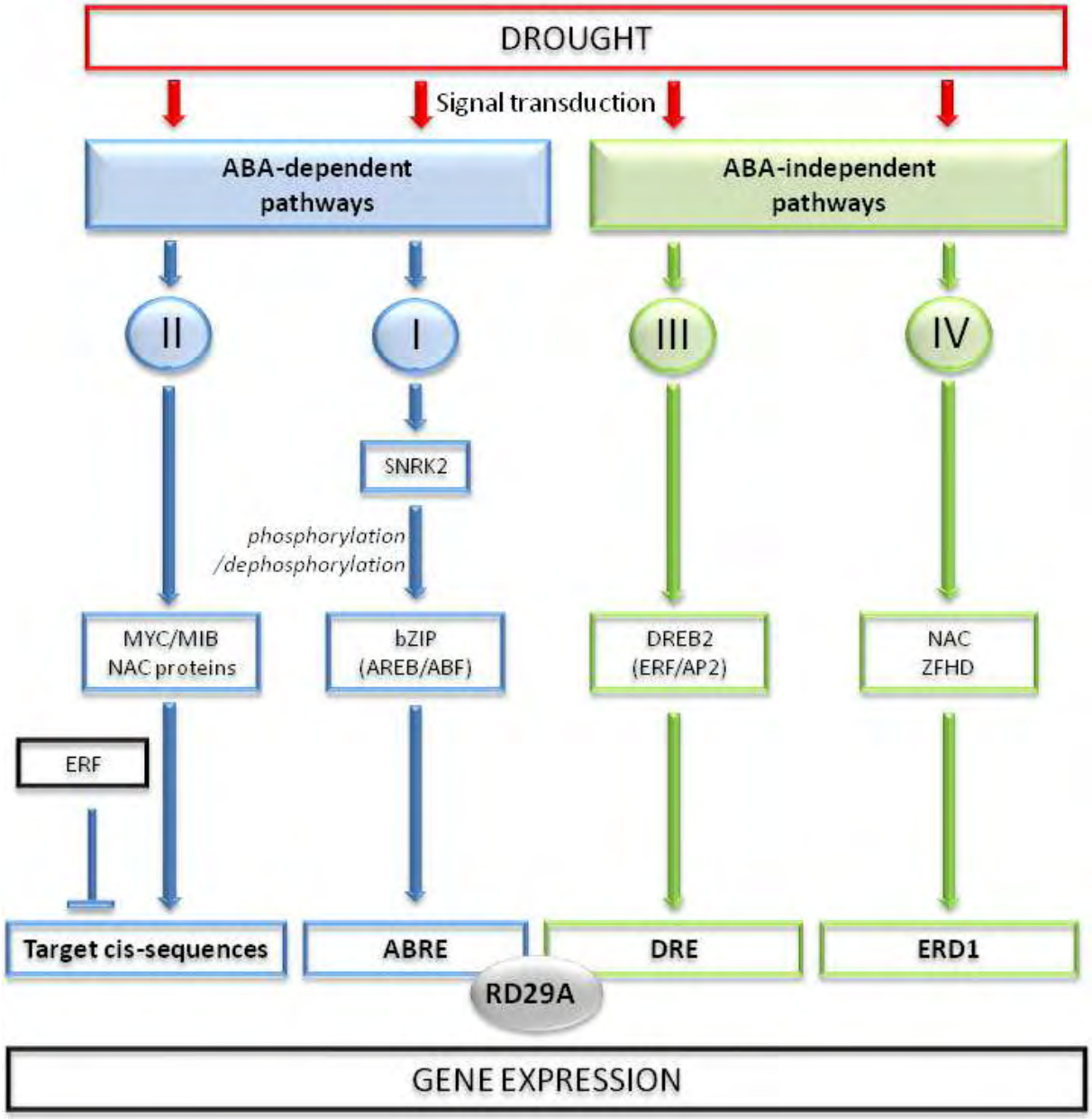


Figure I.6: Four regulatory pathways for drought responsive genes.

Pathways I and II are ABA-dependent. Pathways III and IV are ABA-independent. ERF: Ethylene Responsive Element; SNRK2: type-2 SNF1-related protein kinase; AREB: ABRE-binding protein; ABF: ABRE-binding factors; ABRE: ABA Responsive Element; DREB2: one DRE-binding protein; DRE: cis-acting element with the conserved sequence TACCGACAT; ERD1: early response to dehydration1. RD29A is a drought responsive gene with ABRE and DRE cis-sequences and therefore induced by ABA-dependent and ABA-independent pathways I and III. (Adapted from Shinozaki & YamaguchiShinozaki, 1997; Yamaguchi-Shinozaki & Shinozaki, 2006)

Besides cross-talks between ABA and non ABA pathways during drought stress regulation, some interactions with other hormonal pathways are very likely to happen. An example is the interaction between ABA and ethylene pathways. It has been shown that ethylene can antagonize drought and ABA-induced stomatal closure (Wilkinson *et al.*, 2012). On the contrary, jasmonate is, like ABA, a positive regulator of stomatal closure (Zhu *et al.*, 2012). Again, cytokinins are a class of plant hormones that are known to prevent leaf senescence (Davies *et al.*, 2005) and therefore help maintain photosynthetic activity during drought stress. Brugiere *et al.* (2003) demonstrated that expression of genes coding for cytokinin oxydase and enzymes involved in cytokinins degradation were induced by ABA. Antagonistic signal between ABA and brassinosteroid have also been recently demonstrated. Effectively ABA slows leaf expansion rates during a water soil deficit event (Tardieu, 2013), whereas brassinosteroid biosynthesis promotes leaf cell division and expansion (Zhiponova *et al.*, 2013). However, in most of the cases, knowledge on how the different hormonal pathways interact is lacking.

As for hormonal signals, direct environmental signal is transduced by a various set of genes such as those encoding for calmodulins, G-proteins, protein kinases and transcription factors. This is the case for example, with the *Arabidopsis* genes *AtCDPK1* and *AtCDPK2* (Calcium Dependent Protein Kinases), which are rapidly induced by drought and therefore are involved in the transduction cascade under drought stress. Another example is the genes involved in the MAPK (Mitogen-activated protein kinase) cascade.

I.5.3. Generic genes expression pathway during a drought stress event

Finally, we have seen that various genes are induced during drought stress: genes encoding for effectors proteins that are involved in cell protection against water-deficit damages and genes involved in signal transduction and regulatory pathways. Main regulatory pathways linking all these genes begin to be understood thanks to studies conducted mainly in the plant model *Arabidopsis*. However, in general, relationships between drought stress inducible genes remain largely unknown.

From the analysis of the various drought responsive genes described above, we can however draw a generic pathway of genes involved in responses to water stress and more generally in response to abiotic stresses (Shinozaki & YamaguchiShinozaki, 1997; Wang *et al.*, 2003). Figure I.7 illustrates this cascade. We can class genes in different groups. The first group brings together genes acting as the receptor of environmental signals. Mechanisms and genes involved in drought perception are not clearly known. Several hypotheses can be raised involving osmosensors, an oxidative burst or a change in cytoskeletons tension that could trigger the MAPK cascade and signal transduction. However their functioning in water stress perception by the cell is not entirely demonstrated (Shinozaki & Yamaguchi-Shinozaki, 2007). After the stress perception by the receptor

genes, the environmental signal is transduced by a second class of genes or molecular components (see figure I.7). As described before, transduction of the environmental signal involves secondary messengers such as Ca^{2+}, phosphorylation cascades and/or plant hormones. In a third step, the signal transduction leads to a regulatory network that controls gene transcription. Examples of ABA-dependent and ABA-independent regulatory pathways have been largely described in the previous sections. The fourth class represented in figure I.7 comprehends genes coding for effectors proteins involved in cellular and molecular drought cell responses such as the dehydrins and the aquaporins presented previously. Finally, this generic pathway, triggered by drought stress, results in various physiological and morphological responses to water deprivation that can be read into cellular, tissue and plant phenotypes.

It appears that gene expression corresponds to an important link between the environmental signal perception and the morphological and physiological responses that confer drought stress tolerance. Therefore the study of the gene regulatory networks (GRN) and their cross-talk appears to be a main goal in order to clearly understand drought stress tolerance. Systems biology approaches could be interesting and provide a better knowledge of the implication of the different signaling pathways (Ahuja *et al.*, 2010). Indeed, these approaches at the systems level permit to examine the structure and dynamics of the cellular and organismal functions instead of studying the characteristics of isolated parts of a cell or of an organism (Kitano, 2002) such as the work presented previously from *Arabidopsis* studies. The application of this new strategy has been allowed by the recent and simultaneous progresses in genotyping technologies on one hand, transcriptomic tools (such as microarray technology) on the other hand, and high-throughput phenotyping.

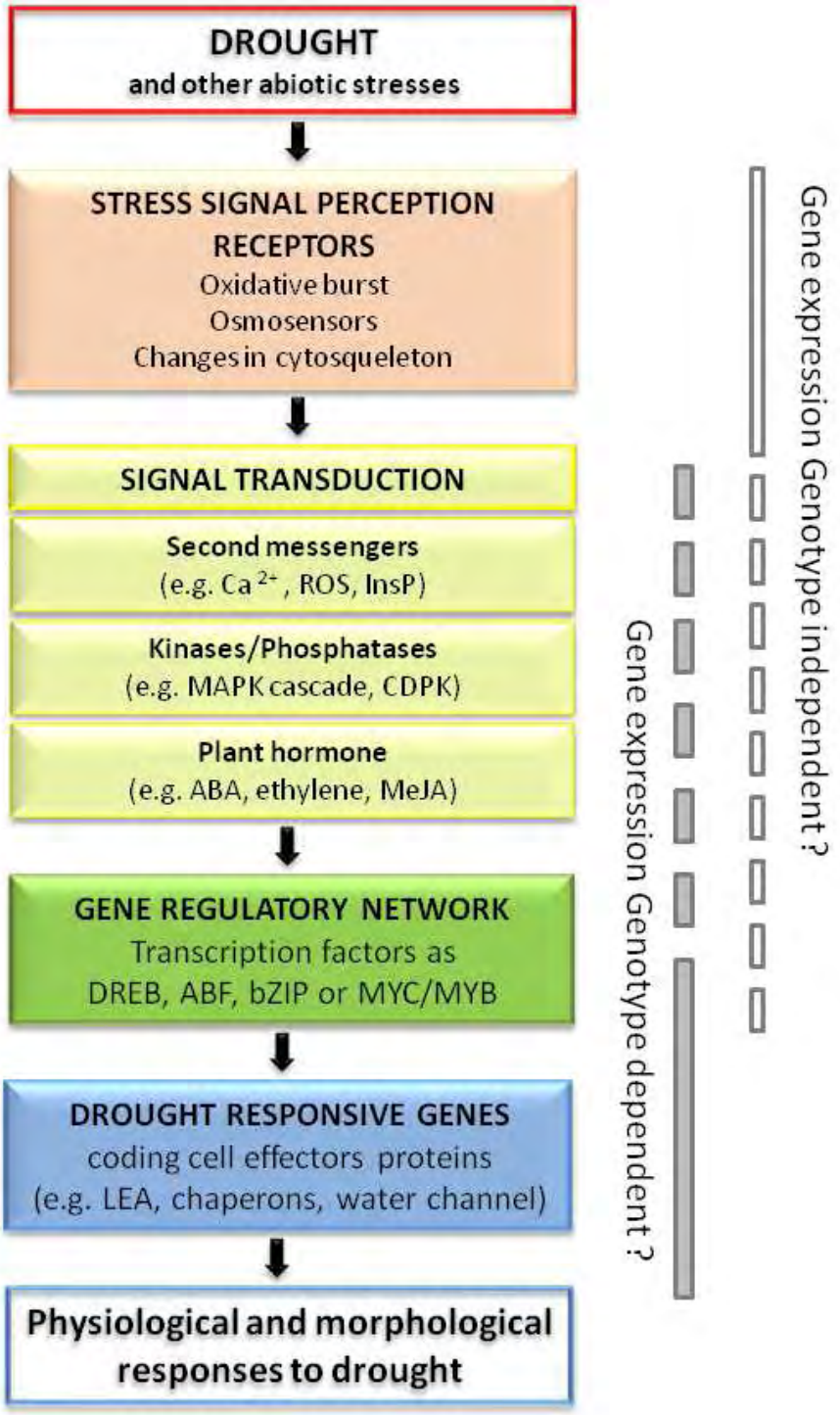


Figure I.7: Generic pathways for plant responses to drought stress (adapted from Huang *et al.*, 2012).

Ca^{2+}: calcium; ROS: Reactive oxidative species; InsP: Inositol Phosphate;CDPK: Calcium dependent protein kinase; MeJA: Methyl Jasmonate; LEA: Late Embryo genesis-Abundant

I.6. Genotype x Environment interactions during drought stress responses

In the previous sections, we have seen that the plants are able to adapt to various environments, as for example, environments with limited water resources. It is done by changing their phenotype, a phenomenon called “phenotypic plasticity” (El-Soda *et al.*, 2014) and that allows implementation of tolerance mechanisms for drought stress. Genotypes do not have identical phenotypic responses for the same environmental constrains. This can be clearly demonstrated by the comparison of a drought-sensitive genotype with a drought-tolerant genotype but also of different drought tolerant genotypes which have different strategies. Although, the plant phenotype is dependent on the genotype and on the environmental factors, phenotypic plasticity itself also depends on the genotype, i.e two genotypes do not present the same variation of phenotypes between two environments (Des Marais *et al.*, 2013). This last phenomenon is called genotype x environment interaction (GxE interaction) and can be identified by the variation of the reaction norms. Reaction norms are graphical representation of phenotypes expressed by a genotype under varied environmental conditions. Figure I.8 illustrates the different cases that reaction norms could highlight: phenotypic plasticity only and phenotypic plasticity with different types of GxE interaction models.

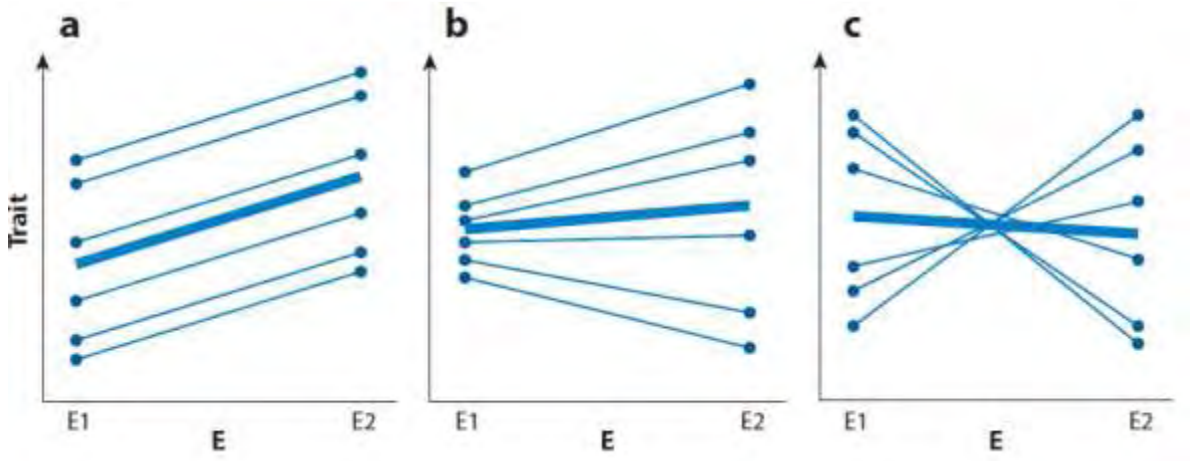


Figure I.8: Reaction norm plots for various patterns of phenotypic plasticity (Des Marais *et al.*, 2013)**.**
Thin lines show the plastic response of a single genotype, whereas the thick line represents the population average plasticity. The x-axis represents different environmental conditions and the y-axis represents the trait of interest; the series represent different genotypes. (a) Plasticity without genotype x environment interaction (GxE); (b) plasticity with variance changing GxE; (c) plasticity with rank changing GxE.

As described above, drought stress responses run under a complex genetic control involving various genes that interact between them and whose expressions depend on the environment. Another important aspect of this gene regulatory network is the GxE interaction effects that play a role in the variations of these gene expressions. Up to now, molecular geneticists have studied GxE interactions using the traditional tools of forward and reverse genetics and the evaluation of condition-dependent mutants (Des Marais *et al.*, 2013). This has led to the identification of important signaling pathways for key environmental interactions, as the drought regulatory network discussed in the previous section, and the establishment of hypotheses about crosstalk and pleiotropy of responses across various environmental signals (Todaka *et al.*, 2012). However, many questions remain concerning GxE interactions, their different patterns (illustrated in figure I.8) and their underlying genetic control.

Different genetic architectures cause GxE interactions (Des Marais *et al.*, 2013). First, a change in phenotypic rank between two genotypes can be interpreted as a genetic trade-off through antagonistic pleiotropy. It means that an allele may have an additive effect that increases the phenotypic trait value in one environment and decreases it in another (figure I.9.A). In a second case called differential sensitivity, the magnitude of the allelic effect on phenotype depends on the environment (figure I.9.B). Conditional neutrality is a particular case of differential sensitivity: an allele has a phenotypic effect in one environment and no effect in another (figure I.9.C). Finally, GxE interactions can be also due to a various range of non-additive effects. Among these non-additive effects, we can cite dominance, epistasis, genetic linkage and epigenetics (El-Soda *et al.*, 2014).

Knowledge of GxE interactions of genes involved in the drought regulatory network is important. It can help refine the interaction between genes involved in those responses to the environment. It can also be useful for breeders if they want to select for tolerance in a precise environment (therefore GxE interactions will be important) or if they search for an ideotype tolerant in a various range of environmental scenarios.

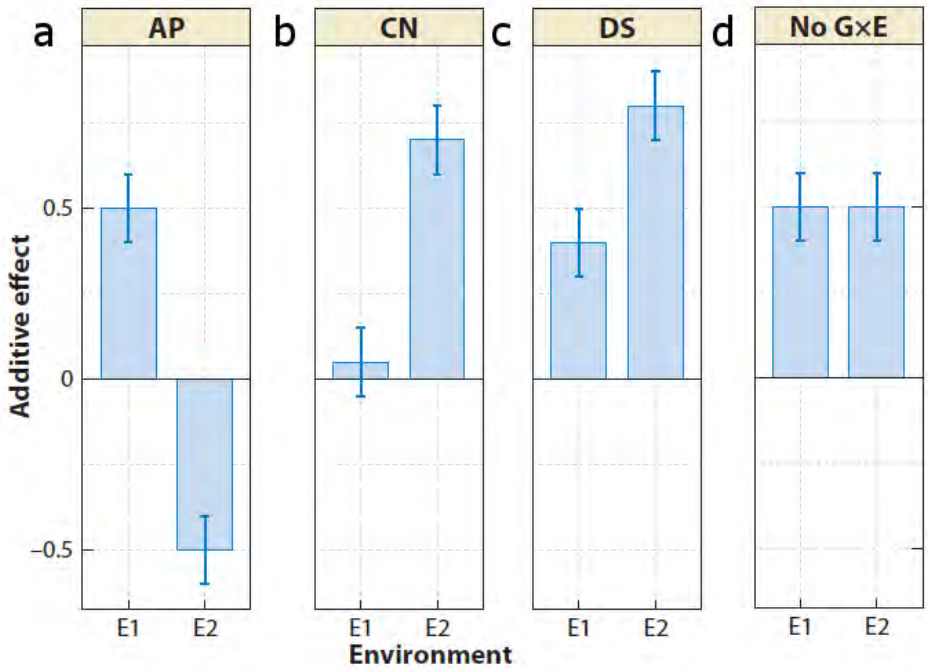


Figure I.9: Patterns of quantitative trait loci additive effects for trait that show genotype x environment interactions (GxE) can fall into four main categories (Des Marais *et al.*, 2013).
(a) Antagonistic pleiotropy (AP): the result of sign changing additive effects; (b) Conditional neutrality (CN): additive effects limited to only specific environmental conditions; (c) Differential sensitivity (DS): the results of changes in magnitude of additive effects; (d) no GxE: no detectable change in additive effects across environments.

I.7. Sunflower and drought stress

I.7.1. Sunflower crop and interest for drought tolerance breeding

Sunflower is a widespread crop cultivated whose products are used in human alimentation (oil and margarine) but also in bio-refinery (i.e. in cosmetics). Sunflower oil is appreciated for its nutritional qualities (vitamin E, tocopherol, omega9 content) (Cetiom, 2014). The sunflower cake is rich in proteins and rare amino-acids (such as methionin) and is therefore very valuable for animal feeding. The development of bio-fuels from sunflower oil was made possible by the recent development of high oleic varieties but its price is not competitive compared to other vegetable oils.

Worldwide industrial use of sunflower is recent. Global sunflower seeds production in world was of 35,568 thousand tons in 2013. Areas of sunflower crop production are largely located in temperate zones. The main producer countries are Ukraine, Russia, European Union, Argentina, and China with respectively 9.8, 9.3, 8.6, 2.9 and 1.7 Mt of seeds harvested in 2013 (National Sunflower Association, 2014).

Sunflower is often reported as a drought tolerant species (Hall *et al.*, 1990). Indeed its root system is efficient to extract soil water. Hence, in southern Europe sunflower crops suffer from a low rainfall and are often cultivated in area with shallow soil. These last 50 years, efficient breeding

efforts allowed the increase of sunflower production. For instance, in Europe the average sunflower yield increased of 6 qx/ha. In France the sunflower yield was 17 qx/ha in 1961 and then reached 23 qx/ha in 2012 with a peak of 25 qx/ha in 2008 (Figure I.10). Despite these progresses and the genetic improvement of sunflower varieties (Vear *et al.*, 2003), there is still a difference between the potential yield and the real ones obtained in the field. Among several factors, including disease attacks, this is most likely due mainly to limited water availability during the yield elaboration. This is why drought tolerance still remains one of the main targets for sunflower breeders.

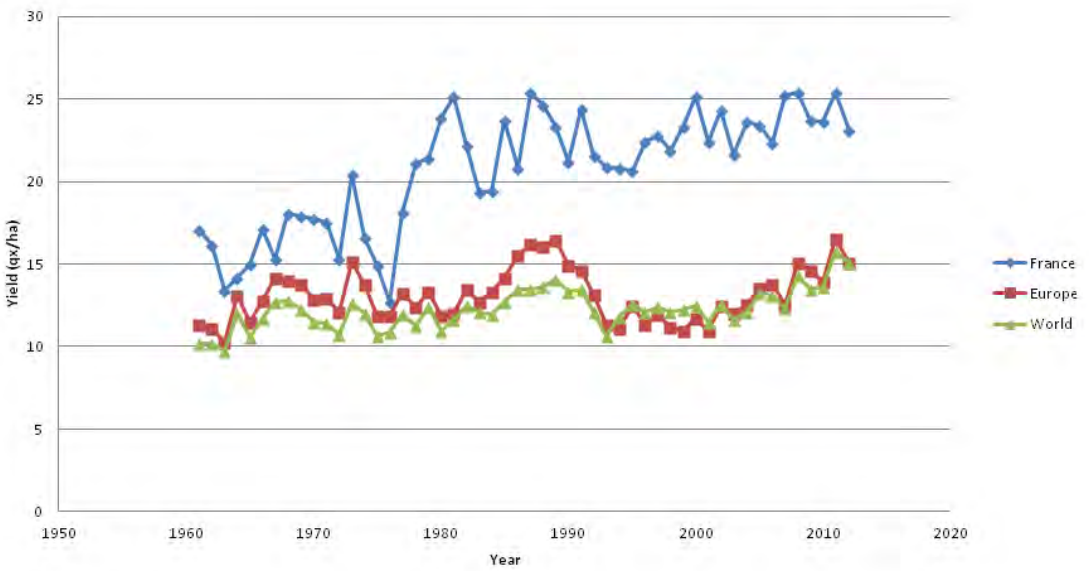


Figure I.10: Evolution of sunflower yield since 1961 to 2012 (from FAOstat, 2014).
Blue curve refers to the average yield in France. Red curve refers to the average yield in all European countries. Green curve refers to the average yield in the world.

I.7.2. Sunflower germplasm adapted to various environments with strong drought constrains

Helianthus belongs to the *Asteraceae* family and includes 62 species with native distributions covering a large part of North America (website: USDA, Natural Resources Conservation Service, 2014). It is a group with a recent history of evolution combining hybridization and polyploidy events. Therefore the phylogenetic reconstruction of this group has been a challenging work for scientists during the last century. In 1981, Schilling and Heiser used extensive crossability information and morphological characters to divide non perennials species in 4 sections: *Helianthus*, *Agrestis*, *Divaricati* and *Ciliares*. To date, the phylogeny of *Helianthus* with the best resolution was generated by analyzing sequence data from the external transcribed spacer of the 18S-25S nuclear ribosomal DNA region (Timme *et al.*, 2007). Figure I.11 represents the hypothetical phylogeny summarized by Kane *et al.* (2013).

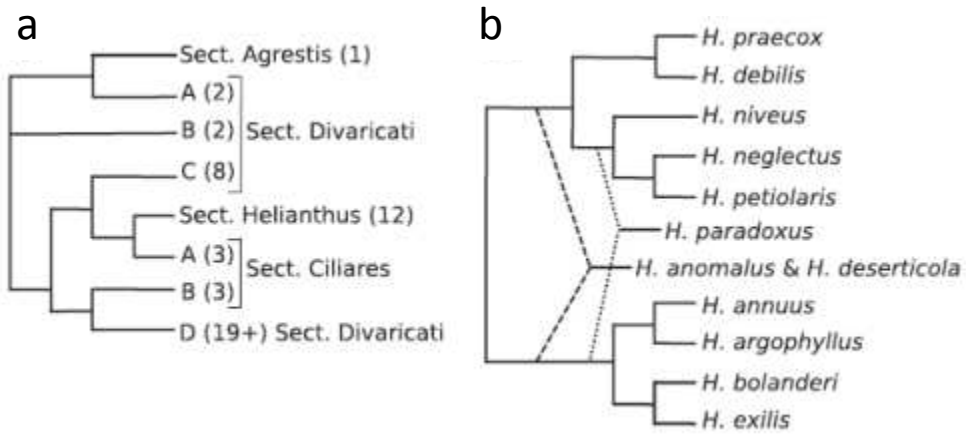


Figure I.11: Phylogenetic trees for *Helianthus* (Kane *et al.*, 2013).
(a) Phylogenetic tree for sections of the genus. Number of species in each clade are given in parentheses following the section name. Sections *Ciliares* and *Divaricatus* are polyphyletic. (b) Phylogenetic network for section *Helianthus*. Putative hybrid speciations are indicated in dashed lines.

The section *Helianthus* comprehends 12 species that occupy a diverse range of habitats. For example, *H. argophyllus* is native to the dry, sandy soils in Southern Texas, an arid environment that imposes strong selection for tolerance to drought stress (Seiler & Rieseberg, 1997). *Helianthus annuus* is the most widespread species of *Helianthus* and is adapted to numerous habitats from Mexico to Canada (Rieseberg *et al.*, 1999). It is sympatric to the *H.petiolaris* species. Hybridization between these two last species occurred three times during *Helianthus* evolution, creating the three hybrid species *H.anomalus*, *H.deserticola* and *H.paradoxus* particularly adapted for sand dunes, desert floors and salt marshes respectively (Rieseberg *et al.*, 1999). Hybridization as described above allows a large distribution of *Helianthus* and a better adaptation to specific environments with various and strong constrains for drought tolerant traits. Therefore, wild sunflower germplasm is a potential source of abundant genetic resources for drought tolerance.

On the contrary, due to its breeding history sunflower elite lines show a narrower genetic variability than wild species. The sunflower family (*Asteraceae*) comes from North America (Stuessy, 2010). The origin of its domestication has been a subject of controversy for a long time, and evidence would place the origin of domestication either in Mexico (Lentz 2008) or in eastern North America (Heiser 2008) 6300 years ago. More recently, evidence from multiple evolutionarily important loci and from neutral markers supports a single domestication event in eastern North America (Blackman *et al.*, 2011). Sunflower was introduced in Europe by the Spanish during the 16th century mainly as ornamental flower. It is only at the beginning of the 19th century that important breeding programs have been developed in Russia in particular to improve oil production thanks to the work of

(Pustovoit, 1964). Until the middle of the 20th century, sunflower production was based on population variety. The discovery of the cytoplasmic male sterility (CMS) (Leclercq, 1969), then of the fertility restoration genes (Kinman, 1970) led, since 1978, to the spread of hybrid sunflower crops throughout Europe. On the one hand, hybrid culture has allowed important genetic progresses in productivity and tolerance to diseases. However, on the other hand breeding history of cultivated sunflower has led to a reduction of the genetic diversity. Utilization of wild sunflower lines could be an important asset in breeding programs where wild alleles of tolerant species can be introgressed in elite lines in order to improve their drought tolerance. Moreover, the wild sunflower germplasm, described above, constitutes an adequate object of study that can help decipher the different mechanisms allowing drought tolerance.

I.7.3. Sunflower morphological and physiological responses to drought stress

The main morphological and physiological responses to drought stress previously reviewed were discovered and described in model plants and other crops (such as rice) and are generally valid in all higher plants such as sunflowers. However, because of the interest of the sunflower community in developing drought-tolerant genotypes and understanding ecological adaptation of wild sunflower species, some knowledge was gathered specifically on the species' responses to water deficit. Indeed, the sunflower has a strong capacity to extract and conduct water from soil to leaves, although in the meantime its water consumption is important due to a high photosynthesis potential and stomata location on both leaf surfaces (Herve *et al.*, 2001).

Few studies report specific sunflower responses to drought stress. For example, (Casadebaig *et al.*, 2008) observed leaf expansion and transpiration rates of different sunflower genotypes in response to soil water deficit measured through the Fraction of Transpirable Soil Water (FTSW). Leaf expansion of sunflower decreases for FTSW values inferior to 0.6 and transpiration rate diminishes for FTSW values inferior to 0.4. Moreover, these two variables show different thresholds for their decreasing in response to drought depending on the genotype. This highlights two different strategies among sunflowers for drought stress responses. The first strategy called *conservative* distinguishes sunflower genotypes that reduce leaf expansion and close stomata when water deficit is still low. Genotypes adopting this strategy can keep high water content to the detriment of biomass production. This *conservative* strategy can be related to the *avoidance* strategy described above that has for goal to maintain the water content. On the contrary, genotypes of the second strategy called *productive* maintain leaf expansion and stomatal aperture even for low values of FTSW and therefore favor biomass production over water content. These two strategies are not efficient in the same drought scenarios: the first is better adapted for environments with long and severe drought events and the second works better in environments with short and moderate water

deficit. Leaf development under water deficit has been particularly studied because leaf area is important for radiation interception and therefore biomass synthesis and final yield (Aguirrezabal *et al.*, 2003). However, a reduced leaf area can lead to a lower transpiration rate and a better tolerance to water deficit. In sunflower, it has been demonstrated that cell expansion and division during leaf development is affected by water stress (Tardieu, 2013) and the effects are more important if drought occurs during the early stage of leaf development when cell division happens (Granier & Tardieu, 1999).

I.7.4. Molecular and genetic responses to drought stress in sunflower

The molecular and genetic architecture controlling the different drought tolerance strategies are far from being well described and understood in most plants. Sunflower is a case in point. The main molecular responses described for the plant model *Arabidopsis* or other crops could be inferred to sunflower. However, there is no species close enough to sunflower to infer the genetic architecture through synteny. Several studies attempted to describe the genetic basis of physiological traits associated with drought tolerance in sunflower. For example, quantitative trait loci (QTL) were found for Relative water content (RWC), water potential and stomatal conductance using a RIL population under drought conditions in greenhouse (Kiani *et al.*, 2007b). Several gene expressions were found to be correlated to different physiological variables used to estimate drought tolerance: stomatal conductance, osmotic adjustment, RWC, Carbon Isotopic Discrimination (CID) (Kiani *et al.*, 2007a; Kiani *et al.*, 2007b; Rengel *et al.*, 2012). However, a complete view of the relationships between those genes and detailed mechanisms of their role in drought tolerance is still to be determined. Plant hormones in sunflower, as in *Arabidopsis*, seem important in signal transduction and gene regulatory network for drought stress responses. The sunflower HD-Zip protein gene HaHB4 was shown to be induced by water stress and ABA treatment. Moreover, a DRE motif was identified in its promoter suggesting that an ABA-dependent pathway involved in drought response signaling is similar between *Arabidopsis* and *Helianthus* (Dezar *et al.*, 2005). It shows evidence that at least a part of regulatory mechanisms involved in water deficit tolerance is conserved between these two species. In another study, modified *Arabidopsis* plants over-expressing HaHB4 were found to be less sensitive to external ethylene treatment. Identification of the potential target of HaHB4 revealed genes related to ethylene synthesis and ethylene signaling suggesting that HaHB4 may improve drought tolerance through the control of leaf desiccation. Therefore, in sunflower, a crosstalk between ABA and ethylene pathways seems to be involved in plant responses to drought events (Manavella *et al.*, 2006).

These works at the genetic and molecular levels start to unveil the complex networks of molecular sensing, signaling and responding to drought stress in sunflower. Together with the

evidences collected in other plants, they suggests a complex system, whose mechanisms are challenging to understand and thus to harness in order to develop more tolerant sunflower varieties in the future.

I.8. Objectives of the PhD works:

Plants are sessile organisms and thus have to cope with the pressures of their environment. Among these constraints are abiotic stresses and in particular drought stress that occurs when water supply is not sufficient to compensate water losses due to evapotranspiration. Throughout this introductive chapter, we have seen that plants develop complex responses to drought in order to become tolerant to water deficit. These responses involve not only morphological and physiological mechanisms that occur at the whole plant scale but also molecular events at the cell level. Genetic control of these responses engages various genes. Up to now, thanks mostly to studies on model plant as *Arabidopsis*, some key regulatory pathways have been described such as, for instance, the ABA-independent pathways described above. A generic pathway of genes involved in drought stress responses can be drawn from these results (see section I.5.3 and figure I.8). It identifies various classes of genes that allow the environmental signal perceived (here drought stress) to be linked to morphological and physiological mechanisms that allowed water deficit tolerance. However, this schematic representation does not translate the complex underlying gene regulatory network that involves many genes, not always identified, interacting between them and with the environment. A better understanding of this network could explain the differences of behavior between species under limited water supply and also differences of phenotypes between tolerant and sensitive genotypes among the same species. It could also help to understand the genetic control of drought tolerance mechanisms observed at the whole plant level and clarify the genetic processes of environmental signal perception by the plant.

The aim of this PhD work is to study the gene regulatory network that leads to morphological and physiological mechanisms developed by plants in order to cope with water deficit. We propose to use the cultivated sunflower *Helianthus annuus* as object of our studies because it has been shown that drought stress is one of the major issues that impact yield for this crop. Therefore breeding for drought tolerance is still one of the main goals in the sunflower selection programs. Moreover, due to the history of its evolution, the genus *Helianthus* offers a wide germplasm interesting for drought tolerance studies.

To achieve our goal, we chose to articulate our work on the generic pathway described in figure I.8 and successively study the different classes of genes identified in this gene cascade.

The first category of genes that we would like to study is composed of genes involved in environmental signal reception. Indeed, among the questions still opened, stands the perception of drought stress by the plants, or in other words the description of the mechanisms that trigger genes involved in the regulation of drought stress responses. Practically, the knowledge of this class of genes could lead to a better estimation of the drought stress perceived by the plant and help with genotype comparisons. One can also wonder if the expression of this kind of genes is dependent of the genotype or if general sunflower mechanisms can be identified. Therefore, the first part of our work describes the identification of genes whose expression is correlated to plant water status. We then discuss their place in the generic pathway for drought responses and how they can improve our understanding of the dialog between the plant and its direct environment.

In a second time, we will identify genes involved in the genetic control of physiological mechanisms for drought stress tolerance. These genes are represented at the end of the generic pathway for water deficit responses (figure I.8). As the behavior regarding drought stress is not identical between genotypes, the expression of those genes depends on the genotype, the environment, and also on the interaction between genotype and environment. The gene regulatory network formed by those genes can help understanding the phenotypic plasticity observed among genotypes of the same species. Thus, another aim of this work is, through an association study, to reconstruct the underlying gene regulatory network and the pattern of the genetic control due to genotype x environment interactions.

Finally, in the last section we will study genes involved in the environmental signal transduction and transcriptional regulation parts. Through the study of these genes and of the gene regulatory network they belong to, we propose to answer two questions. First, how this GRN is involve in phenotypic changes that allow drought stress tolerance and second, how the specific design of the inferred drought GRN could have played a role in *Helianthus* evolution and sunflower breeding? We will try to answer these questions with a systems biology approach that allows us to reconstruct the gene regulatory network and organize relationships between genes involved in the environmental signal transduction.

Chapter II: The expression of genes possibly involved in the perception of the drought stress signal can help to characterize the plant water status via a biomarker construction

II.1. Challenges and issues in the studies of genes receptors of the drought environmental signal

Responses to drought stress are triggered by plant water deficit perception. Then, the understanding of the mechanisms of environmental signal reception and the genes underlying them is an important goal in the process of drought tolerance study. In this chapter, we propose to focus on those genes which are possible receptors of the stress signal or tightly connected to the environmental signal in the regulatory cascade leading to water deficit responses (Figure II.1). Several genes that suit this definition, i.e. with expression levels correlated only to the applied water stress intensity but not to major physiological and morphological plant responses to drought, have already been identified in sunflower (Rengel *et al.*, 2012). Additional knowledge for this type of genes could lead to a better estimation of the water status perceived by the plant. So, in this chapter we describe the construction of a biomarker for plant water status based on the expression of genes involved in perception of water deficit environmental signal.

The term biomarker was first used in the field of human medicine and therefore defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Strimbu & Tavel, 2010). Thereafter, various biomarkers have been developed in different scientific fields and then can be defined more generally as an indicator of specific or general stresses. They concern different biological levels, from molecules to ecosystems. Depending of the studied biological object, a biomarker could be a biochemical, physiological or a morphological change that can be used as a proxy for the tested environmental variable (Ernst & Peterson, 1994), for instance water stress. Such biological indicators are now widely used in human cancer research. For example, beta-galactosidase concentration was found as a biomarker for aging cells in the skin and therefore allows distinction between healthy tissues and tumor (Dimri *et al.*, 1995). Biomarkers were also used more recently to characterize ecosystems in the field of ecology. In this case, microbial biomarkers were used as an indicator of ecosystem recovery following surface mineral exploitation (Mummey *et al.*, 2002).

Biomarkers based on gene expression levels are now available thanks to progresses in high throughout transcriptomic technologies and meta-analyses of these large transcriptomic datasets.

A tool as a biomarker for sunflower water status would be useful in genetic studies that involve an important number of genotypes studied in field conditions with variable water availability. Moreover, in these studies, each genotype, due to its specific development and physiology, have a specific water status and drought stress perception, even if other environmental variables are identical for all the genotypes. Therefore, one major issue of our study is to develop a practical tool that can be used for a range of sunflower genotypes. That is why, genes introduced in such a biomarker should have an expression not only correlated to water stress intensity but also independent of the considered genotypes. In addition, the number of genes used should be parsimonious in order to obtain a tool easy to exploit.

The present chapter describes the identification of such genes and the water status biomarker construction. From results of this work, we also have tried to address different issues about genes involved in environmental signal perception. Where in the regulatory pathway is the location of (1) the genes involved in the perception of the environment and (2) the genes with genotype-independent expression (i.e candidate genes for the biomarker construction)? Could knowledge about the latter genes' distribution give us some insight about which part of the regulatory cascade is genotype-dependent or independent? Indeed, genes located upstream the genotype independent genes are supposed to be genotype-independent as well.

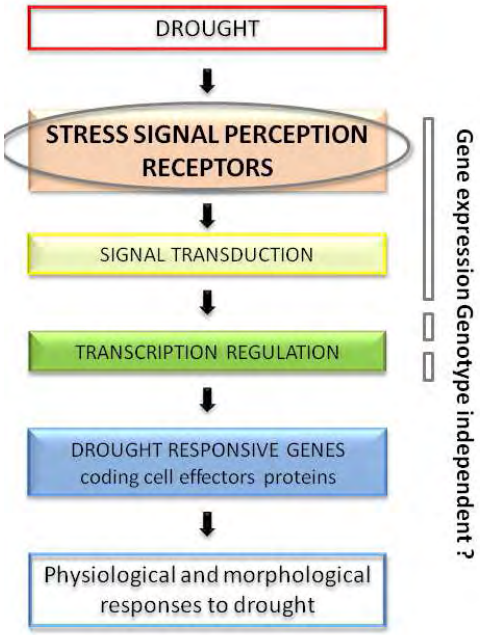


Figure II.1: Genes studied for the water status biomarker construction and hypothesis about their gene expression independent of the genotype.

II.2. Article: A biomarker based on gene expression indicates plant water status in controlled and natural environments

Article status: published in Plant Cell and Environment (December 2013)

List of authors

Gwenaëlle Marchand [1,2], Baptiste Mayjonade [1,2], Didier Varès [1,2], Nicolas Blanchet [1,2], Marie-Claude Boniface [1,2], Pierre Maury [3,5], Fety Andrianasolo Nambinina [4,5], Philippe Burger [5,6], Philippe Debaeke [5,6], Pierre Casadebaig [5,6], Patrick Vincourt [1,2,7], Nicolas B. Langlade [1,2,7]

Laboratories of origin

1 INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, F-31326 Castanet-Tolosan, France

2 CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, F-31326 Castanet-Tolosan, France

3 Université Toulouse, INPT ENSAT, UMR1248 AGIR, F-31320 Castanet-Tolosan, France

4 CETIOM, Centre INRA de Toulouse, F-31320 Castanet-Tolosan, France

5 INRA, UMR1248 AGIR, F-31320 Castanet-Tolosan, France

6 Université Toulouse, INPT, UMR AGIR, F-31029 Toulouse, France

7 These authors contributed equally to this work.

II.2.1. Abstract

Plant or soil water statuses are required in many scientific fields to understand plant responses to drought. Because the transcriptomic response to abiotic conditions, such as water deficit, reflects plant water status, genomic tools could be used to develop a new type of molecular biomarker.

Using the sunflower (*Helianthus annuus L.*) as a model species to study the transcriptomic response to water deficit both in greenhouse and field conditions, we specifically identified three genes that showed an expression pattern highly correlated to plant water status as estimated by the pre-dawn leaf water potential, fraction of transpirable soil water, soil water content or fraction of total soil water in controlled conditions. We developed a generalized linear model to estimate these classical water status indicators from the expression levels of the three selected genes under controlled conditions. This estimation was independent of the four tested genotypes and the stage (pre- or post-flowering) of the plant. We further validated this gene expression biomarker under field conditions for four genotypes in three different trials, over a large range of water status, and we were able to correct their expression values for a large diurnal sampling period.

II.2.2. Keyword index

Soil water deficit, sunflower, *Helianthus annuus L.*, transcriptomic, drought, biomarker, leaf water potential, FTSW, soil water content, indicator.

II.2.3. Introduction

Water deficit in plants can be defined as the imbalance between the actual evaporative demand resulting from climatic conditions and the available water in the soil (Tardieu *et al.*, 2011) This major environmental stress affects the growth and physiology of the entire plant and can therefore dramatically reduce crop yield and quality (Bhatnagar-Mathur *et al.*, 2008). Recent intensification of drought events in Europe, Australia and North America, together with climatic model forecasts, suggest that drought will continue to dramatically affect crop productivity in the 21st century (Moriondo *et al.*, 2010). At the same time, the reduction of arable land area, the scarcity of water resources and the development of the human population all amplify the need to develop agrosystems that are more tolerant to water deficit or less water-consuming.

In this context, plant molecular physiologists, eco-physiologists and agronomists conduct experiments to estimate the plant response to water stress. One major requirement of these experiments is the estimation of soil water available to crops, at least at the most critical points of the developmental process. The direct estimation of water accessible to individual plants can be very difficult or even impossible in natural environments such as the field. Water status indicators can be of two types: soil- or plant-based measurements.

Soil-based measurements use either thermogravimetry, which requires samples for over-drying, or physical measurements of soil properties varying with its water content (Dobriyal *et al.*, 2012). Neutron probes have been widely used since the 1960s (reviewed by Gardner *et al.*, 1991; Klenke & Flint, 1991), and dielectric methods, such as time-domain reflectometry (Topp & Davis, 1985) or capacitance sensors (Whalley *et al.*, 1992), have been used since the 1980s. Generally, these techniques can only describe limited soil regions that may not correctly represent the plant rhizosphere (Ferreira *et al.*, 1996). The access tube used for the probe measurement is difficult to install and often modifies the water circulation and root dynamics surrounding the tube. Furthermore, these tools are time-consuming and labor-intensive and cannot be scaled up for high-throughput evaluation, which is needed in genetic analyses.

Therefore, plant-based measurements are often preferred. These measurements are based on the fact that the plant status reflects soil water availability. Morphological indicators can be used to evaluate drought stress. For example, breeders often score leaf rolling to estimate plant water status in monocots (O'Toole *et al.*, 1979). For perennial species, trunk diameter reflects water fluxes

in the plant and can be used to manage irrigation (Goldhamer & Fereres, 2001). In eco-physiological studies, different indicators of plant status are commonly measured at different organizational levels, such as whole-plant transpiration, leaf water potential or stomatal conductance. For example, water transpiration cools the leaves and can therefore be monitored through thermal infrared measurements such as in wheat (Blum *et al.*, 1982). The measurement of pre-dawn leaf water potential (Ψ_{PD}) has been largely used for decades and is considered to be a standard. It is based on the fact that at dawn, water equilibrates between the rhizosphere and leaves, reflecting the water available to the plant. Therefore, it is subject to some limitations for heterogeneous soils (Ameglio *et al.*, 1999) and its determination in numerous plants is restricted by the need of operating at pre-dawn.

Another method that is successfully used in controlled environments is based on the measurement of the daily transpiration of water-stressed plants relative to well-watered controls. This method defines a scale for water available for plant transpiration between an upper and lower limit of soil water content (SWC). The upper limit matches the SWC at field capacity and the lower limit is the SWC where relative transpiration decreases to less than 0.1 (Sinclair, 1986). According to (Sinclair, 2005), plants respond to the progressive drying of soil in a similar manner across a wide range of environmental conditions when this scale is used, where water stress is expressed as the fraction of transpirable soil water (FTSW). Variability in the leaf expansion and plant transpiration rate in response to water stress has been previously reported in different sunflower genotypes by using this method (Casadebaig *et al.*, 2008). When dealing with field conditions, several main drawbacks limit the applicability of this method. First the need for a control plot (in order to measure transpiration in well-watered conditions) doubles the experimental space required, which thus precludes the use of this indicator for large-scale genetic programs in crops. More importantly, to estimate FTSW in field environment, measurements of soil depth (soil profile, endoscopy), portion of soil explored by roots and soil water content (probes, soil cores) are required but are often unrealistic for genetic studies at microplot scale or poorly estimated.

In the context of the development of high-throughput phenotyping platforms, the need to develop a tool that would allow the early quantification of water deficit in a dose- and time-dependent manner is even more acute. Following the definition of (Ernst & Peterson, 1994) a soil water content biomarker could be a biochemical, physiological or morphological change in plants that measures their exposure to the environment (i.e., water deficit). This biomarker could therefore be used to reveal the status and trends in environmental assessment and also to predict crop responses to other biotic and abiotic stresses that interact with drought.

The transcriptomic response to water deficit is a widely described molecular process that allows plants to adapt to the water imbalance between supply and demand, and to develop a large

range of morpho-physiological changes (Shinozaki & Yamaguchi-Shinozaki, 2007). The gene regulation cascade begins from the composite molecular perception of the environmental signal (the biophysical water imbalance) and moves via signal transduction down to the level of enzymes and structural proteins to produce biochemical compounds, fluxes and developmental adaptations. In accordance, some transcript expression levels are correlated only to FTSW but not to other major plant responses (Rengel *et al.*, 2012). Such genes correspond to the definition of biomarkers that reflect soil water status. Assembling several genes robustly correlated to soil water content appears to now be an attainable goal given recent progress in the description of the transcriptomic response to water deficit at the interface of molecular biology and eco-physiology (Ingram & Bartels, 1996; Ramanjulu & Bartels, 2002; Bartels & Sunkar, 2005; Harb *et al.*, 2010; Aasamaa & Sober, 2011).

In fact, gene expression biomarker search and development is a long-standing goal of the reference meta-analysis platform Genevestigator (Zimmermann *et al.*, 2008). Recently, a successful meta-analysis of a large transcriptomic data set in maize allowed the development of a composite gene expression scoring system to quantitatively assess the response of maize to nitrogen conditions (Yang *et al.*, 2011). Importantly, this first gene expression biomarker for *in planta* nitrogen status is independent of genotype, does not vary throughout plant development and was validated in field and greenhouse conditions.

The development of such tools has certainly been hampered by the rapid variation of plant transcriptome in response to many external factors, such as illumination and handling/wounding, as well as internal factors, such as the circadian clock. However, whole transcriptomic studies now show that part of the transcriptome robustly reacts to the plant environment in a dose- and time-dependent manner, which allows statistical models to be built, notably for the sunflower (Rengel *et al.*, 2012).

In this context, we used the sunflower as a model to develop a composite gene expression biomarker that is independent of genotype, developmental stage and time of day and that allows the estimation of soil water constraint in greenhouse and field experiments. This biomarker was standardized using the pre-dawn leaf water potential, FTSW, soil water content and fraction of total soil water when available.

II.2.4. Material and methods

II.2.4.1. Plant material and growing conditions

Four experiments, i.e., one in greenhouse conditions and three in field conditions, using the four sunflower (*Helianthus annuus L.*) inbred lines XRQ, PSC8, their F1 named Inedi and another cultivated hybrid Melody were conducted in 2012 near Toulouse (Haute-Garonne, France).

For the greenhouse experiment conducted from May to June 2012, bleach-sterilized seeds were germinated on Petri dishes with Apron XL and Celeste solutions (Syngenta, Basel, Switzerland) for 3 days at 28°C. Plantlets were transplanted in 236 individual pots, and each pot contained one single plant.

Pots were filled with 15 L of a substrate composed of 10% sand, 40% P.A.M.2 potting soil (Proveen distributed by Soprimex, Chateaurenard, Bouches-du-Rhône, France) and 50% clay loam from the INRA site in Auzeville-Tolosane (Haute-Garonne, France).

Plantlets were sown on two different dates to obtain plants at two different stages (before and after flowering) respectively, 10 weeks and 4 weeks before the beginning of the stress treatment.

For each phenological (pre-flowering or post-flowering) stage with, respectively, 144 and 92 plants, the pots were arranged in a split-split-plot design with three blocks. The stress intensity (i.e., FTSW values of 0.8, 0.7, 0.5, 0.35, 0.20 and 0.12) was the main factor within the block, the genotype was the second factor, and finally the treatment (control plants were well-watered and treated plants were water-deprived) was the third factor. After an assessment of Alternaria blight evolution, plants of genotype PSC8 were not considered in the experiment dedicated to the post-flowering stage. Each pot was fertilized and irrigated as in Rengel *et al.* (2012) before the beginning of the water stress application.

In field conditions, the same genotypes were sown at three different locations: Samatan (Gers, France), Fleurance (Gers, France) and Auzeville-Tolosane. In Samatan, the four genotypes were sown on April 20, 2012 and grown without irrigation. In Fleurance, PSC8, Melody and Inedi were sown on April 6, 2012 and grown without irrigation. In Auzeville-Tolosane, Inedi and Melody were sown on May 25, 2012 and grown in both irrigated (163 mm) and non-irrigated conditions.

The field experiments were designed in six randomized blocks for each location or location*condition combination. Each plot consisted of 12 rows with a length of 10 m, 12 rows with a length of 6 m and 9 rows with a length of 5.2 m for each genotype in Samatan, Fleurance and Auzeville-Tolosane, respectively, at the same plant population density (6.5 plants.m^{-2}).

Soil analysis

An 800-g soil sample was taken at depths of 60 cm and 30 cm in each trial and sent to the INRA LAS laboratory (Arras, Pas-de-Calais, France) for physical and chemical analyses.

Water stress treatment

In the greenhouse experiment, the pots were saturated with water 31 and 73 days after germination, respectively, for pre-flowering and post-flowering plants. The following morning, excessive water was drained for two hours and pots were weighed to obtain the saturation mass.

From this point, irrigation was stopped for water-deprived (WD) plants. Both control and WD plants were weighed every day between 16:00 and 17:00 to determine the daily evapotranspiration. The water lost was added back to the control plants. To prevent soil evaporation, pots were covered with a 3-mm-thick polystyrene sheet. However, soil evaporation could not be neglected.

In the field experiments, 53, 70 and 40 mm of water were provided, respectively, on June 29, July 11 and August 13 2012 in the irrigated condition in Auzeville-Tolosane.

Soil evaporation estimation in the greenhouse

Six pots without plants that represented the different water content were also weighed every day for two weeks during the experiment. Climate conditions, such as relative humidity and average temperature in the greenhouse, were monitored daily. The soil evaporation could, therefore, be estimated by performing a linear regression with pot water content, relative humidity and average temperature using the function *regress* (MATLAB version 7.13.0.564, Statistics Toolbox 7.6). This model is detailed in Appendix II.1 and was used to estimate the soil evaporation during the greenhouse experiment.

Plant leaf area and transpiration in the greenhouse

For all plants, the length and width of odd leaves were measured every other day. The total leaf area was calculated from these measurements as described in (Casadebaig *et al.*, 2008)

The plant transpiration (E in $g.mm^{-2}$) for each pot was calculated every day as the difference between the water lost by the pot and the water lost by soil evaporation divided by the total plant leaf area.

The normalized transpiration (EN) for each WD plant was calculated every day as the ratio between its transpiration and the average transpiration of control plants of the same genotype in the same block.

II.2.4.2. Estimation of the fraction of transpirable soil water (FTSW) in the greenhouse experiment

The total transpirable soil water (TTSW) is the maximum amount of soil water available to the plant. In our experiment, 8 treated plants reached EN values less than or equal to 10% and were used to estimate the TTSW. This weight ($W_{10\%}$) corresponded to the dry soil plus 26% (w/w) of the water contained in the saturated pot. The fraction of transpirable soil water (FTSW) was finally calculated as follows:

$FTSW = (W_d - W_{10\%})/TTSW$, where W_d is the weight of the pot at day *d*.

The FTSW value was used to determine whether a plant had reached the target stress intensity.

II.2.4.3. Estimation of the soil water content (SWC) in the greenhouse experiment

At the end of the greenhouse experiment, a soil core was collected from each pot. The soil samples were weighed to obtain the fresh weight and then dried for 48 h at 120°C before a second weighing to obtain the dry weight. The soil water content (SWC in percentage w/w) was calculated daily as follows:

$SWC_j = (W_{fi}-W_d)/(W_d-W_d')$, where W_{fi} is the weight of the fresh soil and plant at day *i*, W_d is the weight of the dry soil and plant and W_d' is the weight of the dry plant.

The weight of the fresh plant at day *i* is estimated using the dry weight of the plant and based on the assumption that the plant water content was, on average, 81% for post-flowering plants and 87% for pre-flowering plants and was the same for the different plant tissues.

II.2.4.4. Estimation of the fraction of total soil water (FtotSW) in the greenhouse experiment

Another soil water content indicator is the fraction of total soil water (FtotSW), which was estimated as follows: $FtotSW=(W_{fi}-W_d)/(W_{sat}-W_d)$, where W_{sat} is the weight of the water-saturated pot.

II.2.4.5. Measurement of leaf water potential (Ψ) in greenhouse and field experiments

In the greenhouse experiment, the harvested plants for transcriptomic analysis (WD and control plants) were placed in a dark room until the next morning. The water status at the time of harvest (between 11:00 and 12:30) was noted Ψ_{PD}' and estimated as the leaf water potential, after equilibrium with the soil was reached. Ψ_{PD}' was measured on the n^{th} leaf for each harvested plant using a Scholander's pressure chamber (Soil Moisture Equipment Corp., California, U.S.A.), where n was 2/3 of the total leaf number N_{tot}.

In the field experiments, the water status at dawn was estimated as the classical pre-dawn leaf water potential Ψ_{PD} and was measured for one plant per plot between 4:00 and 5:30 once a week for three weeks. Measurements began when plants were at the F1 stage (CETIOM nomenclature). In the Fleurance and Samatan trials, the first measurement occurred, respectively, on July 18 and 19, 2012. In Auzeville-Tolosane, the measurements began on July 31, 2012. The water potential was measured for the 5^{th} leaf from the head (N_{tot} -5) using a Scholander's pressure chamber.

It is important to note that contrary to the transcriptomic harvests that were always performed at noon (except for the diurnal variation study), Ψ_{PD} and Ψ_{PD}' are slightly different measurements of plant water status. Ψ_{PD} was measured at dawn (between 4:00 and 5:30) after a normal night, thus representing soil-plant water status the day before leaf harvest (Figure II.2.a). In

contrast, Ψ_{PD}' was measured after the soil-plant water equilibrium had been reached during artificial night in dark room representing the exact plant water status at leaf harvest time (between 11:00 and 12:30) (Figure II.2.b).

In the Auzeville-Tolosane trial, to study the influence of the diurnal variations on leaf water potential and gene expression, leaves were harvested in each of 3 blocks for each genotype, both in irrigated and non-irrigated conditions, in the same order, between 4:00 and 5:30, 7:00 and 8:30, 10:00 and 11:30, 11:30 and 13:00, 13:00 and 14:30, 16:00 and 17:30, 19:00 and 20:30, 22:00 and 23:30, and 1:00 and 2:30 (Figure II.2.c). Separate leaves were used from the same plant for the leaf water potential measurement and for transcriptomic analysis. This study occurred on August 9 and 10, 2012 under high evaporative demand and constant sunny conditions.

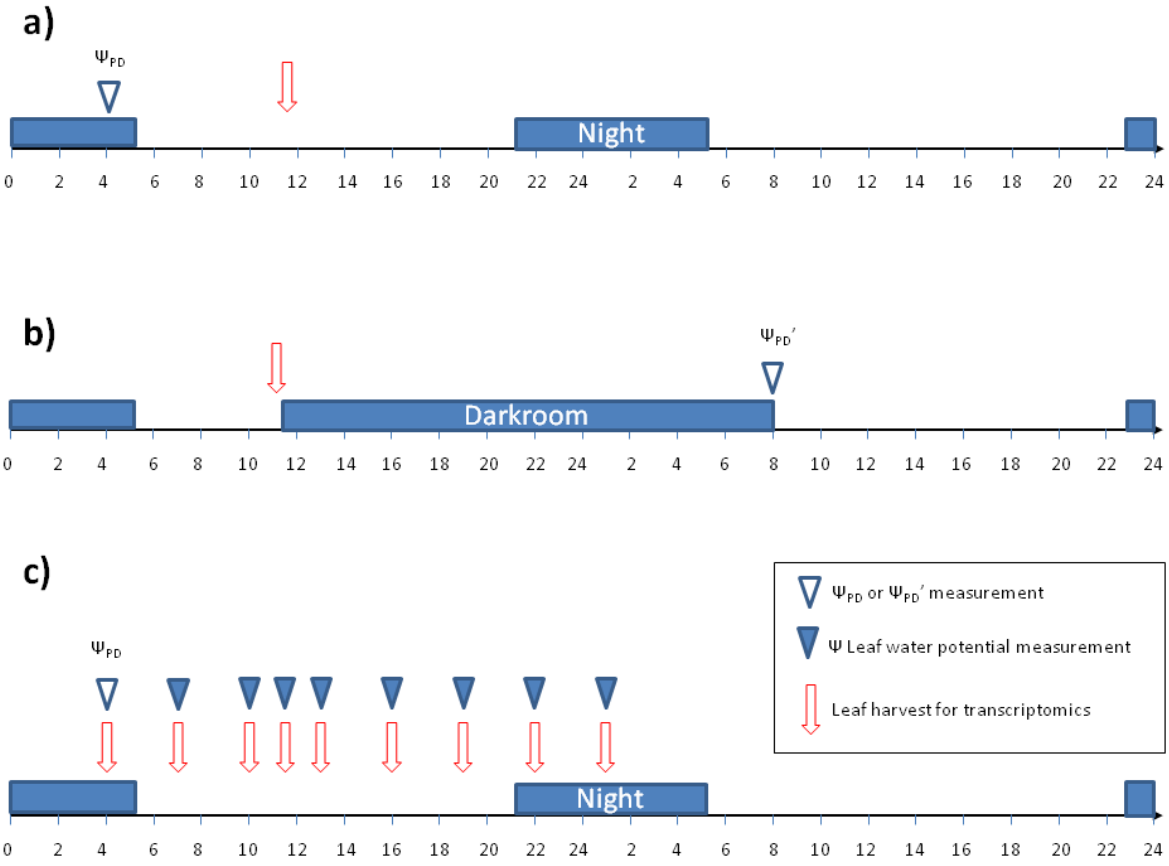


Figure II.2: Sampling and leaf water potential measurements timing.

a) Field trials: Leaf water potential was measured before leaf harvest for transcriptomic data. b) Greenhouse experiment: leaf harvest for transcriptomic occurs at day1 before artificial night and pre-dawn leaf water measurement. c) Diurnal variation study: Pre-dawn leaf water potential was measured at day 1 at 4:00. Eight leaf water potential measurements were performed during the next 24 hours. At the same time of each measurement of Ψ or Ψ_{PD} leaves were harvested for transcriptomic study.

II.2.4.6. Transcriptomic analysis

Selection of genes

Gene indicators of water status

From the study of Rengel *et al.* (2012), we selected genes that were found to (1) be correlated to the integrated transpired water (ITW) in fixed duration stress and in fixed intensity stress (R^2>0.65) and (2) show a treatment or a treatment*genotype interaction effect in the ANOVA for either type of stress. From this list of 143 genes, we chose to focus on sunflower homologues of *Arabidopsis* genes that have been described in the literature to be involved in abiotic stress responses. We finally kept 28 genes. Detailed descriptions of these genes are presented together with all the genes in this study in Appendix II.2.

Circadian clock-related genes

Numerous genes have been identified to vary according to the circadian clock. We chose the four *Arabidopsis* circadian clock regulators *TIMING OF CAB EXPRESSION 1 (TOC1)*, *LATE ELONGATED HYPOCOTYL* (*LHY*), *CONSTANS* (*CO*) and *ZEITLUPPE* (*ZTL*) (Alabadi *et al.*, 2001; Wilkins *et al.*, 2010) and identified the best BLAST hits in the sunflower transcriptome (https://www.heliagene.org/HaT13l). *HaDHN1* and *HaDHN2* of the sunflower were first described by (Cellier *et al.*, 2000) to vary during the circadian cycle and were re-examined in this study. All correspondences are summarized in Appendix II.2.

Genes showing a genotype*treatment interaction effect in ANCOVA

In addition to the 28 genes correlated to water stress intensity, we studied the gene expression levels of four transcripts: HaT13l002164, HaT13l009999, HaT13l009995 and HaT13l020030. The expression of these transcripts was found to be correlated to three other morpho-physiological variables in (Rengel *et al.*, 2012): carbon isotopic discrimination (CID), evapotranspiration (ET) and osmotic potential (OP). The identification of *Arabidopsis* homologs was performed according to the best BLAST hits. These four genes and a fifth gene originally correlated to ITW in Rengel *et al.*, (2012) study, were used to illustrate a genotype*WSB interaction effect in ANCOVA analysis explained below. Detailed descriptions of the corresponding genes are presented in Appendix II.2.

Primer design

Primers were designed using the HaT13l transcript sequence and Primer3 web tool (http://probes.pw.usda.gov/batchprimer3/index.html) using the default parameters with an optimal product size of 60bp (min=50bp, max=80bp). All primers are summarized in Appendix II.3.

Tissue harvest and RNA extraction

One non-senescent and non-growing leaf by plant was harvested between 11:00 and 13:00, except during the diurnal variation study. Each leaf was sampled and treated separately. After freezing and grinding the samples, RNA was extracted and checked for quality and quantity. Detailed protocol of these steps and the cDNA synthesis is provided Appendix II.4.

Estimation of gene expression by qRT-PCR

Gene expression was estimated by qRT-PCR. and was normalized according to the amplification efficiency and the expression levels of seven reference genes identified in Rengel *et al.* (2012). Detailed protocol of these steps is provided in Appendix II.4. and reference gene information is summarized in Appendix II.2.

Gene expression correction following the time of the day

The linear regression of gene expression as a function of the hour of the day was performed on the diurnal variation study's data between 10:00 and 20:00 for genes chosen for the biomarker model using the *robustfit* function in MATLAB. The linear regression was set to pass by means of the expression levels of samples harvested between 11:00 and 12:00, to match with the harvest time observed in the field and greenhouse experiments that were used to calibrate and validate the biomarker models. We corrected the gene expression for samples harvested at different times of the day to obtain an estimated gene expression at 11:30 using linear regression parameters.

II.2.4.7. Statistical analysis

Test of genotypic effect on the models

For each selected gene correlated with water stress intensity, we performed a covariance analysis (*aoctool* function in MATLAB) by testing genotype-dependent (1) and non-genotype-dependent (2) models for the gene expression level as a function of water stress status as follows:

(1) $Y_{i,t}=a_i+b_iX_{i,t}+Z_{it}$: genotype-dependent model and

(2) $Y_{i,t}=a+bX_{i,t}+Z'_{it}$: genotype-independent model,

where $Y_{i,t}$ is the expression level of the gene for genotype *i* and for the actual water stress intensity *t* (with different values in each of the three blocks), $X_{i,t}$ is the value of the stress intensity, and $Z_{i,t}$ and $Z'_{i,t}$ are the residues.

The gene expression was considered to not have no genotypic effect if the F-test performed as follows was not significant ($p>0.01$):

$F = (SSE_2-SSE_1/(2*(G-1))/(SSE_1/df_1)$,

where SSE_1 and SSE_2 are, respectively, the sum of the squared errors for model (1) and model (2), G is the number of genotypes and df_1 is the number of degrees of freedom attached to the error in the

model (1).

Statistical calibration and validation of the water status biomarker (WSB)

All combinations of 3, 4, 5 or 6 genes that were correlated to stress intensity and presented no genotypic effect were tested to construct a model to estimate the pre-dawn leaf water potential. The model fitting was performed by the *GeneralizedLinearModel.fit* function in MATLAB using the greenhouse data as the calibration set. For each of the four types of models, we selected the 50 best models according to the AIC criterion.

Selected models were then tested using the *predict* function in MATLAB, and field data served as the validation set. For each of the four types of models, we selected the best model according to the R^2 of the correlation between $WSB_{\Psi PD}$ predictions and the corrected values of observed Ψ_{PD} using the *regress* function in MATLAB (corrections are described in Appendix II.4.and Figure II.3). We compared the four types of models and chose the best one according to the R^2 of the correlation.

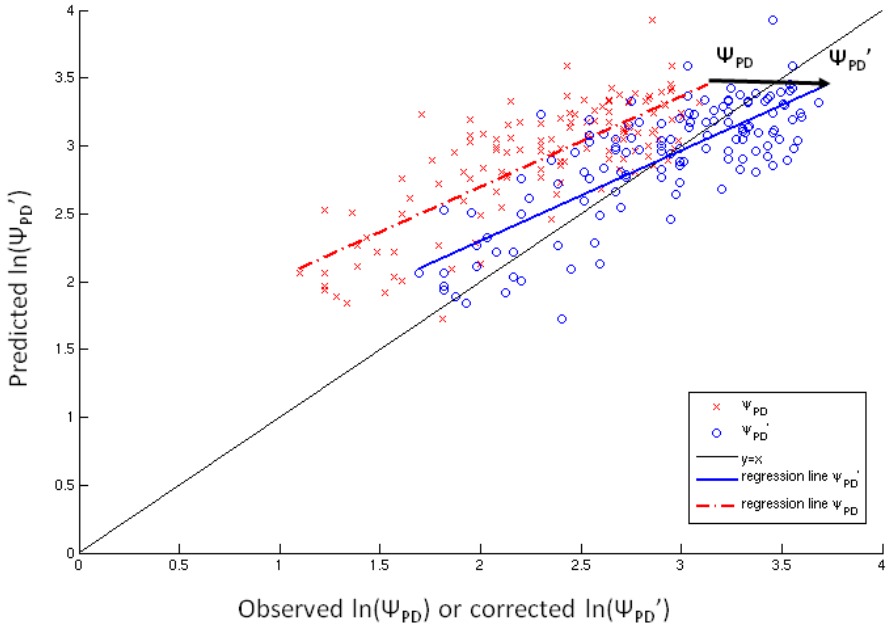


Figure II.3: Comparison ln(-Ψ_{PD}) with biomarker prediction of ln(- Ψ_{PD}) for the four best models.
Red points show comparison between model prediction and raw observation of Ψ_{PD} observed in field experiments. Blue points show comparison between model prediction and corrected observations similar to Ψ_{PD}' used for model calibration.

*II.2.4.8. Genes with Genotype*WSB interaction*

Test of trial effect

We considered samples harvested only in non-irrigated conditions for the three field trials. For the four genes correlated to morpho-physiological traits and and one gene to ITW, we performed an ANOVA using the *anovan* function in MATLAB to test the genotypic, trial and genotypic*trial interaction effects.

Test of the genotype*WSB interaction

Genes found to have no trial effect were tested for the genotype*WSB interaction. We performed a covariance analysis using the *aoctool* function in MATLAB with the following model:

$Y_{i,b} = a_i + b_i X_{i,b} + Z_{i,b}$,

where $Y_{i,b}$ is the expression level of the gene for genotype *i* and biomarker level *b,* $X_{i,b}$ is the value of the WSB level and $Z_{i,b}$ is the residue.

II.2.5. Results

II.2.5.1. Greenhouse results

Selection of candidate genes

Based on our previous results (Rengel *et al.*, 2012), we selected 28 genes that were found to be correlated to the integrated transpired water (ITW) in fixed duration stress and in fixed intensity stress (R^2>0.65). As the expression of these genes was independent of the tested genotypes, they were strong candidates to build a biomarker for plant water status. To assess a particular level of gene expression that reflects stress intensity, we needed to study these genes through a larger range and at a finer scale of drought stress.

Establishment of a fine scale of drought stress

To study changes in gene expression at different stress levels, we established a large range of drought stress with a fine scale. For the treated plants, the water status indicators ranged from 0.97 to -0.087 for the FTSW, from -0.2 to -2.4MPa for the pre-dawn leaf water potential (Ψ_{PD}'), from 54.3% to 5.98% for the soil water content (SWC) and from 0.13 to 1 for the FtotSW. The four genotypes and the two stages were represented through the entire range.

The four water status indicators measured during the greenhouse experiment were highly correlated with the R^2 values, ranging from 0.65 to 0.96 (Figure II.4). Interestingly, the Ψ_{PD}' was only correlated with FTSW values below 0.4, SWC values below 25% and FtotSW values below 0.5. This selective correlation reflects that, in our data, Ψ_{PD}' did not discriminate high water status levels.

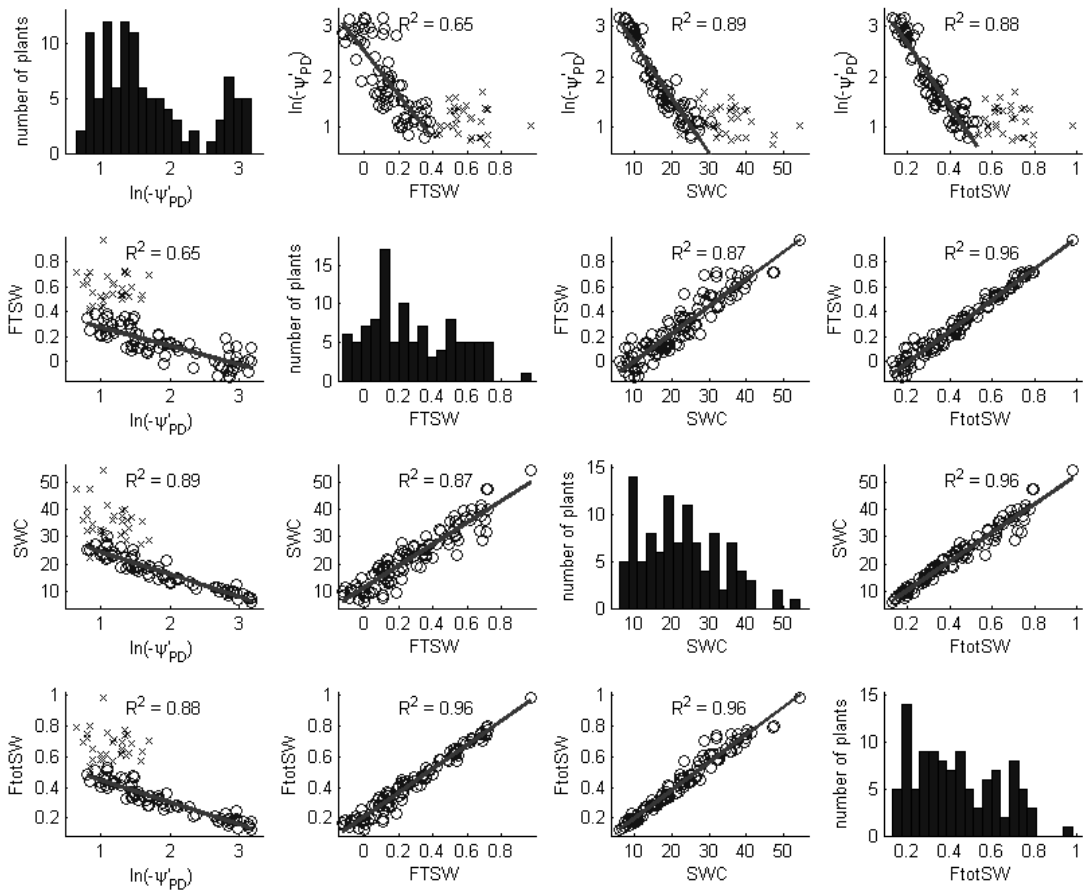


Figure II.4. The distributions and correlations between the four water status indicators measured in the greenhouse experiment: Ψ_{PD}' expressed in bar (1bar=0.1MPa), FTSW, SWC and FtotSW.

Correlation between gene expression and water status indicators

As a confirmation of our previous results using the FTSW (Rengel *et al.*, 2012), we estimated the correlations of 28 selected genes to the four water status indicators over the new finer and larger scales of water status, considering together the treated plants of all genotypes and growth stages. Raw data of the gene expression level compared to FTSW level for the 28 candidate genes are shown in Appendix II.5. We identified 18 genes whose expression was correlated ($p<0.01$) to the FTSW, 20 for Ψ_{PD}', 21 for the SWC and 18 for FtotSW (Table II.1 and Appendix II.6).

A covariance analysis was used to test genotype-dependent correlations ($p<0.01$). Among the correlated genes, we found two genes (according to the water status indicator) whose correlations were genotype-dependent (summarized in Table II.1 and Appendix II.7).

Finally, among the 28 initial genes, we retained 14 genes that showed neither genotype nor stage effects in the greenhouse experiment and that were technically robust in both greenhouse and field experiments. These first steps of gene selection for biomarker construction are summarized in Figure II.5.

	FTSW	Ψ_{PD}'	SWC	FtotSW
Number of genes correlated to water indicators	18 $0.22<R^2<0.75$	20 $0.19<R^2<0.91$	21 $0.18<R^2<0.0.8$	18 $0.18<R^2<0.77$
Number of genes correlated to water indicators without genotype effect	16 $0.22<R^2<0.71$	19 $0.19<R^2<0.83$	21 $0.18<R^2<0.77$	21 $0.18<R^2<0.8$

Table II.1: The number of genes correlated ($p<0.01$) to each water deprivation indicator and genotype effects to be used in gene combinations for biomarker fitting.

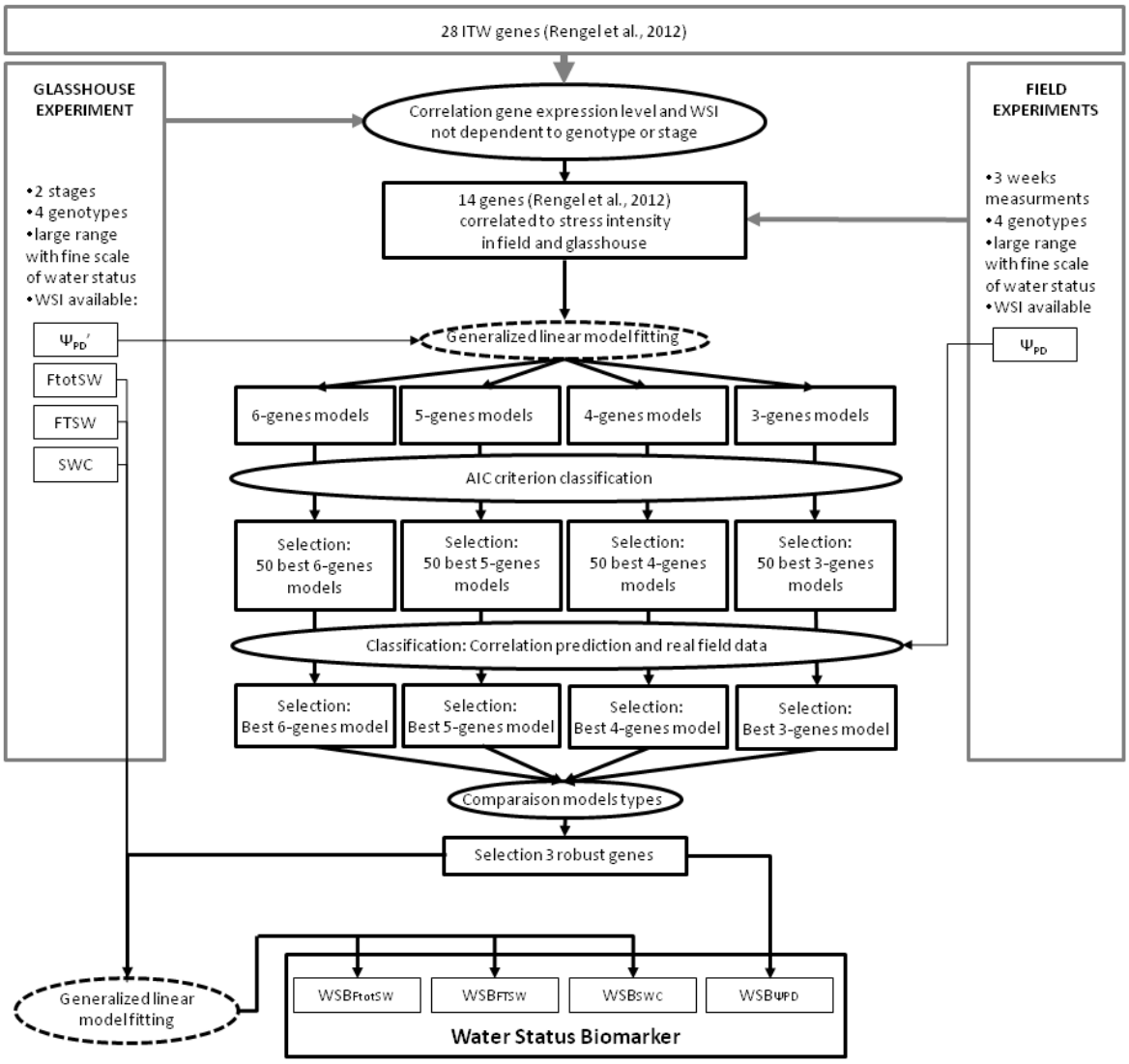


Figure II.5: A schematic description of the water status biomarker construction.

$WSB_{\Psi PD}$ was developed in greenhouse conditions and validated in the field. WSB_{FtotSW}, WSB_{FTSW} and WSB_{SWC} were built in the greenhouse environment using field-robust genes used for $WSB_{\Psi PD}$.

II.2.5.2. Construction of Generalized Linear Models to estimate plant water status

Model for Ψ_{PD}' in glasshouse

According to the AIC criterion, we selected the 50 best generalized linear models with 3, 4, 5 and 6 genes to fit the Ψ_{PD}' from the greenhouse data (Figure II.5). The adjusted R^2 and RMSEc for the different types of models are presented in Table II.2. Considering only these glasshouse data, the adjusted R^2 increased with the number of genes introduced in the model, and the RMSEc values for the four types of models were similar.

	3 gene models	4 gene models	5 gene models	6 gene models
Adjusted R^2	0.73-0.82	0.80-0.83	0.82-0.85	0.84-0.86
RMSEc	0.61-0.66	0.64-0.66	0.64-0.67	0.65-0.68

Table II.2: The range of adjusted R^2 and RMSEc for the 50 best linear models with 3, 4, 5 and 6 genes fitting the Ψ_{PD}' in the greenhouse experiment.

Field experiment validation with Ψ_{PD}

We used the results of three field trials to select the best predictive model for Ψ_{PD}' (Figure II.5). The field trial experiments were implemented in environments with deep (Auzeville-Tolosane) or shallow (Fleurance and Samatan) soils. The Fleurance and Samatan trials had clay soils (respectively in average 52.5% and 52.8% of clay) with a low water-holding capacity. The Auzeville-Tolosane trial had soil with an equilibrate texture between the silt loam and sand (composed in average of 24.3% of clay, 29.8% of silt and 45.9% of sand), and therefore, with a high water holding capacity and therefore with a high field capacity (Table II.3). Trials were chosen with different soil characteristics to ensure a wide range of plant water statuses. For the same reason, we harvested samples and measured Ψ_{PD} over 3 weeks for six repetitions per genotype. Finally, the Samatan data over the last week was disturbed by an important rain event and was discarded. Overall, we obtained a good range of water stress across the entire experiment: Ψ_{PD} ranged from -0.5 to -2.2 MPa in the Fleurance trial and from -0.7 to -2.3 MPa in the Samatan trial, whereas in Auzeville-Tolosane, where we set up irrigated and non-irrigated conditions, Ψ_{PD} ranged from -0.3 to -1.5 MPa.

	Clay (< 2 µm)	Fine silt (2-20 µm)	Coarse silt (20-50 µm)	Fine Sand (50-200 µm)	Coarse sand (200-2000 µm)	pH
AUZ 0-30 cm	203	185	96	180	336	5.85
AUZ 30-60 cm	283	221	95	138	263	6.91
FLE 0-30 cm	531	301	88	45	35	8.43
FLE 30-60 cm	520	327	56	30	67	8.6
SAM 0-30 cm	534	336	68	35	27	8.38
SAM 30-60 cm	522	316	71	48	43	8.38

Table II.3: Soil analyses. AUZ: Auzeville; FLE: Fleurance; SAM: Samatan.

Using field experiment data, we compared the models' prediction ($WSB_{\psi PD}$) and Ψ_{PD}' estimated from the measured Ψ_{PD}. We observed that models with 6 genes that were better in the greenhouse environment introduced errors in field predictions. We selected the three-gene model that showed the best correlation between the observed and predicted Ψ_{PD}' with an R^2 equal to 0.61 (Figure II.6) and an RMSEp of 0.67. With this model, the $WSB_{\psi PD}$ was estimated as follows:

$WSB_{\psi PD}=\ln(-\Psi_{PD}') = 1.53 + 0.35*\ln(dCt_{HaT13I002207}) - 0.39*\ln(dCt_{HaT13I002636}) +0.16*\ln(dCt_{HaT13I5199})$.

where ψ_{PD}' is expressed in 0.1MPa.

In the greenhouse, this model had an adjusted R^2 of 0.78 and an RMSEc of 0.64 and therefore offered better prediction in both controlled and field environments.

Models for water stress indicators not accessible in field conditions

The three genes used in the model to predict Ψ_{PD}' appeared to be robust enough in predicting the stress intensity in both the greenhouse and field environments. We used these same genes in the construction of models for water stress indicators that are not accessible in field conditions. FTSW, SWC and FtotSW were estimated by generalized linear modeling using gene expression levels of HaT13I002722, HaT13I002636 and HaT13I005199 as follows:

$WSB_{FTSW} = 0.42 - 0.0618*\ln(dCt_{HaT13I2207}) +0.21*\ln(dCt_{HaT13I002636}) - 0.04*\ln(dCt_{HaT13I005199})$,

$WSB_{SWC} = 27.70 - 3.83*\ln(dCt_{HaT13I002207}) + 8.51*\ln(dCt_{HaT13I002636}) - 1.79*\ln(dCt_{HaT13I005199})$, and

$WSB_{FtotSW} = 0.54 - 0.06*\ln(dCt_{HaT13I002207}) + 0.17*\ln(dCt_{HaT13I002636}) - 0.03*\ln d(Ct_{HaT13I005199})$.

Models had adjusted R^2 values of, respectively, 0.69, 0.72 and 0.74 for FTSW, SWC and FtotSW, and their RMSEc values were, respectively, 0.20, 9.31 and 0.18.

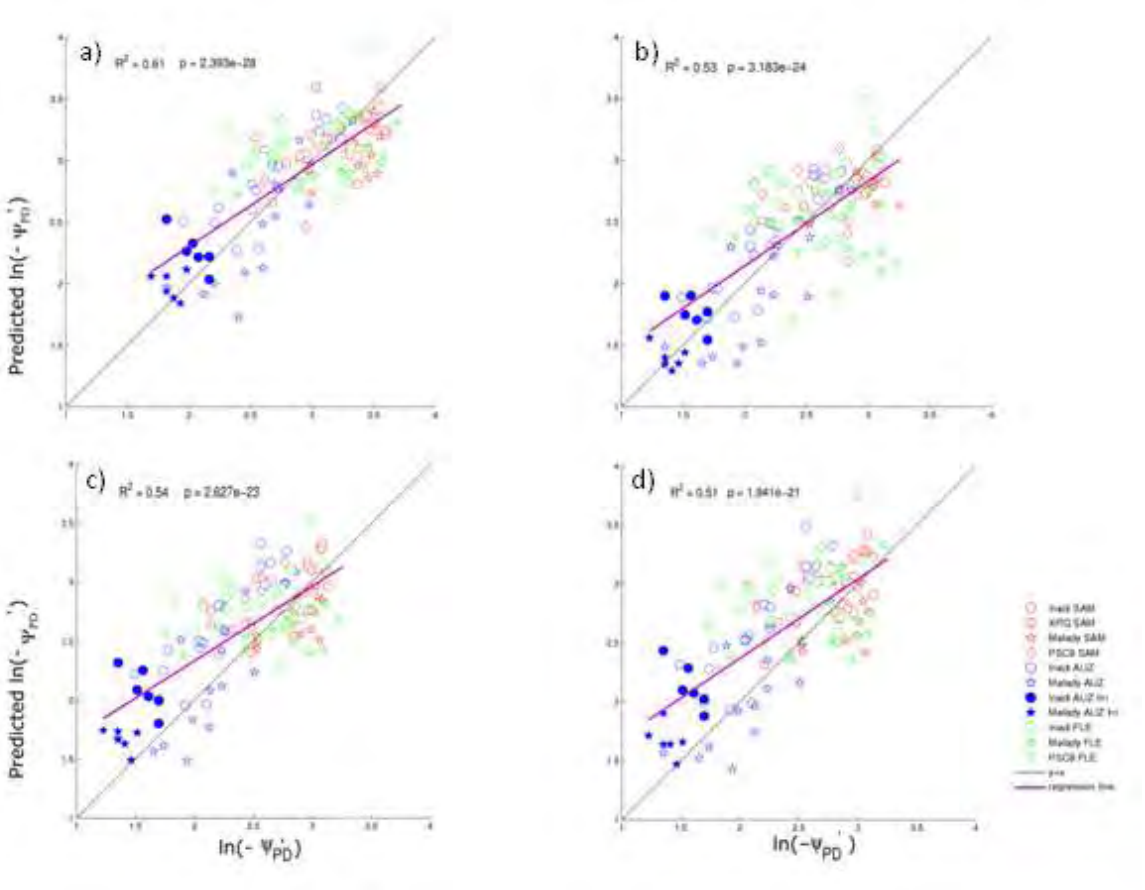


Figure II.6: Correlations between corrected field data ln(-ΨPD') and predicted ln(-ΨPD') by WSBΨPD of the corresponding best model with 3, 4, 5 or 6 genes (n=142 individual plants), where ΨPD' is expressed in bar (1bar=0.1MPa). Validation of the models were performed with samples harvested between 11:00 and 12:30 without correction for harvest time. a) Correlation with predictions of the best three-gene model. b) Correlation with predictions of the best four-gene model. c) Correlation with predictions of the best five-gene model. d) Correlation with predictions of the best six-gene model. Field data of the three trials are represented: Samatan (SAM, in red), Fleurance (FLE, in green) and Auzeville-Tolosane (AUZ, in blue). Note that the three-gene model, represented by the regression line in violet, produced better predictions, with an R^2 value of 0.61.

Correction of gene expression level for diurnal variation

To build these models, we used samples harvested between 11:00 and 12:30. So, the biomarker was calibrated and validated only for samples harvested during this period of the day. To use this model as a biomarker and a practical tool in experiments involving large numbers of genotypes or conditions, it appeared to be useful to obtain a biomarker valid for a larger sampling time period. Therefore we needed to correct for the time of the sampling, at least for genes showing modification of their expression according to the diurnal variation. This variation of expression of the three genes included in the models (shown in Figure II.7.a-c) throughout a 24-hour period could not be neglected in comparison to the variation of known circadian genes (Figure II.8). The kinetic curves of gene expression levels over 24 hours showed that between 10:00 and 20:00, the variation of gene expression could be estimated through a linear regression. We used kinetic curves over 24 hours to estimate the expression at 11:30 from the expression at any time over this specific timeframe as shown in Figure II.7.d-e. The correction was efficient for samples harvested from 10:00 to 17:30; however, for samples harvested out of this timeframe, the correction was not sufficiently reliable to estimate the gene expression at 11:30. Sampling out of this timeframe should therefore be avoided.

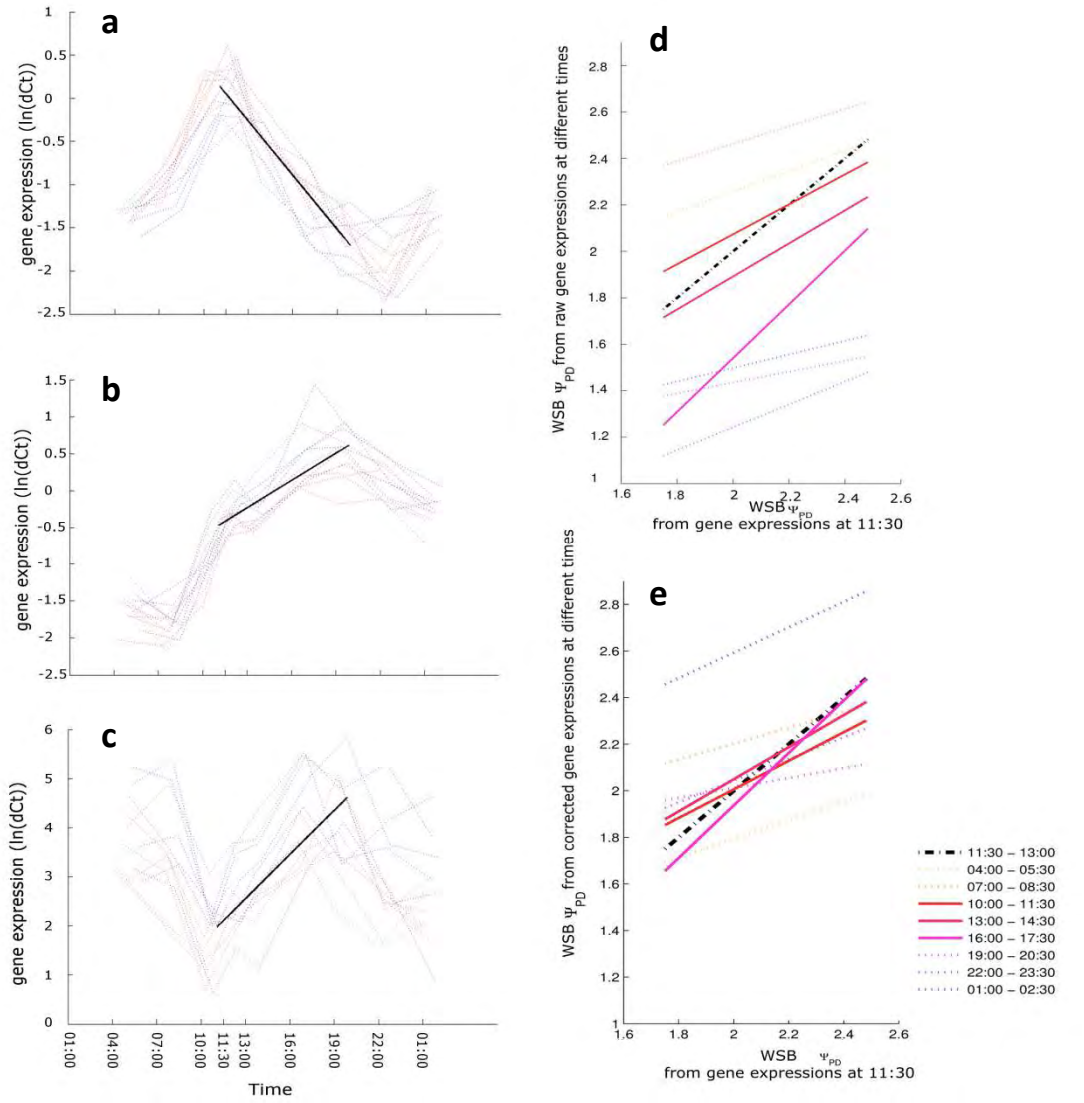


Figure II.7. Diurnal variation of biomarker genes expression and correction efficiency for biomarker prediction from samples harvested at different times of the day.

a) Twenty-four hours kinetic curves of expression level for transcript HaT13l002207. B) Twenty-four hours kinetic curves of expression level for transcript HaT13l002636. c) Twenty-four hours kinetic curves expression level for transcript HaT13l005199. Dotted lines represent kinetic curves for 6 irrigated plots and 6 non-irrigated plots. One plant by plot was harvested for each harvest time. The solid line is the regression line between 10:00 and 20:00 used for transcript expression correction. d) Comparison between biomarker predictions using gene expression between 11:30 and 12:00 and biomarker predictions using raw gene expression at different times of the day. e) Comparison between biomarker predictions using gene expression between 11:30 and 12 and corrected gene expression at different times. As biomarker model was calibrated and validated with samples harvested between 11:30 and 12:00, correction aimed at estimating gene expression at 11:30 from samples harvested between 10:00 and 20:00 and showing a linear variation. Note that the correction is efficient only for samples harvested between 10:00 and 17:30. WSB $_{\Psi PD}$ is expressed as ln(- Ψ_{PD}) with Ψ_{PD} in bar.

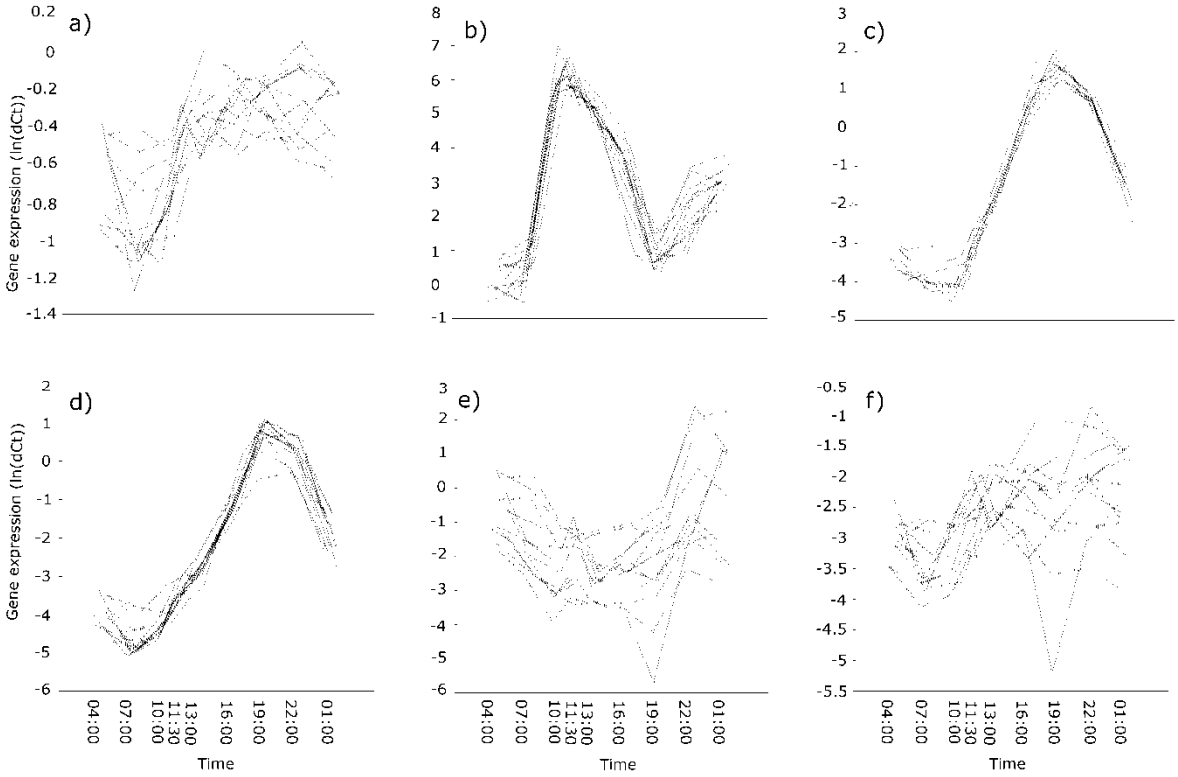


Figure II.8 : Kinetic curves of circadian genes expressions during 24 hours.

Dotted lines show gene expression over 24 hours for 12 plots representing two genotypes (Inedi, Melody) and two conditions (irrigated and non-irrigated conditions). a) Transcript HaT13l000567 homologous to the circadian *Arabidopsis* gene ZTL. b) Transcript HaT13l007116 homologous to the circadian *Arabidopsis* gene TOC1. c) Transcript HaT13l011336 homologous to the circadian *Arabidopsis* gene LHY. d) Transcript HaT13l015763 homologous to the circadian *Arabidopsis* gene CO. e) Transcript HaT13l005099 homologous to the sunflower dehydrin HaDN1 (Cellier *et al.*, 2000). f) Transcript HaT13l011509 homologous to the sunflower dehydrin HaDN2 (Cellier *et al.*, 2000).

II.2.5.3. Use of the Water Status Biomarker

Identification of gene expression profiles showing a genotype*WSB interaction in field conditions

The water status biomarker (WSB) built in this study could be applied to characterize the environments for water stress for different genotypes. This application would allow the identification of genes showing genotype*water status interactions that could explain the genotypic variation for drought tolerance. To illustrate this, we chose five genes correlated to other morpho-physiological variables or water stress intensity in (Rengel *et al.*, 2012), that did not show a trial effect in field experiments ($p>0.05$ in ANOVA over the three non-irrigated trials) as summarized in Table II.4.

Genes	AGI	trial effect (p-value>0.05)
HaT13l009400	AT1G64230	0.054555846
HaT13l002164	AT5G12840	0.141579354
HaT13l009999	AT5G15600	0.14513769
HaT13l009995	AT5G58070	0.180341098
HaT13l020030	AT5G39720	0.098184115

Table II.4: Genes with no trial effect over the three non-irrigated trials.

For these five genes, a covariance analysis showed significant genotypic and genotype*WSB interaction effects as shown in Table II.5 and Figure II.9. These results exemplify a possible use of the WSB when searching for genetic variation of the drought response.

Gene	Genotype effect (p-value)	WSB effect (p-value)	Genotype*WSB interaction effect (p-value)
HaT13l009400	4.98E-12	1.62E-17	1.90E-02
HaT13l002164	9.00E-04	1.52E-04	3.90E-04
HaT13l009999	8.35E-05	1.25E-08	1.75E-02
HaT13l009995	5.70E-01	6.37E-08	4.83E-0.2
HaT13l020030	3.41E-05	3.78E-03	7.23E-03

Table II.5: Results of covariance analysis for five selected genes.

These genes shown G*WSB interaction effect ($p<0.05$) and illustrate the use of the biomarker to detect differential plant drought responses according to the genotype and the plant water status as it is identified by the WSB.

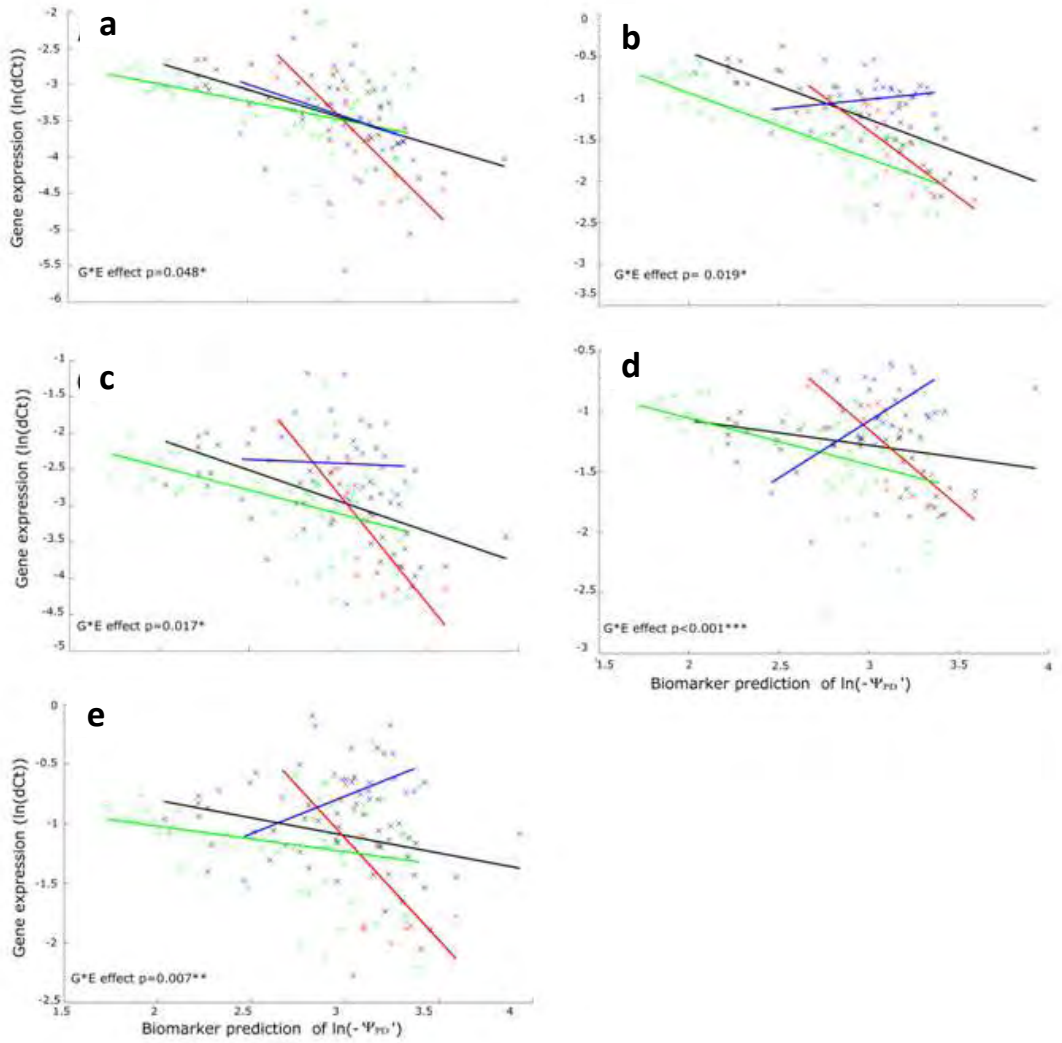


Figure II.9: Gene expression showing genotype*WSB interactions in ANCOVA study (p<0.05).
The four genotypes Melody (green), Inedi (black), PSC8 (blue), and XRQ (red) showed different interactions with the environment described by the $WSB_{\Psi PD}$. $WSB_{\Psi PD}$ is the prediction of $ln(-\Psi_{PD}')$ with ΨPD expressed in bar (1 bar=0.1MPa) (a) HaT13I009995, homologous to the *Arabidopsis TIL1* transcript. (b) HaT13I009400, homologous to the *Arabidopsis UBC28* transcript. (c) HaT13I009999, homologous to the *Arabidopsis SP1L4* transcript. (d) HaT13I002164, homologous to the *Arabidopsis HAP2A* transcript. (e) HaT13I020030, homologous to the *Arabidopsis AIG2L* transcript.

II.2.6. Discussion

II.2.6.1. Description of the three genes selected for the water status biomarker

The water status biomarker was defined from the expression levels of three genes normalized by the expression levels of reference genes. HaT13I002207 is homologous to the *Arabidopsis* transcript of *TUA5* (AT5G19780). This gene encodes a tubulin. Microtubules are polymers of tubulin heterodimers. The relationship between microtubules and ABA in plant cells has been extensively studied, although the exact mechanisms involving the microtubule response to drought stress remain largely unknown. Dynamic microtubules in guard cells are sensitive to extracellular

stimuli and drought stress, which affect both the microtubule dynamics and ABA accumulation (Marcus *et al.*, 2001). Moreover, Lu *et al.* (2007) demonstrated that changes in microtubule dynamics have an effect on ABA accumulation in root cells of *Zea mays*.

HaT13l005199 is homologous to the *Arabidopsis* transcript of *XTR7* (AT4G14130). This second gene encodes for a concanavalin that is a xyloglucan endotransglycosylase (XET). XETs form a large family, of which some members are involved in cell wall biogenesis or rearrangement (Van Sandt *et al.*, 2007), which are processes inherent to growth. Moreover, a relationship between the response of the growth rate under water stress and XET activity has been previously suggested (Thompson *et al.*, 1997).

HaT13l002636 is homologous to the *Arabidopsis* transcript of *GBF3* (AT2G46270). This last gene is a transcription factor that encodes a bZIP G-box binding protein, and its expression was found to be induced by ABA, cold and water deprivation (Lu *et al.*, 1996).

These three genes were shown to have direct or indirect links with water deficit or ABA, which is the drought stress hormone. Although these links were demonstrated in *Arabidopsis* and in the microtubule dynamics of maize, our biomarker gene selection and model calibration may be specific to the sunflower. Accordingly, new WSB development would be required for other species.

II.2.6.2. Comparison between the WSB and classical water status indicators

From the correlations observed between the Ψ_{PD}', FTSW, SWC and FtotSW, we confirmed that plant-based water status indicators, such as Ψ_{PD}' and FTSW, reflect the soil water content. The expression levels of genes selected to construct the WSB were highly correlated to the water status indicators, especially to the plant-based indicator Ψ_{PD}'. Therefore, the expression of these three genes reflects the environmental water status as integrated by the plant. As the expression of these genes was independent of genotype and stage, the determination of a given drought stress based on the water status biomarker was the same for all genotypes and stages tested.

We obtained a better WSB for Ψ_{PD} because we selected genes using the Ψ_{PD}' measurements but also because gene expression levels and Ψ_{PD}' are both plant-based measurements. Therefore, to characterize the water available for the plant in the field, $WSB_{\psi PD}$ was the most robust biomarker.

II.2.6.3. Advantage of WSB over environmental data

The WSB built in this study represents the environmental water status perceived by the plant. The expression of the selected genes was correlated to the environmental water status and independent of genotypic diversity and stage. However, the water status described by the biomarker was not exactly the same as the water status described by soil-based water indicators. For example, the SWC reflected the water status of the soil, but according to the type of soil, the availability of

water to the plant could vary. As a plant-based water status indicator, gene expression levels have the advantage of offering a better representation of the environment perceived by the plant than raw soil and climatic data.

II.2.6.4. WSB genes and environmental signal integration

Gene regulatory networks integrate the environment to drive morphological and physiological responses (Shinozaki & Yamaguchi-Shinozaki, 2007). Any genotypic variability in a step of the response cascade would translate in further genotypic variations in the overall drought response process. Therefore, we can argue that the genes used in our genotype-independent biomarker are involved in upper steps of this cascade close to the integration of the environmental signal or in general genotype-independent pathways (maybe illustrated by the presence of the tubulin gene in our WSB).

II.2.6.5. Validity of the WSB

The water status biomarker was developed using the correlation with Ψ_{PD}' in a greenhouse and validated with field Ψ_{PD} data (Figure II.5). Because they are not easily tractable in field conditions, the estimation of FTSW, SWC and FtotSW using our gene expression biomarker was only validated in the greenhouse experiment. However, to account for greenhouse and field variations, we built WSB_{FTSW}, WSB_{SWC} and WSB_{FtotSW} with the same genes used for the $WSB_{\Psi PD}$, as they were shown to be robust in both environments. This robustness allows us to be confident in our predictions of the WSB_{FTSW}, WSB_{SWC} and WSB_{FtotSW} in the field and makes these indicators accessible in this environment; providing a new tool for plant biologists.

The circadian regulation of gene expression has been documented for a couple of sunflower genes (Cellier *et al.*, 2000). However, to construct the biomarker, we did not take into account that the diurnal variation of our genes expression could be important. To calibrate and validate the water status biomarker, we sampled plant tissues between 11:00 and 12:30, so the problem of sampling period was not crucial. With the diurnal variation study, we were also able to infer the gene expression level at 11:30 from samples harvested between 10:00 and 17:30, which thus defined a valid timeframe for sampling. This correction was specific to the day of study (sunny and warm) and might at least improve the prediction in most of the drought studies.

The biomarker construction was based on samples harvested from May 31 to June 15 2012, including average relative humidity varying from 63 to 80% on cloudy and sunny days reflecting variable evaporative demands. It selected genes that were not affected by this kind of climatic variations. This was confirmed in the field experiments performed two months later in shorter day

conditions, in three different locations with higher evaporative demands.
In these conditions, the biomarker gene expression was not correlated to diurnal water potential and was not influenced by the specific climatic conditions of the day of harvest. So in all these different relative humidity and average temperature conditions, we could validate our model for both experimental conditions.

Both in the greenhouse experiment and in field trials, we managed to obtain a large range of water stress, and we showed that the WSB was able to characterize the environment for this entire range of water statuses. However, in all experiments, we only applied a continuous deficit of water. We did not test the ability of our WSB to describe plant recovery if, for example, rain events occur after a long period of water deficit.
Importantly, the expression of the three genes chosen to the build biomarker appears to be independent of the tested genotypic diversity. This independence is a particular characteristic of the selected genes and makes them distinct from many other genes as we describe below.

II.2.6.6. Use of the WSB

Breeding for a trait affected by drought (G*E)

For crop breeding, environmental characterization is critical to understand the genotype*environment interaction. Climate and crop management data alone are not sufficient to obtain a good definition of the environment perceived by the plant. Therefore, the WSB developed in this study could be useful for characterizing the environment with regard to water availability, allowing breeders to better understand genotype*drought stress interactions. Accordingly, this WSB could be a powerful tool to study any trait affected by drought and help to breed drought tolerance in sunflower.

Example: identification of gene expression responses depending on the water status biomarker

Following this G*E identification strategy, we looked for genes showing significant genotype*WSB interaction effect and whose expression levels were independent of trials.
Our results suggest that five genes could show this pattern (Figure II.9). Among these genes, four were found correlated with morpho-physiological variables and the last one to water stress intensity in (Rengel *et al.*, 2012). These genes were examples of genes related to plant drought responses and whose expression changed according to the genotype and the plant water status as predicted by the WSB. The expression of HaT13l002164 and HaT13l009999 was correlated with carbon isotopic discrimination (CID). These transcripts are respectively homologous to the *Arabidopsis* transcripts of *HAP2A* (AT5G12840), which codes for a subunit of the CCAAT-binding complex, and *SP1L4*

(AT5G15600), which regulates cortical microtubule organization. The third gene encodes HaT13l009995, whose expression is correlated with evapotranspiration (ET). This transcript is homologous to the *Arabidopsis* transcript of *ATTIL* (AT5G58070) involved in thermotolerance (Chi *et al.*, 2009). The fourth gene encodes HaT13l020030, whose expression is correlated with the osmotic potential (OP). Its *Arabidopsis* homolog is the avirulence-induced gene *AIG2L* (AT5G39720). Finally, HaT13l009400 expression was found to be correlated with water stress intensity but was not used to build the biomarker because it showed a genotype effect in the greenhouse experiment. This transcript is homologous to the *Arabidopsis* transcript of *UBC28 (*AT1G64230), which codes for a ubiquitin-conjugating enzyme.

Following the signaling cascade from the environmental signals down to the genotype-specific responses, these genes would be responsible for the final responses and belong to the end of the gene signaling cascade. Because we were able to identify environment-related genes and response-related genes, this approach could allow us to model the gene regulatory network from the global gene expression dataset.

Crop model

Crop models represent dynamic crop processes and are used to simulate crop development and behavior as a function of the environment, management conditions and genetic variations (Sinclair & Seligman, 2000). Such tools may also benefit from the use of the WSB. For example, SUNFLO (Casadebaig *et al.*, 2011) is a crop model that is able to simulate biomass yield and transpiration of the sunflower genotypes in contrasting environments. In this model, the FTSW is an output variable of a water budget module based on climatic and management input variables and plant parameters (expansion and transpiration sensitivity to water stress, soil depth and water holding capacity). The simulated FTSW is thereafter used to model the effects of water stress on crop growth and performance.

In this context, WSB_{FTSW} could be a tool to readjust the simulated FTSW values with observations to improve crop performance prediction for a specific site. It appears impossible to harvest plants every day to obtain a daily WSB_{FTSW}. However, harvesting at a few key stages of crop development appears to be a good compromise and could help to perform a more accurate simulation of crop development.

Distinction between traits of interest and drought stress responses

In the field, crops are actually subjected to both abiotic and biotic stresses. These two types of stresses are in interactions. A biomarker characterizing water status could be a tool to distinguish the part of genetic variation in a trait of interest, such as distinguishing the resistance to a disease

from drought stress responses that could interfere in phenotyping. As an example, it has been shown that water deficit conditions were significantly involved in the disease severity of premature ripening induced by *Phoma macdonaldii* susceptibility (Seassau *et al.*, 2010). In this case, the biomarker for characterizing the water status environment perceived by the plant could be used to perform a screening of *Phoma*-tolerant genotypes adjusting for different water statuses.

Drought stress management feasibility: time and cost of the WSB

Our goal in this study was not only to demonstrate the possibility to characterize the water status environment from gene expression levels but also to design a practical tool that could be easily used. To achieve this goal, we paid particular attention to the cost and time needed to run the new biomarker.

Important parameters to consider in the development of a cost- and time-effective biomarker are the sampling time window and the number of genes used.

Concerning the time window for sampling, the diurnal variation study allowed us to estimate gene expression levels at 11:30 from samples harvested between 10:00 and 17:30. Therefore, the WSB can be used with samples harvested during a large diurnal sampling period, in contrast to the Ψ_{PD}, that can only be measured at pre-dawn.

Regarding the number of genes, we developed a WSB based on the expression of only a few genes, i.e., the three genes included in the generalized linear models and reference genes used for normalization. Therefore, it is possible to easily test very large numbers of samples using q-PCR with minimal time and cost.

However, because of the delay between harvest and q-PCR results, the WSB seems more relevant for breeding or studying genotype behavior than for drought stress management during crop production.

II.2.7. Conclusions

In this study, we developed a gene expression biomarker that was able to estimate the plant water status expressed as the Ψ_{PD}', FTSW, FtotSW or SWC (Figure II.5). This tool is independent of the tested genotypes and the developmental stage. A correlation between the WSB and Ψ_{PD}' was validated in greenhouse and in field conditions with different soil properties. Other classical water status indicators showed robust correlations with the WSB in greenhouse experiments. The water status biomarker developed here could be a useful tool in different scientific fields for characterizing the water status in plants.

End of article: "A biomarker based on gene expression indicates plant water status in controlled and natural environments" published in December 2013 in Plant Cell & Environment.

II.3. Conclusion and outlooks regarding the Water Status Biomarker

Several leads can be followed up in order to improve the Water Status Biomarker (WSB) described above or to avoid some problems during the future development of new plant biomarkers. First, as already mentioned, the WSB was not tested for plants recovering from a drought stress, i.e. when the water status improves after a severe drought event. Knowledge about gene expressions in such situation could be useful to circumscribe the validity domain of the WSB. For example, the nitrogen status biomarker for maize developed by Yang *et al.* (2011) was tested and still valid for recovering plants. However, testing this hypothesis for our WSB raises the question of the rate of recovery for the gene expression following a water stress, even if it is very likely that ARN turn over and gene induction should happen faster than the morpho-physiological responses to rehydration. Photosynthesis, transpiration, and stomatal conductance were shown to be correlated to leaf water potential in kidney bean during the recovering following a drought stress (Miyashita *et al.*, 2005). If the recovery of drought stress is similar in sunflower, it suggests that the utilization of a WSB to study physiological traits related to drought stress would remain valid in this drought scenario followed by recovery. However, it does not imply that the expression of genes used in the biomarker will have a similar rate of recovery. A new study, addressing those different issues, will provide more knowledge about gene expression and regulatory network under drought recovery.

Another important aspect of the WSB that needs to be discussed is its stability under a wider genetic variability. Expression of genes involved in the biomarker was found independent of the genotype. However, only a small genetic diversity (8 genotypes in Rengel et al., (2012) and then 4 in this study) was used to build the WSB. We can argue that genotypes used for the biomarker construction have different strategies regarding drought stress. The Melody hybrid closes its stomata under a low water potential whereas Inedi reduces stomatal conductance at a higher water status. Despite this choice in the genotypes, we might expect that the biomarker genes are differentially expressed in a panel of genotypes with a larger genetic diversity, as for example wild species compared to elite lines. This aspect will be discussed again in the chapter III of this PhD work.

Finally, when using the WSB to estimate the plant water status, we have to keep in mind that it is a predictive model for water status with a non-negligible margin of error. However, this inconvenient has to be compared with the precision of the measure performed by the Scholander's chamber used for the pre-dawn leaf water potential measurements. Moreover, measuring each genotype water status with the standard water stress indicators in the experiments set up for wide genetic study involving hundreds of genotypes is not realistic. We agree that using WSB as sole indicator of water stress is probably not sufficient and should be combined with climatic and environmental data as well as water status measurements (Ψ_{PD} for example) on a limited number of

control genotypes. Those measures could be introduced in a crop model to estimate water status of genotypes via another method which could be compared with the WSB predictions. However, the introduction of the WSB in genetic experiments in order to compare genotypes while correcting for their water status will certainly help reducing the Genotype x Environment (GxE) bias even if it does not eliminate it completely.

To conclude, we could plan to build others biomarkers based on gene expression for eco-physiological traits with a high throughput and in field conditions. It could be very useful for complex traits that are difficult and time consuming to measure with traditional tools and methods. It points out two types of biomarkers. The first group, as in our study, aims at characterizing the environment and its perception by the biological organisms (here the sunflower). The second type of biomarkers is used to study and characterized complex physiological traits to characterize the development of the organism, population or community, and/or its response to environmental factors. In crops, there is a clear interest to develop proxies for photosynthesis or transpiration rate for example. Scalable at high throughput, such biomarker would greatly help the identification of the genetic control of these traits and therefore the construction of ideotypes. Biomarker could also be developed in order to characterize biotic stress. Indeed phenotyping for disease resistance in a quantified manner is a difficult work. Therefore, biomarker utilization could be a way to tackle this difficulty.

II.4. Discussion about genes receptor of the environmental signal

The first criterion for WSB genes was the correlation between their expression level and the water stress intensity. On this account, we can state that they are part of the regulatory cascade leading to drought stress responses. The second criterion for those genes selection was that their expressions were not correlated to other main morpho-physiological traits involved in water deficit responses. Finally, their expression was not dependent on the genotype. All together, this suggests that they are (i) part of the receptor system of the environmental signal (ii) involved in signal transduction and/or in particular responses that were neither studied nor genotypically variable.

The hypothesis that they are part of a receptor system is reinforced by the functional annotation of their *Arabidopsis* homologues. Shinozaki and Yamaguchi-Shinozaki (1997) reported that a change in the physical tension of the cytoskeleton during water stress might triggers osmotic responses and that the xyloglucan endotransglycosylase are part of the touch-genes that induce water-stress-inducible genes. Other studies reported in literature, reported similar description of these types of genes in drought responses (Van Sandt *et al.*, 2007; Thompson *et al.*, 1997). Therefore, we can propose the hypothesis that the WSB genes are, at least for the genes encoding a tubulin (TUA5) and a concanavalin (XTR7), part of a receptor system of the environmental signal.

The expression of the genes selected for the WSB are independent of the genotype at least for the genotypic diversity tested. However, drought stress can differ between two genotypes due to the genotype-dependent expression of some genes in the regulatory cascade for drought stress responses. There is no clear knowledge of the stage where some genetic variability arises in the gene regulatory pathway. A hypothesis that needs to be verified is that different cascades regulate drought stress responses: some that involve only genotype-independent genes and the others involving a mix of genotype-dependent and genotype-independent genes. Cross-talks and connections between these cascades would be common. Based on this assumption, genes used for the WSB construction would either be at the beginning of the overall cascades (starting from the environmental stimulus) or anywhere in a genotype-independent cascade that would not control the characterized morpho-physiological responses (Figure II.10).

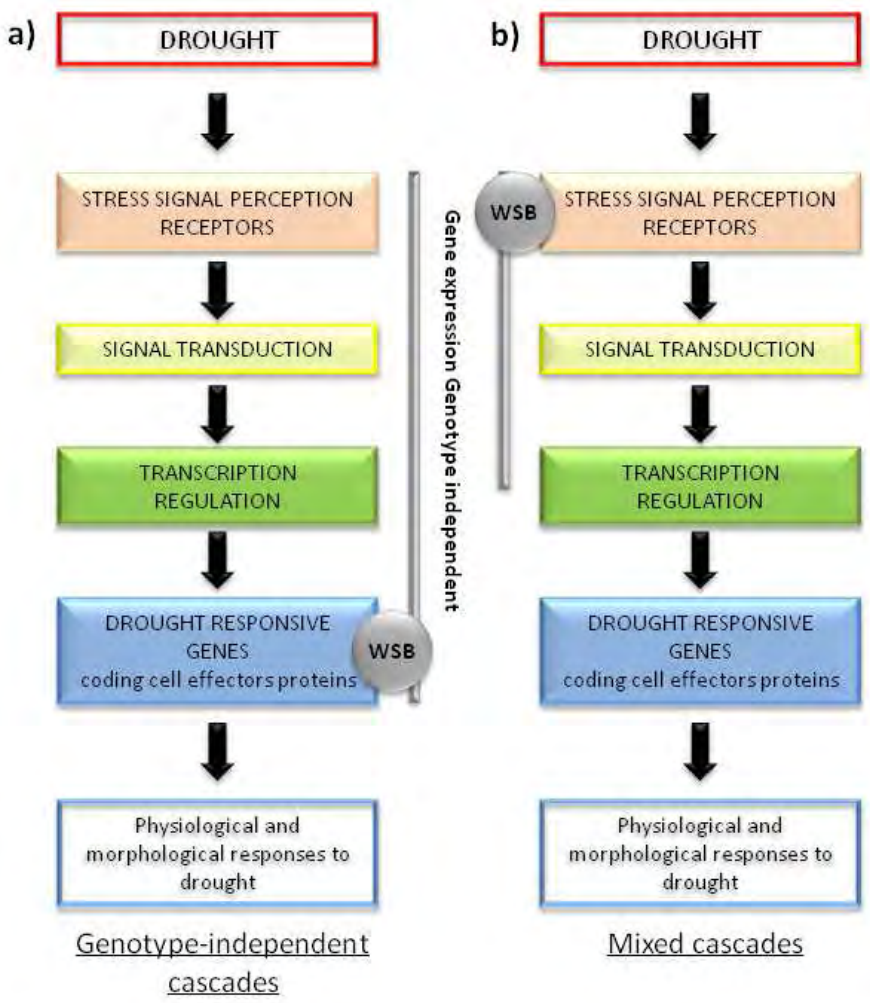


Figure II.10: Localization of WSB genes and hypothesis of two regulatory cascades for drought stress responses regarding genotype dependency of the gene expressions.
(a) case of a separate genotype-independent cascades (b) case of mixed cascades. Grey lines represent the hypothesized localization of genotype-independent genes.

This work about genes that are receptors of the environmental signal raises other questions. Plant water status evolves: it is a constant adjustment between drought stress responses of the plant and water constraint of the environment. A first question that needs to be addressed is the genetic control of the plant water status that allows this adjustment. Then, it would be also interesting to focus our research on the genetic control the genes underlying the major morpho-physiological traits involved in drought stress responses. It could help understanding the GRN that link receptor genes and effector genes for water deficit responses. Moreover, the utilization of the WSB in such analyses allows the comparison of genotypes with the same plant water status.

Chapter III: Linking complex morpho-physiological traits involved in drought tolerance to the underlying genomic loci. Reconstruction of a regulatory network through a genome wide association study of gene expression levels.

In this third chapter, we choose to focus on drought responsive genes encoding for cell effectors proteins as opposed to receptors and signal transducer proteins. This class of genes regulates the main morphological and physiological traits involved in water deficit tolerance. Those traits are complex quantitative traits under the control of many genes interacting between them and with the environment. To link DNA sequences variation to the diversity of phenotypes, two main genetic approaches, described in the next section, have been developed.

III.1. Deciphering the genetic control of complex traits: Quantitative Trait Locus (QTL) analyses or association studies.

Many traits of agronomical interest, such as drought tolerance, are quantitative traits controlled by several genes and their interaction with the environment and between them. Deciphering this genetic control is a major goal for breeder in an objective of varieties improvement. To aim at this goal two main approaches have been developed these last decades: quantitative trait locus (QTL) detection (Lander & Botstein, 1989), also called linkage mapping, and association mapping. Improvement in both methods has been made possible thanks to major progresses not only in the genotyping technologies (Jiménez-Gómez, 2011) but also in the statistical methods used (Yu *et al.*, 2006; Mackay & Powell, 2007). These approaches lead to the improvement of the breeding practices (Morgante & Salamini, 2003). Both approaches have for final goal to find significant statistical correlation between the genotype and the observed phenotype. The two approaches do not use the same type of genetic material. Linkage mapping focuses on families of known pedigrees as for example a RIL (Recombinant Inbred Line) population. On the contrary, association mapping used a collection of individuals whose ancestry is often unknown in plant (Yu & Buckler, 2006). Both methods rely on the principles of genetic recombination and exploit the shared inheritance of two loci: the targeted functional polymorphism and an adjacent marker. This shared inheritance is due to

the linkage disequilibrium (LD). LD is defined by the nonrandom association of alleles at different loci (Flint-Garcia *et al.*, 2003). If a LD occurs, an allele at a locus is found associated to a second locus more often than if the two loci segregate independently in the population. In the case of the QTL detection approach, LD level between two loci is caused by the linkage between them, i.e the physical distance between the two loci. In association mapping, the studied genotypes are the result of a complex and unknown evolution history. This approach exploits historical and evolutionary recombination at the population level (Yu & Buckler, 2006). Therefore LD between two loci is due to linkage but also to various mechanisms related to population history and that influence LD level and decay.

Mutations are the initial mechanisms which provide the polymorphisms that will occur in LD. In addition to recombination events LD is influenced by the mating pattern of the species, the selection, the reduction of the population size, the admixture and the genetic drift (Flint-Garcia *et al.*, 2003). Table III.1 sums up the major phenomena affecting LD.

Mechanisms	Effect on LD
Mutation	Temporary increasing of LD around the locus affected by the mutation
Recombination	Decreasing of LD
Admixture	LD extends even to unlinked sites but breaks down rapidly with random mating
Reduction of population size (bottleneck)	Conservation of few allelic combination induces increasing LD
Selection	Increasing of LD
Mating pattern:	LD decay more rapidely in outcrossing species as compared to selfing species, because recombination is less effective in selfing species that are largely homozygous
Genetic drift	In small population it goes with the loss of rare allelic combinations and therefore with an increase in LD level

Table III.1: Mechanisms that influence LD level and decay

Due to the difference of genetic populations used by the two methods, association mapping has the advantage over linkage mapping that it takes into account a greater number of alleles and

has a broader reference population. Another distinction involved by the difference of genetic material is the mapping resolution for the same size of population. In families with known ancestry, there are only few opportunities for recombination to occur. It results in a low mapping resolution. On the contrary, association mapping exploits historical recombination and natural genetic diversity, resulting theoretically in a higher mapping resolution (Zhu *et al.*, 2008). Figure III.1 illustrates the difference in mapping resolution for the two methods. However, this difference between linkage and association mapping is in reality not so categorical. Indeed, the comparative theoretical high resolution of association mapping is dependent of the structure of the LD across the genome that can limit recombination events.

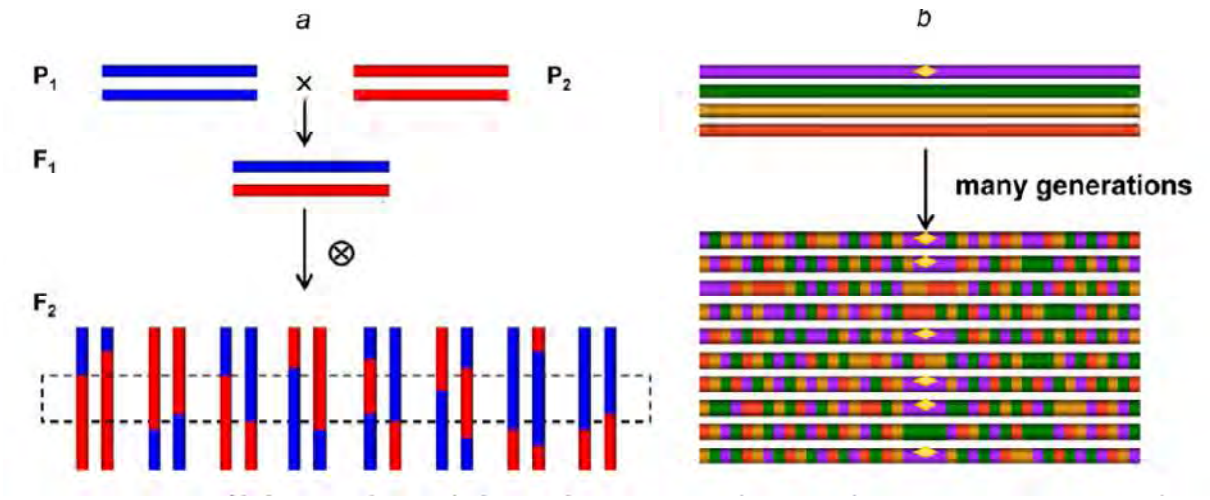


Figure III.1. Schematic comparison of linkage analysis and association mapping
(a) linkage analysis with designed mapping populations and (b) association mapping with diverse collections (Zhu et al., 2008).

So, the decay of LD over physical distance in a population determines the necessary marker density coverage. As LD has been shown to vary between species and within species it is necessary to study LD level and extent in the population before performing an association analysis (Flint-Garcia *et al.*, 2003). LD decay varies also from one locus to another with a larger extent of LD for loci target of the selection: for example, adh1 (alcohol deshydrogenase 1) was shown to have LD extend over 500kb in Maize elite lines (Jung *et al.*, 2004). Several studies on *Arabidopsis thaliana* (Nordborg *et al.*, 2002) and Maize (Tenaillon *et al.*, 2001) have been conducted. Thanks to the new sequencing technologies and high throughout genotyping, a more and more important number of markers is available. It made easier the association mapping development and the transition between the candidate genes approaches and the genome-wide strategies (Rafalski, 2002). The first association mapping studies on plants have been conducted by Buckler and his collaborators on flowering time in maize (Remington *et al.*, 2001; Thornsberry *et al.*, 2001). Since this first study, many were

published in a variety of plant species as rice (Agrama *et al.*, 2007), rapeseed (Honsdorf *et al.*, 2010), soybean (Singh *et al.*, 2008) and even sunflower as detailed below.

Despite some advantages of the association mapping, this method can lead to the detection of numerous false positive associations. However, as already discussed above, LD in an association panel can be due not only to linkage but also to various mechanisms, such as the panel structure and the familial relatedness between individuals. Indeed, if in reality the association panel is composed of several distinct sub-groups, each with different allele frequencies, the union of such sub-groups in one population leads to a modification of the LD and therefore to a risk for false positive detection. This risk increases with the population size (Yu & Buckler, 2006; Thornsberry *et al.*, 2001). Several statistical methods have been proposed to account for population structure and familial relatedness as for example the structure association computed by the software STRUCTURE (Pritchard *et al.*, 2000), the principal components approach (Patterson *et al.*, 2006) or the mixed model approach (Yu *et al.*, 2006) that takes into account the genotypic effect through a random factor and combines structure population estimation (matrix Q) and relative kinship for each genotype pairs (matrix K). Globally, those methods use genotypic information from random molecular markers across the genome to account for structure population and familial relatedness in association tests (Figure III.2).

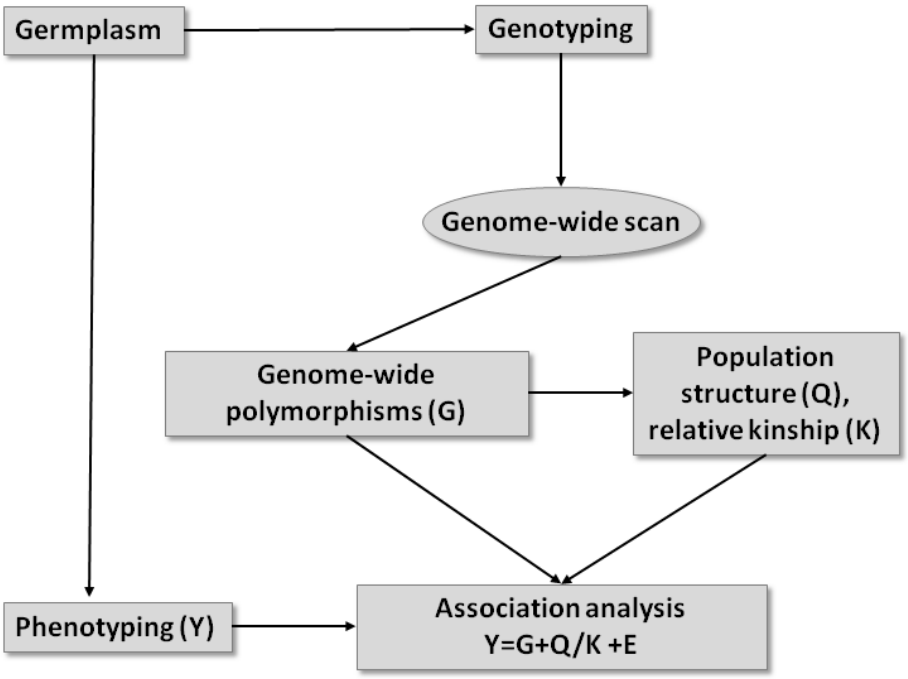


Figure III.2: Schematic diagram of genome-wide association mapping.
The inclusion of population structure (Q), relative kinship (K) or both in final association analysis depends on the genetic relationship of the association mapping panel. E stands for residual variance. (adapted from Zhu *et al.*, 2008)

Linkage mapping and association mapping are two complementary approaches with both advantages and disadvantages. More recently, another type of mapping population, the nested association mapping (NAM) population has been developed in order to overcome, in particular, the issue of marker density for a genome-wide association study. A NAM population consists of a large set of related mapping progenies (for example RIL) developed from diverse selected founders. The sequencing or the dense genotyping of the founders associated to the genotyping of both founders and the progenies thanks to a small number of tagging markers allow projecting the high-density marker information from the founders to the progenies. Therefore a GWAS study can be conduct on the progenies with a high marker density (Yu *et al.*, 2008). Such population is developed, for example, for maize by the Maize Diversity Group (http://www.panzea.org) in order to dissect complex traits. Close to the Maize NAM population, in *Arabidopsis* and in rice, Multiparent Advanced Generation Inter-Cross lines have also been developed to improve the power to detect and localize QTL (Kover *et al.*, 2009; Bandillo *et al.*, 2013). A MAGIC population is initiated by intermating the founder accessions during several generations (for example, four in the *Arabidopsis* MAGIC lines). In a second step the outcrossed families produced by the intermating are inbred for several generations in order to produce a stable panel RIL composed of nearly homozygous lines. The Figure III.3 shows an example of MAGIC population development as realized in the study of Bandillo *et al.* (2013). Analytic methods have been developed to fine-mapping QTL in the MAGIC lines by reconstructing the genome of each line as a mosaic of the founders (Kover *et al.*, 2009).

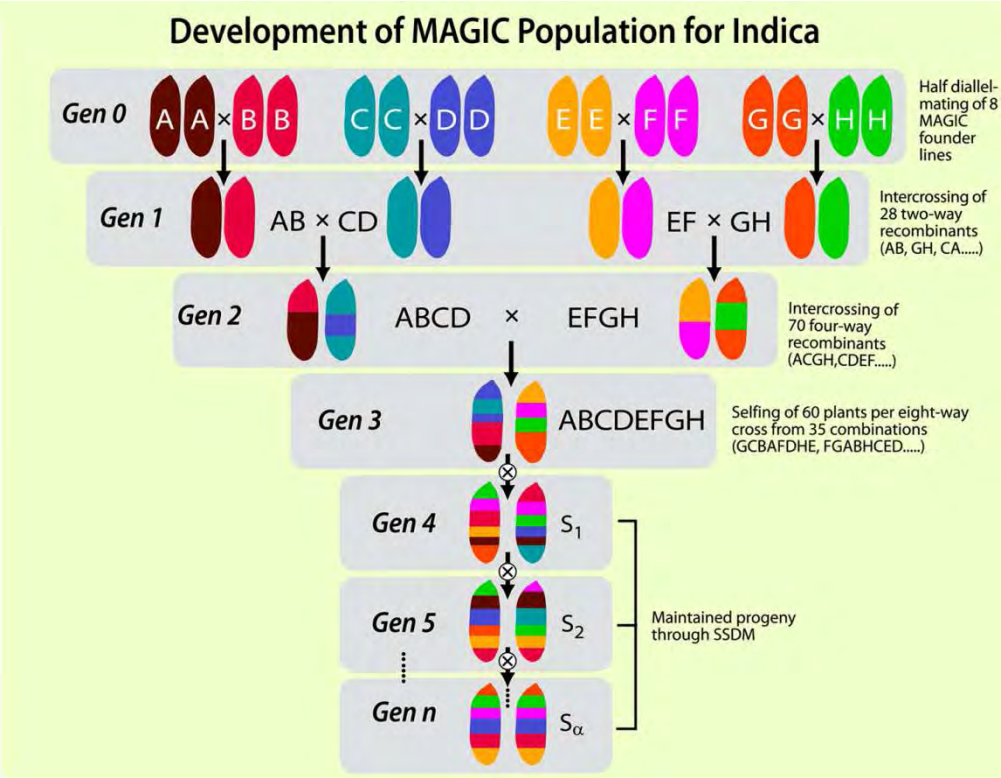


Figure III.3. Crossing scheme to produce multi-parent advanced generation inter-cross (MAGIC) population (Bandillo *et al.,* 2013).

III.2 Association studies in sunflower

III.2.1 Three association mapping researches on sunflower

Various papers have already reported linkage disequilibrium studies on sunflower (Kolkman *et al.*, 2007; Fusari *et al.*, 2008). For example, Kolkman *et al.*, (2007) estimated SNP density across the sunflower genome (~3500 Mbp). They predicted that sunflower harbors at least 76.4 million common SNP among modern cultivar alleles. They also show for their panel that in the inbred lines LD level declined to 0.32 by 5.5 kbp and that this decay happened slower in inbred lines than in wild population due to history of domestication and breeding pressure. These studies revealed the potential of linkage disequilibrium mapping studies on this species thanks to the sufficient SNP frequencies and LD decay in modern sunflower cultivars. Indeed, several association studies conducted on sunflower have then followed and confirm this. Fusari and co-workers (2012) developed an association mapping approach to detect loci involved in Sclerotinia head rot resistance. For this first study, a collection of 94 sunflower inbred lines was used in a candidate gene strategy (43 genes). Another association study with a genome-wide strategy was then developed for the detection of loci involved in branching and flowering time (Mandel *et al.*, 2013). In this work, 271

sunflower lines were genotyped on an Illumina Infinium 10k SNP array. In a third study, Cadic and co-workers (2013) combined results of an association mapping study and of a linkage mapping approach to identify QTL involved in the control of flowering time in sunflower. Using these two complementary approaches allowed the authors to overcome the downsides of each one, explore more environments and therefore produce robust association results. As the same sunflower association panel is used in this work, we will use the next sections to give more detail on results concerning LD and structure of this core collection.

III.2.2 Association panel used and described in the work of Cadic et al., 2013

To achieve their work that concerned a core collection of 384 sunflower genotypes, Cadic and co-workers (2013) evaluated the linkage disequilibrium as well as the structure of their panel and the kinship between each pair of genotypes. Linkage disequilibrium decay ranged from 0.08 to 0.26 cM, after correcting for a structure effect, depending on the linkage group (LG) and the status of the inbred lines (B-lines: maintainers of cytoplasmic male sterility or R-lines: fertility restorers). Figure III.4 shows the LD decay in this panel.

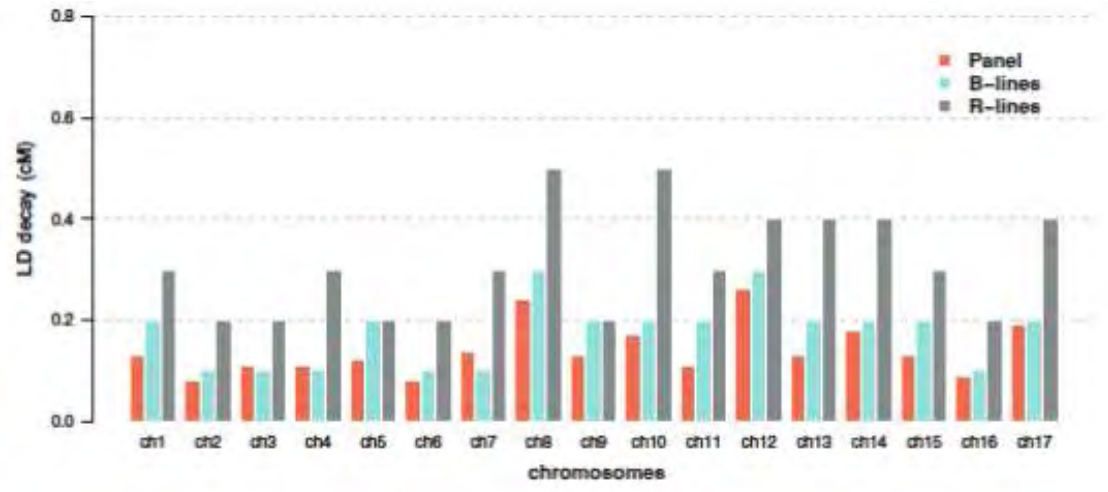


Figure III.4: Distribution of LD decay across chromosomes for the entire panel and for each breeding pool (B-lines: maintainers of cytoplasmic male sterility and R-lines: fertility restorers).
LD decay was calculated using the r2vs statistic (Mangin *et al.*, 2012) that includes correction for the structure effect (Cadic *et al.*, 2013).

Analysis of the panel structure thanks to a Principal Component Analysis (PCA) revealed differences between R-lines and B-lines (Cadic *et al.*, 2013). Indeed the first component of the PCA, explaining 5.91% of the variability, separated the B-pool on the right side and the R-pool on the left side (Figure III.5).

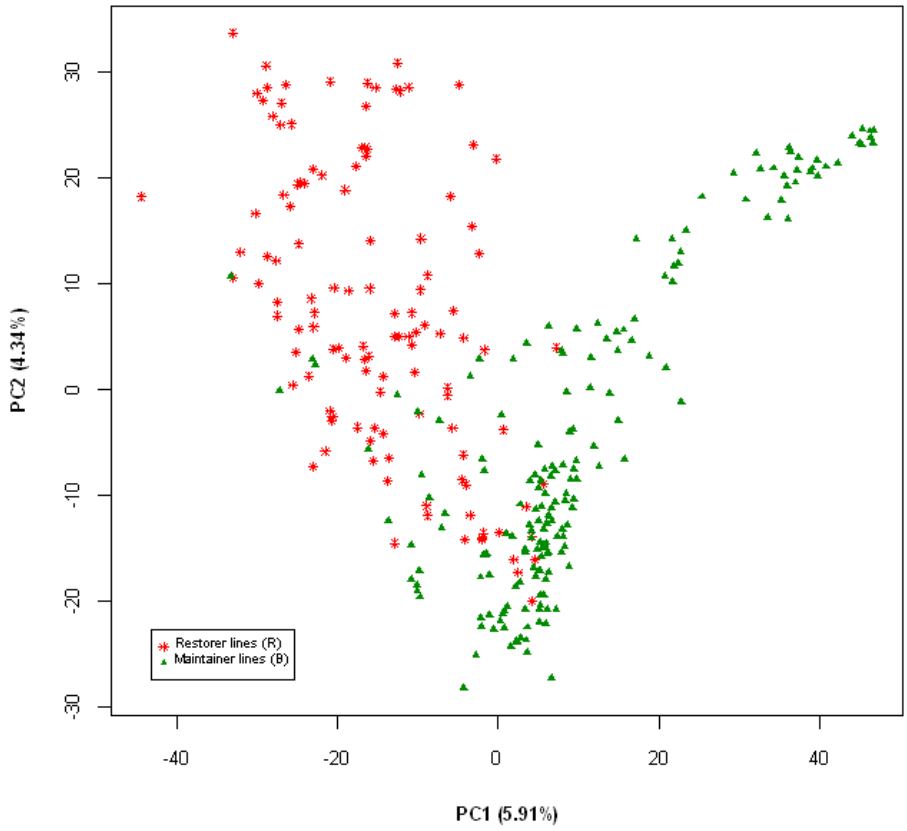


Figure III.5: Structure of the association panel used in Cadic et al., 2013.
The top two principal components of the PCA analysis are represented. Percentages in parentheses refer to the proportion of variance explained by the principal components. Symbols represent the two breeding pools: x for R-lines (red) and triangles for B-lines (green). (Cadic *et al.*, 2013)

Therefore, the structure used in the association model of the study of Cadic *et al.* (2013) reflected the belonging of the lines to the B- or R-pools.

To study the genetic control of drought response through the control of the expression of genes that support physiological and morphological traits involved in water deficit tolerance (Figure III.6), we carry out an association mapping study on a selection of these genes. This work uses a subset (N=275) of the same association panel described and studied in the work of Cadic *et al.*, (2013). Then, we base our work on their results that provides the necessary knowledge about population structure and linkage disequilibrium in this panel and, as well as the best association model to use.

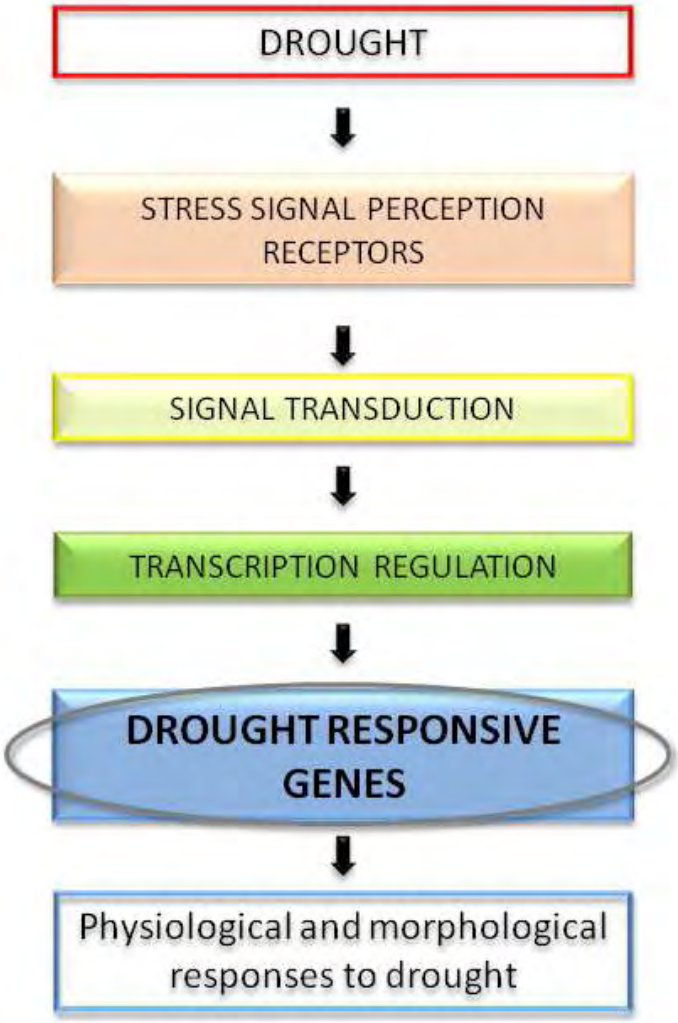


Figure III.6: Genes studied for their response to drought stress and inducing morpho-physiological response to water deficit.

III.3 Issues and challenges in the study of genes correlated to water stress responses

The goal of the work described in the next part of this chapter is to study the genetic control of genes whose expression is correlated to main morphological and physiological traits involved in drought tolerance, such as, for example, relative water content (RWC) or the evapotranspiration (ET).

The gene expression is an upstream phenotype which not accounts for the final, integrated and complex phenotypic response. However, due to the strong improvement of molecular technologies aiming to quantify the transcription for a large number of samples and genes, characterizing the phenotype through a transcriptomic quantification of genes related to the complex traits, appears now as a competitive approach when compared to physiological phenotyping. Moreover, this approach will allow reinforcing the bridge of knowledge between the genotype and its specific phenotypic response.

Therefore, in this chapter we describe and present the results of an association mapping study on these gene expression levels in order to unravel the gene regulatory network that controls them. This will also permit us to identify cis- (proximal) and trans- (distant) regulations of those gene expressions. The second aspect of this work is to highlight the differences between the genotype x environment interaction effect and genotype effect of genetic variants on gene expression. We will then address the distinction between the part of the drought stress response that is plastic (changing in function of the environment water status) and that which is solely genotypic. To answer this question we will use the Water Status Biomarker described in the previous chapter in order to normalize water status of the genotypes and compare them in a similar environment regarding drought stress.

III.4. Article: Integration of the environment in gene regulatory networks. Identification of plastic regulations in the case of drought stress in sunflower via an association study on gene expression.

Article status: **to be submitted**

Authors

Gwenaëlle Marchand, Baptiste Mayjonade, Stéphane Muños, Didier Varès, Nicolas Blanchet, Brigitte Mangin, Patrick Vincourt, and Nicolas B. Langlade

III.4.1.Abstract

Organisms adapt their phenotype to the environment through the gene regulatory networks (GRN) that control their development and physiology. Regulation modifications in these networks can code constitutive expression differences between species or genotypes but also differences in the expression modulation i.e. plasticity. A better understanding of the genetic control of complex traits and knowledge of the part in which the plasticity is involved, is of particular interest for fields as physiology, evolution, and breeding. The approaches that treat transcript abundances as quantitative traits appeared to be a way to decipher the genetic control of those complex traits and their underlying gene regulatory network. We studied the genetic basis controlling the expression variation of 86 genes previously shown to be involved in various morpho-physiological responses to drought stress in sunflower (*Helianthus annuus*). This was achieved through an association mapping approach on a panel of 275 sunflower hybrids grown and studied in field conditions. The water status of each genotype was estimated with a Water Status Biomarker, related to the pre-dawn leaf water potential, and was exploited as an environmental covariate. This allowed to perform the genetic analysis using two association models. The first one did not correct for the environment and compared genotypes at different water statuses. The second model corrected the environmental effect and therefore compared genotypes at similar water status. Comparison of both models gave access to the genotypic and genotype-environment interaction parts of the gene expression genetic control. Indeed, three genes showed significant plastic responses to drought intensity. From this analysis, we constructed a gene regulatory network linking 78 genetic loci to 33 gene expression levels correlated to 6 morpho-physiological responses. This systems biology approach integrated genetic and transcriptomic data to characterize which part of GRN allows phenotypic plasticity and species adaptation to new environments.

III.4.2. Authors summary

Expressions of genes underlying complex traits are modulated in function of (1) genetic variants i.e genotypes under same environment could have different phenotypes, (2) environment variation i.e a single genotype could present different phenotypes under different environmental conditions and (3) genotype x environment variation i.e each genotype do not vary in the same way following the environment. The two last parts represent the phenotypic plasticity of the trait. A better understanding of the genetic control of such complex traits, as for example drought tolerance, is of particular interest for different disciplines. As transcript is the first step between genome information and phenotypes, the utilization of transcript abundances as quantitative traits appears as a powerful way to understand genetic control of complex traits and the underlying gene regulatory network.

In this context, we studied the genetic basis of expression levels variation of 86 genes correlated to drought tolerance on sunflower through an association study on 275 sunflower genotypes. We evaluated the water environment perceived by each genotype and we compared results of two statistic models to estimate the genetically-variable part of the plasticity in the genes' expression regulation. Finally, we could reconstruct a drought gene regulatory network that links the genes, correlated to molecular and physiological processes to drought responses, to the genomic loci involved in their control. Moreover the influence of the environment in the genetic control could be identified.

III.4.3. Introduction

Phenotypic variation within a species shows a large diversity for plant traits as morphology, physiology or disease susceptibility. These complex traits, with very important phenotypic variation, are the product of a genetic control with multiple loci interacting between them and with the environment (Mackay *et al.*, 2009). To link DNA sequences variation to the diversity of phenotypes, different genetic approaches as linkage mapping (Lander & Botstein, 1989) or association mapping (Remington *et al.*, 2001) have been used before. These studies allowed the identification of many regions in the genome involved in complex traits control (Mackay, 2001) but the complex gene regulatory network that links genes and phenotype remains largely unknown.

Gene transcription is the first molecular step between genome information and the final, integrated, complex phenotype. So, changes in transcription levels are generally considered to be essential factors that reflect, at least for a part, the production of different phenotypes. Variation in gene expression levels were shown to be highly heritable (West *et al.*, 2007) and transcript abundance of a gene can be considered as a quantitative trait (Brem *et al.*, 2002). Therefore, Jansen and Nap (2001) introduced the idea of genetical genomics, in which linkage or association mapping

could be applied to gene expression levels. This new approach allowed the identification of loci in the genome underpinning the observed variation in transcript abundance and then the reconstruction of the gene regulatory network that controls the complex physiological traits.

The application of this new strategy has been allowed by the progress of genotyping technologies on one hand and transcriptome arrays on the other hand.

Several linkage mapping studies on gene expression have been performed in a variety of organisms: first on yeast (Bing & Hoeschele, 2005; Brem *et al.*, 2002), as it is a well studied organism for gene expression, but also on plants, with several studies on maize (Schadt *et al.*, 2003) and on the model plant *Arabidopsis thaliana* (West *et al.*, 2007; Cubillos *et al.*, 2012b; Keurentjes *et al.*, 2007). These studies highlighted and described thousands of expression QTL (eQTL) with, as for classical traits, a large variation in the number of controlled transcripts for each locus.

However, Cubillos and co-workers (2012b) demonstrated that transcriptome architectures were moderately conserved between crosses for the plant model *Arabidopsis thaliana*. This emphasizes the need for new studies, taking into account a larger genotypic diversity, in order to better understand the transcriptomic control within a species and produce a regulatory network integrating differences between a large diversity of genotypes.

In this context, association mapping on gene expression appeared to be a promising strategy. This approach has been largely used on Human, for example to better understand the cellular biochemical processes associated to susceptibility loci for complex diseases such as diabetes (Schadt *et al.*, 2008) or degenerative diseases (Dixon *et al.*, 2007). One genome-wide association study (GWAS) has been performed for the plants on *Arabidopsis thaliana* (Gan *et al.*, 2011), in which whole genome seedling transcriptomes were used on a small association panel of only 19 accessions.

If gene expression levels are highly heritable, the interaction between genotype and environment is also an important factor that could modify transcript abundance (Smith & Kruglyak, 2008) and this issue has to be taken into account particularly when studying the expression of genes related to responses to biotic and abiotic stresses. Indeed, understanding the ability of plants to adapt to their environment is a major issue that could have numerous applications in several fields as physiology, evolution or crop breeding. Plant response to environment also referred as phenotypic plasticity can be defined as the ability of a genotype to produce multiple phenotypes in response to the environmental variations (Des Marais *et al.*, 2013). In his study, Bradshaw (1965) highlighted the importance of genetic variation in plasticity, which was then measured as a genotype x environment interaction (GxE). Since this first study conceptualizing GxE interactions, there is now accumulating evidence that GxE interactions are very common and account for a non negligible part of the phenotypic variation (Grishkevich & Yanai, 2013). Because gene expression variation could be considered as phenotypic quantitative traits, study of their plasticity and in particular of their GxE

interactions could be of great interest to analyze gene regulatory network involved in abiotic plant responses. For example, it will help to answer questions about the molecular and genetic mechanisms that underlie gene expression plasticity, since all genes do not show equivalent plasticity. In their review, Grishkevich and Yanai, (2013) reported that promoter architecture, expression level and regulatory pattern correlate with the differential regulation of a gene by the environment. The understanding of these different components could therefore help to breed and construct novel ideotypes based on more accurate model predictions for environmental adaptation.

In this study, we performed a GWAS on expression levels of genes involved in drought stress responses for sunflower (*Helianthus annuus*). Drought is one of the major environmental stresses. It limits the productivity of all major crops, and is expected to become more frequent and widespread in the future. It affects the expression of numerous genes that are the first step toward changes in the morphology and in the physiology of the plant (Bhatnagar-Mathur *et al.*, 2008). The GWAS strategy allowed us to take into account the large genetic diversity of *Helianthus annuus* through an association panel, and we studied the genetic variation of plasticity thanks to a newly developed water status biomarker in order to evaluate the drought stress perceived by each genotype. This allowed us to make the distinction between the genotypic effect, constitutive of the genotype, and the GxE effect, corresponding to the plastic part of the response to the environment i.e. the genetically-variable part of the plasticity.

III.4.4. Results

III.4.4.1. Estimation of drought stress perceived by each genotype in the association panel

A core collection of 384 sunflower inbred lines was built using a nested core collection strategy from an initial set of 752 inbred lines (Cadic *et al.*, 2013). The association panel used in this study contains 275 inbred lines and is a subset of the initial core collection described above. The lines of this panel were crossed with two testers according to their status (maintainers of cytoplasmic male sterility “B-Lines” or fertility restorers “R-Lines”), and grown in agronomic conditions in Villenouvelle (Haute-Garonne, France) during summer 2011. The field experiment design was formed of blocks with 24 entries replicated in two sub-blocks. Each sub-block was randomized separately and contained two check hybrids: Melody and Pacific. The field trial was conducted without irrigation.

As we aim at finding the genetic architecture of drought stress responses, it was important to determine the water stress perceived by the plants. To estimate water status of the 275 genotypes of the association panel, we calculated the Water Status Biomarker $WSB_{\Psi PD,}$ i.e. a biomarker that

estimates the pre-dawn leaf water potential (Ψ_{PD}) from a linear combination of three gene expression levels (Marchand *et al.*, 2013). Ψ_{PD} is a classical indicator of plant water status that evaluates water available in the soil for the plant. Values of the $WSB_{\Psi PD}$ observed for the association panel genotypes in field conditions after a period of moderate drought in a deep soil corresponded to a range of Ψ_{PD} between -0.38MPa and -1.52MPa. This variability could be due to (1) the variability of the environment i.e. the position of the genotype in the trial design, (2) the specific morphology and physiology of the genotype involving that, for the same amount of water in the soil at the beginning of the experiment and the same meteorological conditions, different genotypes had different access to soil water and consumed it more or less rapidly. Consequently, at the day of harvest, the 275 genotypes of our panel had different water statuses and perceived different drought stresses.

The Ψ_{PD} values of the two check hybrids, Pacific and Melody, repeated at each block of the experimental design ranged from -0.55MPa to -0.82MPa and from -0.54MPa to -1.0MPa respectively (Figure III.7.a). This result showed that the variability due to the spatial variation in the field could not be neglected for the association study.

We then corrected the water status $WSB_{\Psi PD}$ of each genotype by the block effect (calculated in an analysis of variance, ANOVA) that captured the spatial variation but was not due to rainfall differences (Appendix III.1). For all the genotypes, the normalized values of $WSB_{\Psi PD}$ corresponded to values of Ψ_{PD} with a reduced standard deviation of 1.09 (instead of 1.34 for the raw data) and range from -0.33MPa to -1.29MPa (Figure III.7.b).

These results showed that the panel genotypes accessed and used water in different ways although they were subjected to the same external environment. The observed variation in water status (i.e. the environmental factor) resulted from the genotype x environment interaction. It results in a range of environments that can be used to correct our variables of interest and access their genetic control, but it also can be exploited to reveal the dynamic response to water availability in sunflower.

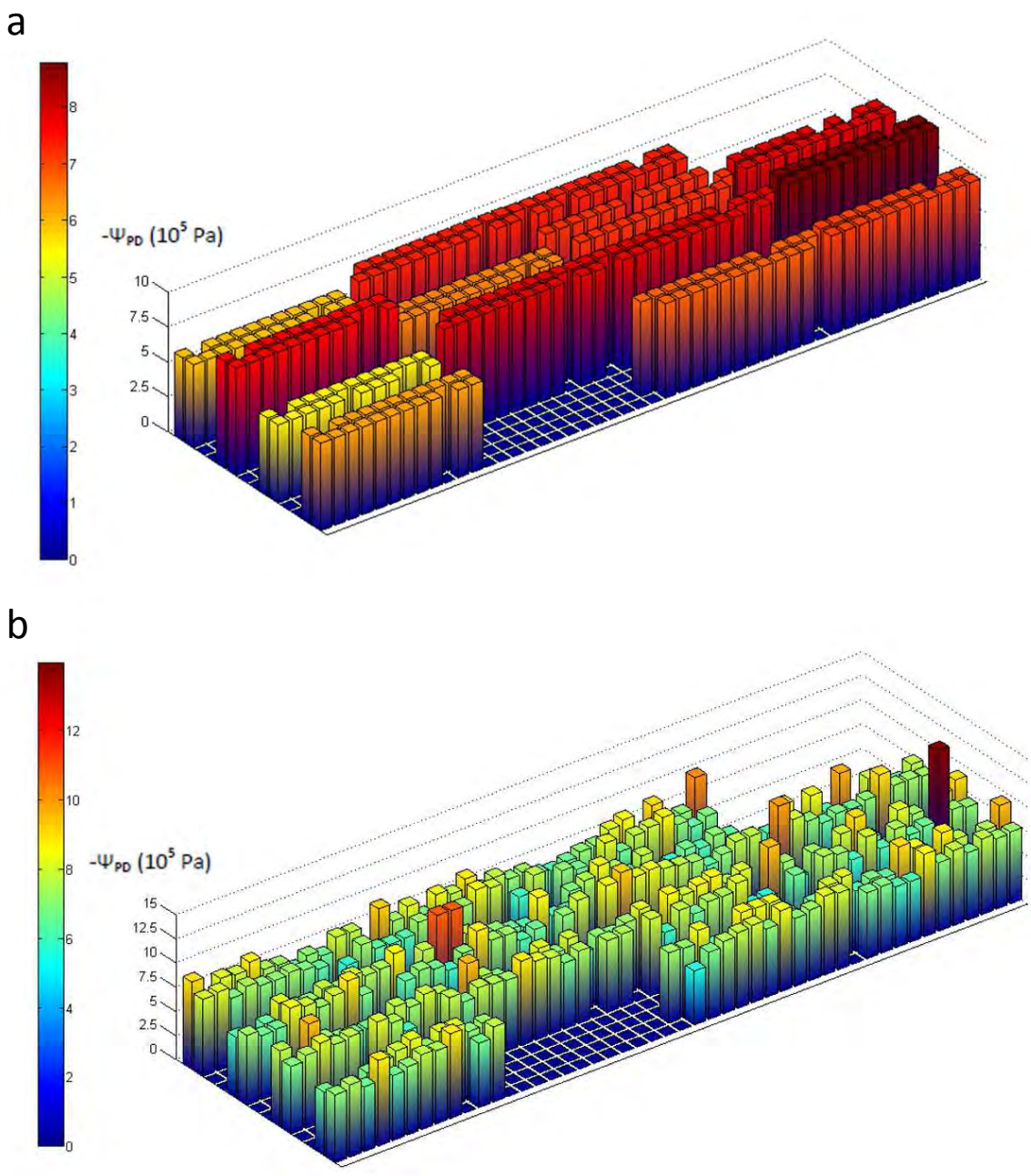


Figure III.7: Variability of the water status perceived by the genotypes in the field trial

Each bar represents a field plot with one genotype. Bar height and color represent the estimated pre-dawn leaf water potential (Ψ_{PD}) i.e water status perceived by the genotype. (a) Water status perceived by the check genotype Melody, without correction for spatial variation in each block. (b) Water status perceived by each genotype after correction for spatial variation.

III.4.4.2. Selection of genes reporting drought responses

From a previous study (Rengel *et al.*, 2012), we selected genes found to be correlated, using a sparse partial least squares approach using mixOmics (González *et al.*, 2011), to (i) Carbon Isotopic Discrimination (CID), (ii) Osmotic Potential (OP), (iii) Evapotranspiration (ET), (iv) Relative Water Content (RWC), (v) Specific Leaf Area (SLA), and (vi) Total Leaf Area (TLA) in two different drought scenarios conducted in controlled conditions (R^2>0.65). In addition to those genes, we also selected genes that showed a Genotype x Environment interaction effect in the ANOVA performed in (Rengel *et al.*, 2012) with 8 genotypes and a fixed intensity of stress scenario. We finally studied the expression of 86 transcripts. Detailed descriptions of these genes are presented together with all the genes in this study in Appendix III.2.

III.4.4.3. Gene expression data analysis

The expression levels of the 86 transcripts were determined by qRT-PCR on the 275 genotypes of the panel and on the check genotypes included in each block. For each gene expression level, the best linear unbiased predictors (BLUP) of genotypes were calculated for two models as described in the Materials and Methods section. The first model only corrects the spatial variation in the field and compares genotypes in different environments. The resulting BLUPs captured together the genotypic, the environmental and the genotype-environment interaction effects on the gene expression. This first model was therefore noted GE. The second model corrects both the spatial variation in the field and the environmental effect and compares genotypes in a similar environment (Figure III.8). Accordingly, the resulting BLUPs reflect mainly the genetic control of the studied gene and the model was noted G.

Using the GE model, we calculated BLUPs for 86 gene expressions and the WSB. Out of the 86 genes, 17 did not vary when corrected spatially, and the genetic control of their expression could not be performed. Similarly, for the G model, we calculated BLUPs for the 86 gene expressions. Among them 70 were found with BLUPs different from zero. All the BLUPs values are reported in the Appendix III.3.

In total, 140 variables were studied: the WSB and 69 BLUPs of gene expression in the GE model, and 70 in the G model.

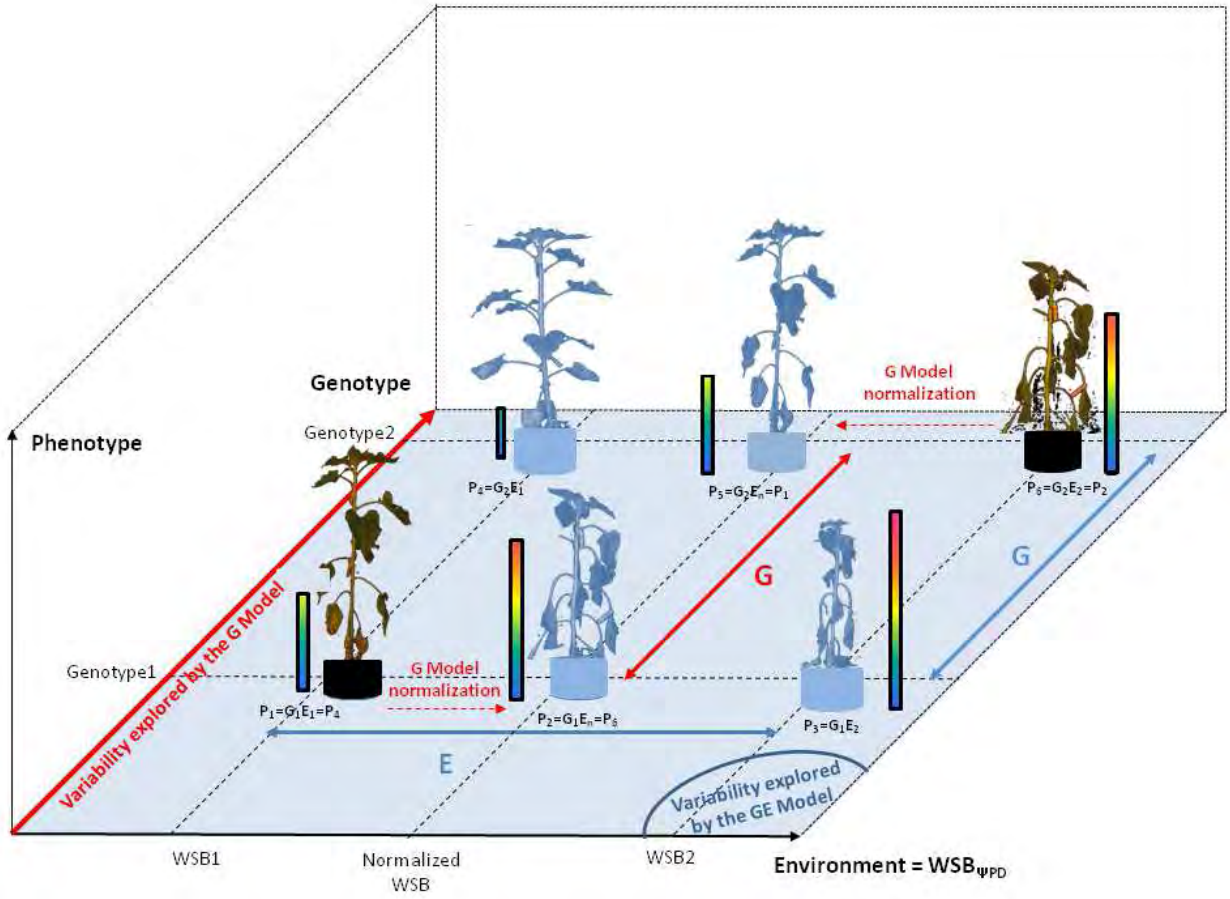


Figure III.8: G and GE models explanation.

The variability captured by the GE model is due to the genotype (y-axis) and to the water environment (x-axis) and therefore represented by the blue plane. Normalization in the G model permits to compare different genotypes at the same water status. Therefore the variability explored by the G model is only due to a genotypic effect and is represented only by one dimension (y-axis). Phenotype is the result of the G and GxE variability and is represented in the z-axis by schema of plants and vertical bars whose height and colors represent expression level of a gene with under the control of genotypic, environmental and GxE effects.

III.4.4.4. Association mapping

Association analysis

To perform association tests between gene expression variation and SNP variation, we selected 62,820 SNPs showing polymorphism for the association panel with MAF > 5% and without redundancy out of the 197,914 SNPs of the sunflower AXIOM array (see Materials and Methods).

The association model was chosen according to the study of Cadic *et al.* (2013), in which several models of association were investigated for the same core collection of sunflower. We used a mixed model that corrects for structure and kinship between the lines of the association panel. This model appeared to be the best association model for our panel according to BIC, p-value criteria, and

reductions of false positives (Cadic *et al.*, 2013). Associations with an adjusted FDR p-value <0.05 were considered to be robust.

Association studies with and without correction for the environment provide distinct results

For most of the gene expression levels, associations found with the GE model (with the correction for the environment effect) and the G model (without the correction) were similar. However, it was interesting to note that depending on the model considered, with or without the WSB correction, the adjusted p-values for the phenotypic traits were not exactly identical. Twenty-six gene expression levels had lower adjusted p-values using the GE model and 42 using the G model. Results of a paired t-test on adjusted p-values (Appendix III.4) showed that 62 gene expressions (91%) had different p-values between G and GE models and four did not shown significant differences between the two models.

Moreover, among the 66 gene expression levels studied in the two models, 33 presented significant associations: 26 with both models, and 4 and 3 only for the G model and the GE model respectively. Among these 26 genes, 18 did not present the same number of significant associations in G and GE model respectively.

These results suggest that the two models give complementary results and allow making a distinction between genotypic, GxE interaction, and environment control of the gene expression in order to understand the genetic architecture of drought stress responses.

Association mapping results

For the G model, 30 expression levels out of the 70 studied presented significant associations with 1 up to 443 markers. For the GE model, 29 expression levels and the WSB presented significant associations with 1 up to 437 markers. Table III.2 and Appendix III.5 summarize results of the association study.

The 1364 SNPs found in association were mapped on a consensus map of two genetic maps generated from two RIL populations named INEDI (XRQxPSC8) and FUxPAZ2 (see Materials and Methods). The SNPs were mapped thanks to different sources of information: INEDI and/or FUxPAZ2 mapping, Linkage Disequilibrium (LD) mapping and alignment comparison of markers context-sequences on genomic and transcriptomic sequences of the sunflower genotype XRQ. Finally, 649 SNPs could be positioned using the genotyping data information on INEDI or FUxPAZ2, 488 were mapped using LD information and 106 using sequence alignment comparison. In total, 1243 SNPs associated to 27 and 26 gene expression level in the G and GE models respectively (91% of the associated SNPs), were mapped on the consensus map and 121 remained unmapped. In each model, three gene expression levels were only associated with unmapped SNPs.

Gene ID	Correlated phenotypic variable	*Arabidopsis* name/ AGI	Model	Number of associated SNP	Most significant FDR adjusted pval	Number of QTL	Number of mapped QTL
HaT13l000239	OP	SEC, SECRET AGENT	GE	58	1.47E-05	6	4
			G	55	1.26E-05	6	5
HaT13l001185	RWC	AXS2, UDP-D-XYLOSE SYNTHASE 2	GE	5	3.66E-02	1	1
			G	0	_	_	_
HaT13l001663	OP	ATFC-II, FERROCHELATASE 2	GE	4	8.77E-03	2	2
			G	4	9.35E-03	2	2
HaT13l002091	CID	EMB1873	GE	41	4.29E-12	3	1
			G	41	4.38E-12	3	1
HaT13l002581	OP	ATSUC2, SUCROSE TRANSPORTER 1	GE	2	7.31E-03	2	1
			G	2	5.75E-03	2	1
HaT13l002627	ET	ASB1; TRP4; WEI7	GE	437	5.71E-10	6	6
			G	443	6.27E-10	6	6
HaT13l002719	GE candidate	KUP10	GE	24	5.47E-05	4	3
			G	25	3.73E-05	4	3
HaT13l002773	GE candidate	SRF3	GE	3	4.93E-05	1	1
			G	3	1.71E-05	1	1
HaT13l002800	CID	ATKRS-1, LYSYL-TRNA SYNTHETASE 1	GE	0	_	_	_
			G	1	6.64E-03	1	0
HaT13l002822	GE candidate	AT2G42490	GE	6	3.73E-03	1	0
			G	6	4.88E-03	2	1
HaT13l003718	ITW.RWC	BETA-6 TUBULIN, TUB6	GE	1	3.70E-02	1	0
			G	2	1.22E-02	1	0
HaT13l004212	ET	AT3G19320	GE	2	4.92E-02	1	1
			G	0	_	_	_
HaT13l005549	RWC	PMSR3	GE	4	1.12E-02	3	1
			G	4	1.10E-02	2	1
HaT13l006786	ET	AT2G22420	GE	114	8.14E-13	4	3
			G	113	1.42E-12	3	2
HaT13l007963	ITW.RWC	AT3G18050	GE	36	3.25E-05	4	3
			G	34	5.60E-05	3	2
HaT13l008198	GE candidate	MBR2	GE	16	1.23E-02	6	5
			G	15	1.63E-02	6	5
HaT13l008549	GE candidate	SARK	GE	0	_	_	_
			G	25	4.05E-02	2	2
HaT13l009999	CID	SP1L4, SPIRAL1-LIKE4	GE	181	3.32E-15	15	10
			G	179	7.92E-15	14	9
HaT13l010540	CID	AT5G47390	GE	1	2.63E-02	1	1
			G	1	2.33E-02	1	1
HaT13l011270	CID	AT1G76020	GE	76	2.30E-07	9	8
			G	76	2.26E-07	9	8
HaT13l011662	CID	FAD2, FATTY ACID DESATURASE 2	GE	3	9.05E-03	2	2
			G	3	8.16E-03	2	2

HaT13l012070	CID	PAT1	GE	10	9.23E-03	1	1
			G	10	8.11E-03	1	1
HaT13l013507	ET	ABH1	GE	3	3.13E-02	1	0
			G	3	3.36E-02	1	0
HaT13l013529	OP	CRK5; RLK6	GE	25	5.48E-04	3	3
			G	31	3.47E-04	3	3
HaT13l016627	CID	AT5G42250	GE	0	_	_	_
			G	5	1.67E-02	2	1
HaT13l025285	OP	TINY2	GE	2	4.41E-03	2	1
			G	6	6.68E-03	4	2
HaT13l033242	ET	AT1G78070	GE	0	_	_	_
			G	6	4.12E-02	3	1
HaT13l059347	ET	PLDALPHA1	GE	16	1.33E-02	6	4
			G	17	2.00E-02	6	4
HaT13l060757	GE candidate	_	GE	1	4.96E-02	1	1
			G	0	_	_	_
HaT13l068709	CID	_	GE	168	9.31E-13	7	5
			G	168	4.82E-13	7	5
HaT13l200063	OP	ACHT4	GE	6	9.96E-06	2	1
			G	6	8.47E-06	2	1
HaT13l200627	GE candidate	EDF4	GE	5	2.83E-04	1	1
			G	4	6.71E-05	1	1
HaT13l201322	RWC	AT1G22930	GE	5	1.64E-02	2	2
			G	10	9.70E-03	3	3

Table III.2: Summary of associations and QTL detected

III.4.4.5. QTL detection and identification of cis- and trans-regulations

On the INEDI, FUxPAZ2 and consensus maps, SNPs associated to the same trait and distant from less than 5 cM from the next associated SNP were considered to form one single QTL. This binning allowed us to account for Linkage Disequilibrium between SNPs and to synthesize the genetic information for further analysis. Again, if two or more genes were controlled by adjacent QTL, those QTL were considered to be a single one if they were distant from less than 5cM.

In total, 50 QTL were found for all gene expression levels (corrected by both models) and were placed on 16 out of the 17 different chromosomes of sunflower. In addition to those mapped QTL, we found 12 QTL with only LG information and we grouped unmapped associated SNP in 16 other QTL depending on the trait associated.

Considering only the mapped QTL, between 1 and 10 QTL per trait were detected (TableIII.2). The expression levels of HaT13l009999 and HaT13l011270 were found associated with the highest number of QTL (10 and 9 for the first one in the GE and G model respectively and 8 in both models for the last one). Mapped associations for these traits are represented in Figure III.9. Linkage group 14 had the highest number of regions in association with 7 QTL spread over 101.8cM. A hot spot spanning 14.3 cM could be identified on linkage group 16 (QTL16_46) where it was associated to 10 gene expression levels (5 genes found with expression associated in both, GE and G models).

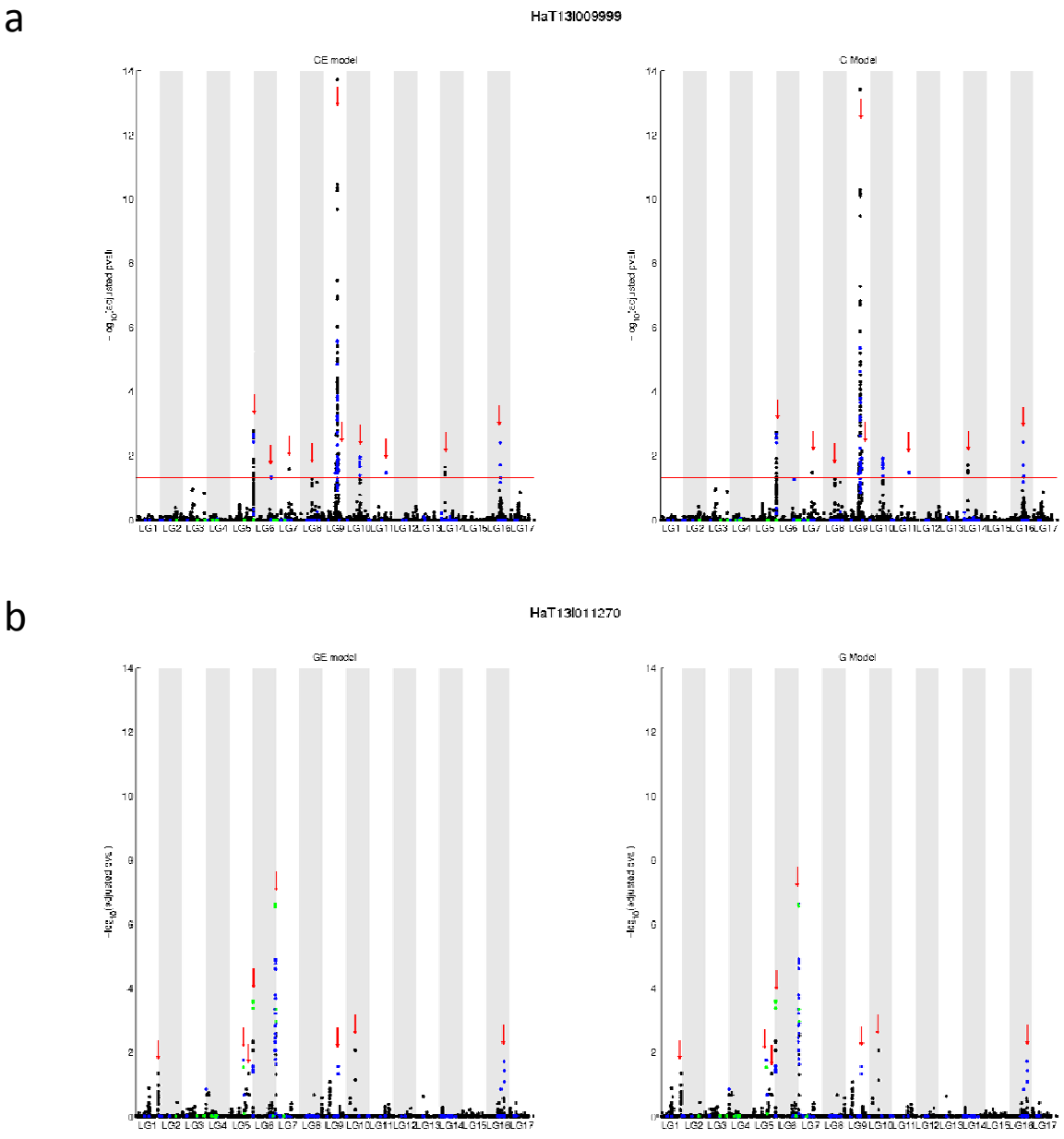


Figure III.9: Manhattan plot representing FDR adjusted p-values for the association between gene expression levels corrected or not by the environment and mapped SNPs.

Point color represents the type of information used to map the SNPs: black points correspond to the INEDI or FUPAZ2 RIL maps information, blue points to linkage disequilibrium information, and green points to context-sequences comparison. Red arrows represent QTL as defined in the Material and Methods. (a) Manhattan plot of the gene HaT13l009999 for the G model (left) and the GE model (right). (b) Manhattan plot of the gene HaT13l011270 for the G model (left) and the GE model (right).

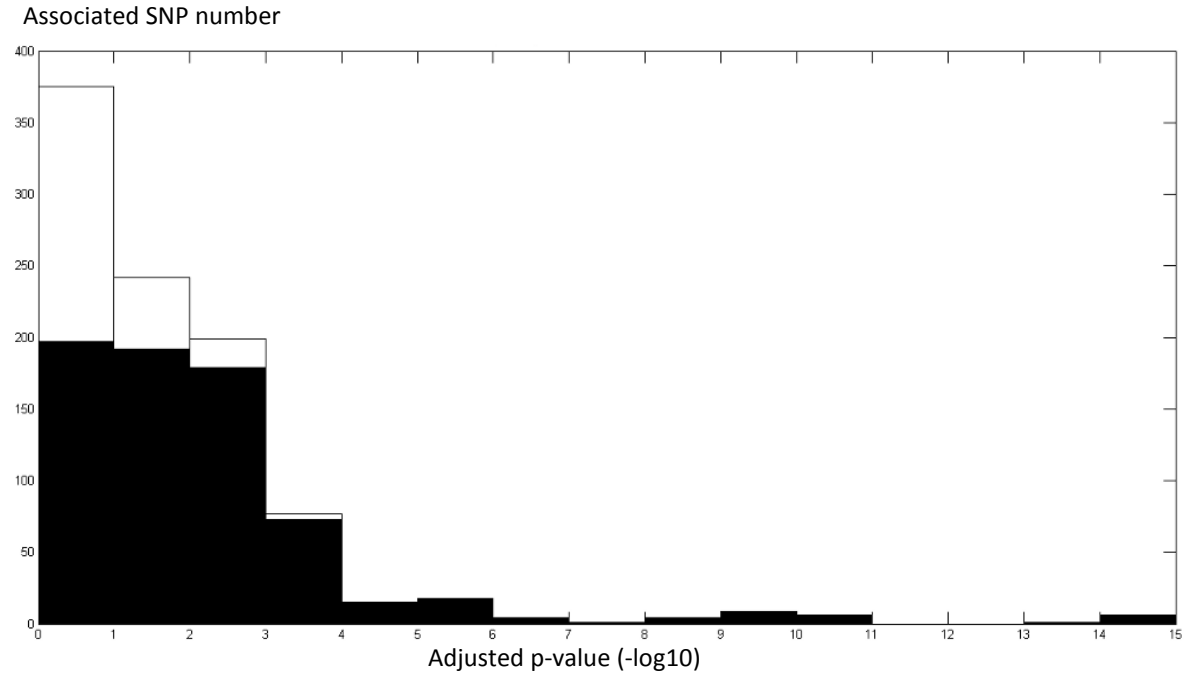
Comparison of the studied gene position in the genome and QTL locations allowed to make distinction between associations with loci near the gene Open Reading Frame (ORF) that indicate local regulations hereafter called cis-regulations and associations with loci far from the regulated ORF that indicate distant regulations, hereafter called trans-regulations.

To group the associations corresponding to cis- and trans-regulations, we either directly mapped the genes studied for expression level or we used the homology between the context-sequence of the markers with sunflower transcriptomic available in the public database Heliagene (https://www.heliagene.org/HaT13l). Among the genes whose expression levels were in association, 21 presented polymorphism between XRQ and PSC8 (INEDI population parents) or FU and PAZ2 and then could be mapped on the consensus genetic map. A regulation was considered to be a cis-regulation if the gene was distant from less than 10cM of the QTL associated to its expression level.

Finally, we found 22 QTL in cis (corresponding to 11 genes with association in the two models), 115 QTL in trans and 64 QTL of undetermined type. QTL of undetermined type are due to the fact that we were missing the exact gene or marker position information and marker context-sequence could not be aligned with sequences on sunflower transcriptomic database Heliagene. However, QTL found in local association had more significant p-values and grouped from 4 and up to 423 SNPs. Less SNPs were found in distant associations. Actually, the trans-QTL included between 1 and 32 associated SNPs to the gene expression.

We observed that the cis-associations were more significant than the trans-associations with adjusted p-value ranging from $9.23x10^{-3}$ to $3.32x10^{-15}$ and from $4.49x10^{-2}$ to $1.71x10^{-5}$ respectively. The Figure III.10 shows the distribution of the adjusted p-value comparing trans- and cis-regulations on one hand and trans- and cis-QTL on another hand.

a)



b)

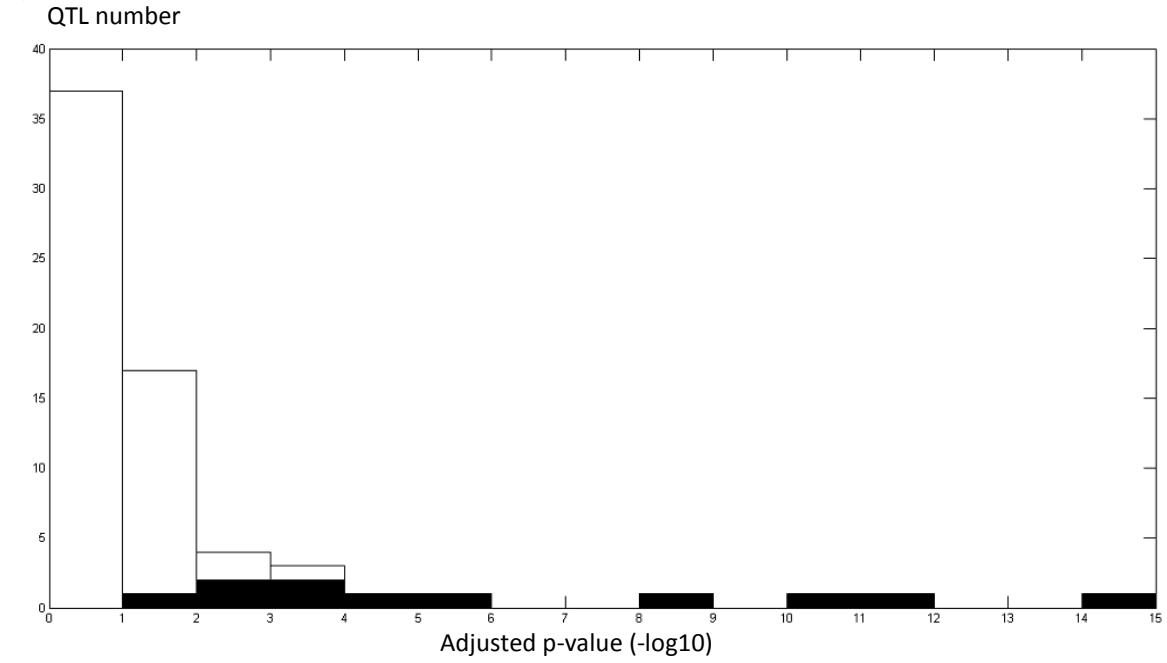


Figure III.10: Distribution of p-values adjusted by FDR of cis- and trans-regulations.

Trans-regulations are represented in black and cis-regulation in white. Results of both G and GE models are grouped. (a) Distribution of associations (SNP). (b) Distribution of the QTL.

III.4.4.6. Comparison of genotypic effect between G and GE models

The genotypic effects of the 1364 associated SNPs and their variance were calculated (see Material and Methods) and are presented in Appendix III.6. Genotypic effect of an SNP was considered significantly different between G and GE model if the confidence intervals of its effects in the two models did not overlap. Among the 33 genes studied for gene expression level and with significant associations in at least one model, we found that the associations between HaT13l002800, HaT13l010540 and HaT13l008549 expression levels and SNPs were significantly different between G and GE models. This shows that for these three genes, the two models allow to identify and evaluate the importance of the genotypic and GxE control of their expression.

III.4.4.7. Building a gene regulatory network

From this overall analysis, we were able to build a network that represented links between QTL on one hand and gene expression on the other hand (Figure III.11).

In this gene regulatory network, the source nodes are the QTL and the target nodes are the genes (expression levels), that we characterized to be correlated to phenotypic traits such as osmotic potential (OP), transpiration (ET), carbon isotopic discrimination (CID) or relative water content (RWC). This GRN was composed in total of 111 nodes (78 for QTL and 33 for genes) and 201 edges. Approximately half of the associations were found using the GE model (98) and half with the G model (103). Among the 78 QTL detected, 65 controlled the expression of the same genes in the G and in the GE models as shown in Figure III.11.

However, we could observe 8 associations that were only detected with the GE and 12 with the G model. But these corresponding QTL were also associated with expression levels of other genes using the two models. Interestingly, 6 different QTL were only found associated to gene expression using the GE model and 8 others only using the G model.

Depending on the physiological variable to which they were found correlated, connectivities between the genes were different (Appendix III.7). For example, genes correlated to CID and ET were grouped respectively in 4 and 3 connected components. Genes correlated to these two physiological traits appeared to be regulated by the same QTLs. On the contrary, genes correlated to OP and RWC were not grouped together, each gene was controlled by a different QTL.

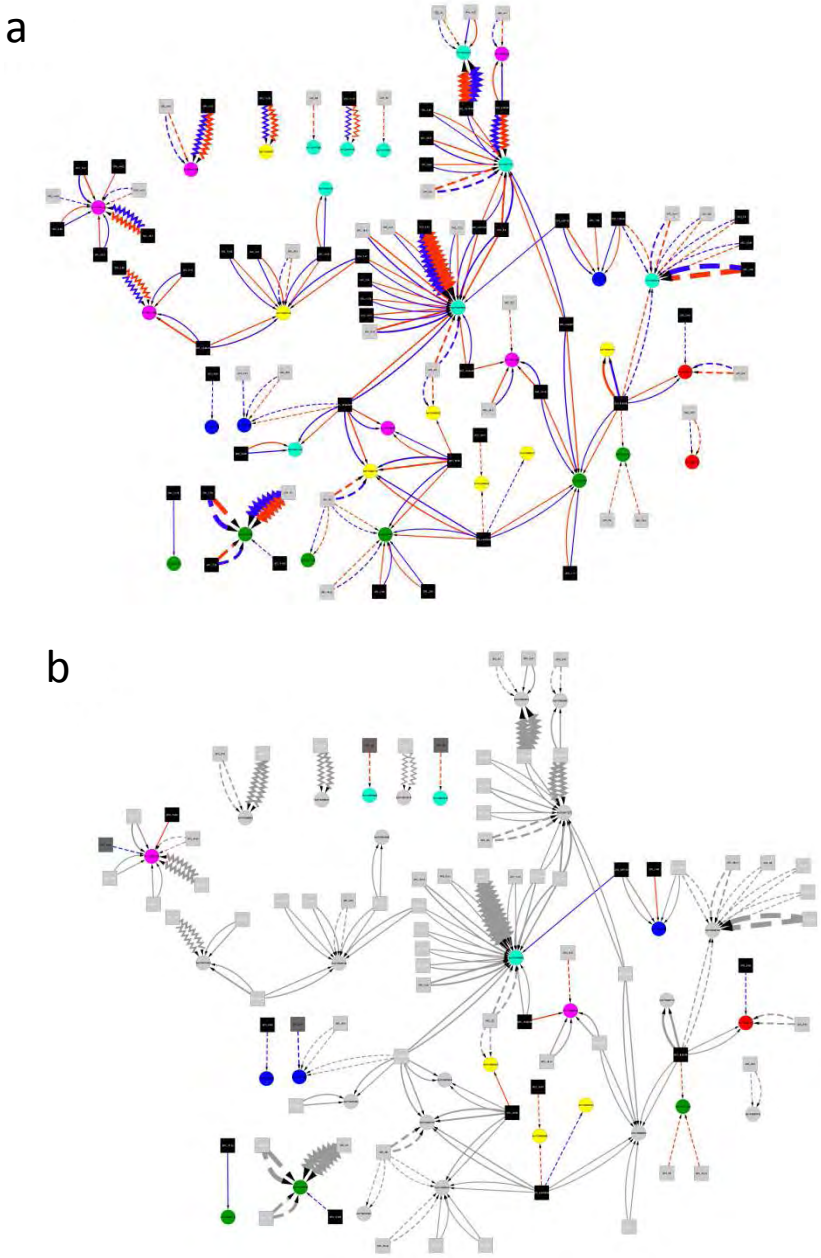


Figure III.11: Gene regulatory network for drought responses reconstructed from associations for gene expression levels corrected or not for the environment.

(a) Entire gene regulatory network. Black squares represent the mapped QTLs and grey squares unmapped ones. Circles represent the genes studied for their expression levels. Circle colors represent the phenotypic variable correlated to the gene in a previous study (Rengel *et al.*, 2012): Light blue represents genes correlated to CID, green represents genes correlated to Evapotranspiration, yellow represents genes with interaction genotype treatment in Rengel et al. 2012), pink represents genes correlated to osmotic potential, dark blue represents genes correlated to the RWC and red represents genes correlated to stress intensity and RWC. Red arrow represents associations found by the G model and blue arrows represents associations found in the GE model. Cis-regulations are represented by lines in zigzag and trans-regulations by solid lines. Finally, dotted lines represent regulations of unknown type. (b) Same gene regulatory network colored only with the associations found in one model only.

III.4.4.8. Association to the Water Status Biomarker $WSB_{\psi PD}$

During this study, we also identified through our genome scan, QTL regions that can account for the observed variation in the water status estimated by the Water Status Biomarker $WSB_{\psi PD}$. The $WSB_{\psi PD}$ is correlated to the pre-dawn leaf water potential (Ψ_{PD}) which is a classical water status indicator that estimates the water available for the plant in the soil. Eleven SNPs were found associated with $WSB_{\psi PD}$ and grouped in one QTL on LG4. As $WSB_{\psi PD}$ is an indicator of water status at the time of measurement, one possibility to interpret these polymorphisms is that they could be associated to genetic variation in the access and use of water before sample harvest.

III.4.5. Discussion

The genome-wide association study performed allowed us to identify several QTL associated to the expression levels of the selected genes. Those genes were chosen from the results of the previous study of Rengel *et al.* (2012) because they had their expression levels correlated to physiological traits linked to drought stress responses and could therefore be considered as proxies for these physiological traits. This strategy of GWAS on gene expression has already been performed in human (Stranger *et al.*, 2005; Cheung *et al.*, 2005; Dixon *et al.*, 2007) and on 19 genotypes in the plant model *Arabidopsis* (Gan *et al.*, 2011). The other studies using quantitative genetics to understand the transcriptomic regulation in plants are based on linkage mapping performed on RILs (on maize (Holloway *et al.*, 2011) and on *Arabidopsis* (Cubillos *et al.*, 2012a)). With this new GWAS, we showed that this approach can be also successful to understand transcriptomic regulation underlying complex phenotypic traits in plants. More interestingly, we aimed at understanding how the environment plays a role in gene regulatory networks. To achieve this goal for the particular case of genetic regulation of drought stress responses, we combined the classical GWAS with an estimation of water status perceived by each genotype in the association panel thanks to the use of a Water Status Biomarker, $WSB_{\psi PD.}$

With this approach, we were able to (1) identify the polymorphism associated to the control of the water status of the plant in the case of the GE model and (2) decipher the genetic architecture of the transcriptome regulation for the selected drought genes and make a distinction between genotypic and GxE interaction effects .

III.4.5.1. Sunflower controls water status of its micro-environment

In our experiment, the water status of the plants estimated through $WSB_{\psi PD}$ (Marchand *et al.*, 2013), showed a variability according to the genotypes. It demonstrated that even placed in a similar “starting” environment and subjected to identical crop management conditions, the genotypes of the association panel exploited their water reserves in different ways and placed

themselves in variable water statuses through their development. So, in the GE model, we used the genotypic variability of water use to generate a range of environmental situation (estimated with the $WSB_{\psi PD}$) and to decipher the drought response plasticity (Figure III.8).

Indeed, the plant water status $WSB_{\psi PD}$ can be considered as a classical phenotypic trait. Our results showed that $WSB_{\psi PD}$ was associated to eleven SNPs co-localized in one QTL. As in the GE model each genotype represented a different drought stress level, we can hypothesize that the identified polymorphisms might play a role in the use and access of water of the genotypes from the sowing to the time of sample harvest of this study. So, this result may highlight the fact that the plants can indirectly control their micro-environment through their physiology and development. Despite the fact that the identified loci cover a large region with very likely a lot of genes, we can hypothesize that causal genetic variants affect processes involved in the plant capacity to access and use the soil water. Therefore, those genetics variants play a role in the modification of the original amount of soil water. This could be seen as a dialog between the plant and its micro-environment to switch from one water status to another. (Figure III.8).

The $WSB_{\psi PD}$ has been calibrated and validated with a narrower genetic diversity than the one explored by the association panel. Therefore, we also have to envisage that the detection of SNPs associated to the $WSB_{\psi PD}$ could be due to the fact that the biomarker is not genotype-independent for the whole genetic diversity presented by the association panel. However, this verification is not experimentally tractable.

III.4.5.2. Identification of plasticity QTL thanks to the G and GE models

Genotype-constitutive and plastic parts of the drought stress response

We used two estimations of gene expression levels in our GWAS. In the G model, we corrected the gene expression levels with the $WSB_{\psi PD}$ in order to compare genotypes in a similar environment considering the drought stress perceived. The part of the regulation detected only with the G model represents the genotypic part, dependent only on the genotype and its alleles, and is not due to the GxE interaction. We propose to call this part of the QTL effect “genotype-constitutive” regulation, as it is not dependent on the drought stress intensity but only on the genetic variant at this locus. On the contrary, the GE model did not include correction for the effect of drought stress perceived by the plant and estimated by $WSB_{\psi PD}$. Associations in this model were therefore based on a genotype-constitutive effect, but also on a genotype x water status (GxE) interaction effect because the different genotypes were not at the same water status. In the GE model, the genotype-constitutive regulation is therefore modulated depending on the environment by this second component of the QTL effect due to the genetically-variable part of the plasticity that we will also call the “plastic”

response. In fact, with this GE model, we used panel genotypes diversity to examine GxE interactions as recently suggested by Grishkevich & Yanai (2013). With this method we try to understand the genetic control of plasticity for drought responses.

Throughout these last decades, the genetic control of plasticity has been the subject of several studies that developed several approaches to measure genetic variation for plasticity based on variation in slope, curvature, and other characteristics of the genotype reaction norms (Gavrilets & Scheiner, 1993). Simultaneously, various genetic models have been theorized in order to explain the genetic control of plasticity. Three models that are not mutually exclusive have been proposed (Scheiner, 1993). The terminology used to name the different models can be confusing as the terms are similar to standard terminology used in genetics but have a different signification. Therefore, the reader has to keep in mind this distinction when reading a discussion about plasticity models. The first model called *overdominance* hypothesized that plasticity is a function of homozygosity. In the *pleiotropy* model, plasticity is a function of differential expression of the same gene in different environments. Finally, in the *epistasis* model, plasticity is due to genes that determine the level of response to the environment and that interact with genes that determine the average expression of the trait: therefore plasticity of a trait is independent of its average value. The *pleiotropy* and the *epistasis* models are both supported by results of several studies, whereas, on the contrary, no evidence was found to support the *overdominance* model (Scheiner, 1993).

Our approach of GWAS on gene expression level, allowed us to decipher the genetic control of the phenotypic plasticity in a different way that the one described above (Scheiner, 1993). In our study, comparison of the results obtained with the G and GE models allows to answer the following questions: (1) is expression of the studied genes regulated by the environment and does it present GxE interactions? (2) what part of the genetic variability of the expression is due to the genotypic effect and what part is due to the GxE interaction effect in the GE model?

If the two models (G and GE) show association with the same SNPs, or QTL, but with significant differences in their effects, we can conclude that regulation of the gene is responsible for the plasticity of the controlled trait. In the case when the QTL showed no significantly different effects in both models, we can consider that, in the present experimental conditions and for the selected gene, the identified regulation was only genotype-constitutive and not involved in genetic variation of the plasticity of response to drought stress.

In our study, we used the two models and detected several types of QTL. The first class is represented by one QTL detected equally with the G and the GE model and that showed genotype-constitutive and plastic responses to drought stress. The second class regrouped 3 QTL that were detected only with the G model and showed genotype-constitutive and plastic responses to drought stress as their effects in the two models had significant differences. In those cases, the genetically-

variable part of plasticity for drought stress response (the plastic response) introduced some noise that did not allow the associations detection by the GE model. Therefore, the G model was also able to reveal regulations that were not detectable with the GE model only. So, even if the two models detected the same associations and gave similar SNP effect for the majority of the genes, the comparison revealed some associations with small genotypic effects that were concealed in the GE model due to the loss of detection power. The third class regrouped other QTL detected either by G, GE models or both but they did not show a significant difference in SNP effects between the two models. The non-detection of a genetic variation for plasticity in response to drought stress can be explained by different likely causes. First, we can suppose that for some genes the experimental conditions did not represent a large enough range of drought stress. Then, the selected genes were not regulated by G x water status interactions as all the genes are not equally likely to exhibit GxE interactions due for example to their promoter architecture or their regulatory complexity (Grishkevich & Yanai, 2013).

A last class of possible QTL was not detected in our study. It groups QTL detected only in the GE model and with significant differences between the effects of the two models. If such QTL were found, it would have indicated that for those genes the genotype-constitutive response should be not significant. One possible cause would have been the too small genotypic variability explored by the association panel. Therefore, only the genetically-variable part of the plasticity of the drought stress responses would have been revealed by the GE model. The absence of this last class of QTL is also consistent with the results of Lacaze *et al.,* (2009). In this study about genetics of phenotypic plasticity for barley, Lacaze and co-workers compared the localization of traits QTL (yield and its components) showing QTLxE interaction effect to the localization of plasticity QTL i.e QTL for slope and variance of reaction norm for the same traits. All plasticity QTL were co-localized with trait QTL. Therefore, as in our results, there were no QTL that only affected plasticity.

Biomarker utilization advantages and application on drought tolerance selection

Thanks to the estimation of a Water Status Biomarker, genotypes of the association panel were shown to perceived different water status even if they were placed in the same “starting” environment. Therefore, the gene expression variation across genotypes could be used to explore the responses to different water status variation and therefore to identify the probable GxE interactions. This approximation has its the real advantage in drastically decreased experimental costs in comparison to Multi Environment Trials that are usually set up to identify these interactions (Grishkevich & Yanai, 2013).

Moreover, Water Status Biomarker utilization in the G model and comparison with the GE model allowed the distinction between the genotype-constitutive and the plastic parts of the

regulation of the drought responsive genes. Knowledge on the importance of these two components of the regulation could be useful in several goals. Considering one locus independently of the others, a QTL with only a genotype-constitutive regulation of drought responses will be useful in the construction of ideotypes adapted to different water status environments. In addition, the knowledge of the relative importance of the G and GxE effects of a QTL for drought stress responses could be exploited in order to breed genotypes efficient for this particular water status. Moreover, considering several loci, the knowledge of the importance of their different genotype-constitutive and plastic effects could be even more important. In this context, we could combine several loci with QTL presenting strong plastic effects and breed ideotypes adapted to non-predictable environments.

Ratio of local and distant regulations in the drought regulatory network

Knowing localization of the studied genes for transcripts level and localization of the associated markers permitted to group QTL in two classes: local (cis) and distant (trans) regulations. Results in the present study concerning the ratio of cis/trans-regulations and their effects on transcript level variations are consistent with findings reported in studies based on linkage mapping for RILs in other plants such as *Arabidopsis* (Cubillos *et al.*, 2012a), maize (Swanson-Wagner *et al.*, 2009) and rice (Wang *et al.*, 2010). In the maize and rice studies, approximately 70% of the expressions QTLs were distant and explained a small fraction of the variation of each transcript (Cubillos *et al.*, 2012a). This is comparable to our results: among the identified cis/trans-regulations, 85% were distant but showed smaller genotypic effects. Concerning the GWAS, in humans, Dixon and co-workers (2007) found similarly numerous trans-regulations but with weaker effects than those in cis. On the contrary, in *Arabidopsis* on a study with 18 accessions, (Gan *et al.*, 2011) found more local associations than distant. However, this uniquely small ratio of trans- over cis-regulations could be due to the weaker effect of the trans-regulations and very small size of the studied panel that did not provide enough statistical power to detect trans-regulation effects (Gan *et al.*, 2011).

In our study, identical cis-regulations were found in both, G and GE models, and comparison of the effects of the associated SNPs between the two models showed that there was no significant difference. This suggests that local genetic variation affects gene expression in a consistent manner over a large range of environments. This finding is consistent with the results obtained in yeast (Smith & Kruglyak, 2008) as they showed that variation in local-regulatory sites induced change in transcripts levels that are less condition dependent than those induced by trans-acting factors.

On the contrary, the three genes previously demonstrated as to be involved in a GxE interaction were found to be controlled by distant QTL. This type of results was also found in *C.elegans* (Li, Y *et al.*, 2006). Therefore we can suppose that in sunflower as well, trans-regulations

are more likely involved in the genetically-variable part of the plasticity of response to drought stress. Our experimental conditions and results were not able to highlight a GxE interaction in the other distant QTL. However, we could suppose that with more stringent drought stress conditions, more genetic variability, and more repetitions, several other trans-regulations with GxE effects could be revealed. Indeed, several studies (Grishkevich & Yanai, 2013; Brem *et al.*, 2002; Des Marais *et al.*, 2013) demonstrated accumulating evidence that GxE likely accounts for the greater part of the phenotypic variation, and therefore gene expression, seen across genotypes.

III.4.6. Materials and Methods

III.4.6.1. Plant material

A core collection of 384 sunflower inbred lines has been built by a nested core collection strategy from an initial set of 752 inbred lines (Cadic *et al.*, 2013). It includes 176 public lines, whereas the others are private lines of the breeding companies: Soltis, R2N and Syngenta Seeds. Association panel used for the present association study contains 275 inbred lines and is a subset of the core collection described above.

Testcross progeny were obtained by crossing association panel lines with two testers according to their status (maintainers of cytoplasmic male sterility "B-Lines" or fertility restorers "R-Lines"), as described in Cadic *et al.* (2013). The R-Lines were crossed with the tester FS71501 and the B-Lines with the tester 83HR4gms.

III.4.6.2. Tissue harvest and RNA extraction

In the field, each sub-block of the first repetition was harvested on the 12th July 2011 when the plants were at post-flowering stage. For each genotype, the fourth leaf from the head was harvested on four plants and pooled. Samples of different genotypes were treated separately.

The leaves were cut without their petiole and immediately frozen in liquid nitrogen. Grinding was performed using a ZM200 grinder (Retsch, Haan, Germany) with a 0.5-mm sieve. Total RNA extraction of samples was performed using Qiazol (Qiagen, Hilden, Germany) and following the manufacturer's instructions. The quantity of RNA was estimated using a ND-1000 spectrophotometer (Nanodrop, Wilmington, DE, USA). cDNA synthesis was performed from 1g of total RNA using Invitrogen Super Script VILO cDNA synthesis Kit with random hexamer N6.

III.4.6.3. Gene expression quantification by qRT-PCR

Primers for qRT-PCR were designed using the sunflower reference transcriptome HaT13l (https://www.heliagene.org/HaT13) and Primer3 web tool (http://probes.pw.usda.gov/batchprimer3/index.html) using the default parameters with an optimal

product size of 60bp (min=50 bp, max=80 bp). We checked the target sequences of the primers according to the best BLAST hits in the sunflower transcriptome HaT13l.

Gene expression was estimated by qRT-PCR using the BioMark system (Fluidigm Corporation, San Francisco, CA, USA) with a 96.96 Dynamic Array IFC and EvaGreen (Bio-Rad, Hercules, CA, USA) as the DNA binding dye (Spurgeon *et al.*, 2008). The expression level of gene *i* expressed as the threshold cycle (Ct) was normalized according to the amplification efficiency (noted eff_i below) and the expression levels of seven reference genes (noted *r* below) identified in (Rengel *et al.*, 2012) and was estimated as follows:

$$dCt_i = \frac{(1+eff_i)^{Ct_i}}{\frac{\sum_{r=1}^{N_r}(1+eff_r)^{Ct_r}}{N_r}}$$

with N_r the number of reference genes.

III.4.6.4. Two models to analyze gene expression data

We proposed to analyze gene expressions with two different mixed models using the function lmer in the R package lme4.

The first model is called GE model:

$$E_{ij}^{GE} = \mu + G_i + b_j + \varepsilon_{ij}$$

where E_{ij}^{GE} is the phenotypic observation for the i^{th} genotype in the j^{th} block, μ is the intercept term, G_i is the genetic effect of the i^{th} genotype and is considered to be a random effect, b_j is the effect of the j^{th} block and is considered to be a fixed effect, and ε_{ij} is the residual error.

The second model is called G model and introduced a correction to take into account the water status of the plant using $WSB_{\psi PD}$ as a covariable in the model:

$$E_{ij}^{G} = \mu' + G'_i + b'_j + WSB_{ij} + \varepsilon'_{ij}$$

where E_{ij}^{G} is the phenotypic observation for the i^{th} genotype in the j^{th} block with the WSB value WSB_{ij}, μ' is the intercept term, G'_i is the genetic effect of the i^{th} genotype and is considered to be a random effect, b'_j is the effect of the j^{th} block and is considered to be a fixed effect, WSB_{ij} is the corresponding WSB value of the i^{th}genotype in the j^{th} block and ε'_{ij} is the residual error.

The GE model corrects only the spatial variation in the field and the genotypes are compared at different water statuses. The G model corrects both the spatial variation and the water status; therefore the genotypes were compared in similar water environment.

III.4.6.5. Genotyping of the association panel

An AXIOM chip (Affymetrix, Santa Clara, CA, USA), with a total of 197,914 single nucleotide polymorphism (SNP) markers was used to genotype the association panel. These SNPs were selected from either genomic re-sequencing or transcriptomic experiments. An additional set of 6800 non-polymorphic sequences were added as controls. Combined with internal technical controls, the AXIOM chip was designed with a total of 445,876 probesets. The 275 panel lines were genotyped with the AXIOM chip. All hybridization experiments were performed by Affymetrix and the genotypic data were obtained with the GTC software (Affymetrix). In total, 62,820 SNPs that showed polymorphism for the association panel with MAF > 5% and no redundancy between them were used as genotyping data for the association study.

III.4.6.6. Association analyses

Association between SNPs and traits was performed using Emma R package (Kang *et al.*, 2008). According to the study of Cadic *et al.* (2013), we used the mixed model that corrects for structure and kinship between the lines of the panel association:

$$G_i^{BLUP} = \sum X_{ic} a_c + M_{il}\theta_l + u_i + \varepsilon_i$$

where G_i^{BLUP} is BLUP for the i^{th} hybrid, X_{ic} is the tester category, a_c is the effect of the tester category *c*, M_{il} is genotype of the i^{th} hybrid at locus *l,* θ_l is the effect of locus *l*. a_c and θ_l are considered to be fixed effects, and ε_i is the residual. u_i is the random polygenic effect modeling kinship between panel lines with:

$$var(u) = \sigma^2 uK$$

where K is kinship matrix.

The kinship matrix K used in the association model (Cadic *et al.*, 2013) is estimated with Emma version 1.1.2 R package (Kang *et al.*, 2008) using the 62,820 SNPs set. It is an Identical By State (IBS) allele-sharing matrix. The population structure taken into account in the model is the structure due to the testers (FS71501 and 83HR4gms crossed with the R- and B-lines of the panel respectively) and is a binary covariate.

A False Discovery Rate (FDR) (Benjamini & Hochberg, 1995) was applied on p-values to correct for multiple testing using the *p.adjust* function in R. Associations with an adjusted FDR p-value <0.05 were considered to be robust.

III.4.6.7. G and GE models comparison

Comparison of the adjusted p-values for the two models

We first tested that G and GE models gave results with statistically significant differences considering the whole set of 62820 SNPs. We performed for each expression gene a paired t-test to compare FDR adjusted p-values. The t-test was performed using the function *ttest* in MATLAB (version 7.13.0.564, Statistics Toolbox 7.6. The Mathworks, Natick, MA, US)

Estimation of the SNP effects and comparison of G and GE effects

Using the Emma package (Kang *et al.*, 2008) in R and the association model described above, for each SNP in the G and GE models, we retrieved the predictors of the genotypic variance ($\hat{\sigma}_g^2$)and of the variance of the residual error ($\hat{\sigma}_\varepsilon^2$). We estimated the predictor of $\hat{\theta}_l$, matrix of the effect of the locus l, as follow:

$$\hat{\theta}_l = [M_l^t V^{-1} M_l]^{-1} [M_l^t V^{-1} G^{BLUP}]$$

where,

$$V = \hat{\sigma}_g^2 K + \hat{\sigma}_\varepsilon^2 Id$$

with K the Kinship matrix, Id the Identity matrix, M_l the genotyping data at locus *l*

Effects of a same SNP in the G and in the GE model were considered to be with significant differences if their confidence intervals (CI) at 95% did not overlap. The CI at 95% were calculated as follow:

$$CI = \hat{\theta}_l \pm 1.96\, \sigma_{\hat{\theta}_l}$$

where $\sigma_{\hat{\theta}_l}$ is the standard deviation of $\hat{\theta}_l$

III.4.6.8. Building genetic maps

To map the SNPs found in association with the gene expression traits, two genetic maps from two RILs populations were built with CarthaGène v1.3 (de Givry *et al.*, 2005). INEDI and FUPAZ2 populations, obtained from the cross between XRQ and PSC8 lines (180 samples) and from the cross between FU and PAZ2 lines (87 samples) respectively, were genotyped with the same AXIOM chip as for the association panel. From the 197,914 SNPs, 35,562 were polymorphic between XRQ and PSC8

and 28,529 between FU and PAZ2. To build the INEDI genetic map, we first added the genotypic data of markers from a consensus map, described in a previous study (Cadic *et al.*, 2013) to the set of AXIOM SNPs of the INEDI population to assign AXIOM markers to the appropriate LG. The INEDI genetic map consisted of 31,757 markers that were located on the 17 LGs for a total genetic distance of 1487.7 cM and grouped in 1861 different loci. We then built a FUPAZ2 genetic map using the AXIOM markers. We attributed all markers to the appropriate LG thanks to the previous genetic map from the INEDI population. The FUPAZ2 genetic map consisted of 17,901 markers that were located on the 17 LGs for a total genetic distance of 1425.3 cM and grouped on 807 different loci. We built a new consensus map to compare positions in FUPAZ2 and INEDI genetic maps. First, we selected the common SNPs polymorphic in both populations in order to obtain a consensus map by merging the two data sets in CarthaGène. This first-step produced a first consensus genetic map that was composed of 7076 markers in 1113 different loci located on the 17 LGs for a total genetic distance of 1471.1 cM. It was used as a skeleton on which we projected the INEDI and FUPAZ2 maps to produce the final consensus map. This latter map comprised 45,566 markers in 2711 different loci for a total genetic distance of 1,794.19 cM.

III.4.6.9. SNP mapping by Linkage Desequilibrium

Not all the SNP found in association with the gene expressions were mapped on the final consensus genetic map. LD was calculated between associated SNPs that were unmapped in one hand and all the 17,902 and 30,066 SNPs respectively mapped on FUxPAZ2 and INEDI genetic maps in the other hand. We used the statistics r^2vs and r^2v (Mangin *et al.*, 2012) that correct for biases caused by structure and kinship between individuals. For each unmapped marker, in each genetic map, we selected the ten mapped markers with maximum LD according to r^2vs statistic and the ten mapped markers with maximum LD according to r^2v. If the positions of these 20 markers were not more than 5 cM distant from the position of the marker with the maximum LD statistic (all methods considered), unmapped SNP was assigned to the same position as the mapped SNP that was in maximum LD.

III.4.6.10. SNP mapping using marker context-sequence alignment

The context sequences of the associated SNPs (71 bp-long) were aligned on the genomic and transcriptomic sequences of the sunflower genotype XRQ available on the Heliagene web-portal (https://www.heliagene.org/HaT13l). Transcripts and genomic scaffolds corresponding to the best BLAST hits were retrieved for each SNP context-sequence. If the context-sequences of an unmapped SNP and of a mapped SNP had the same best BLAST hit, we placed the two SNP, mapped and unmapped, at the same locus on consensus map.

III.4.6.11. Genes mapping

Mapping of the likely drought responsive genes was necessary in order to characterize the associations found in local or distant regulations. Transcript sequences of the genotype XRQ for each gene were retrieved from sunflower transcriptomic database (https://www.heliagene.org/HaT13l). Using alignments of these sequences with PSC8, FU and PAZ2 transcriptomic databases and XRQ and PCS8 genomic sequences, we looked for polymorphisms between INEDI population parents in one hand and FUPAZ2 population parents in the other hand.

Primers for Kaspar markers were designed on the HaT13l transcript sequence with Primer3 web tool (http://probes.pw.usda.gov/batchprimer3/index.html) using the parameters for allele specific primers and allele flanking primers with an optimal product size of 60bp (min=50 bp, max=80 bp). Genotyping using Kaspar technology (KBioscience UK Ltd., Hoddeston, UK) of 86 RILs for the INEDI population and 44RILs for FUPAZ2 population was performed. We mapped Kaspar markers on INEDI or FUPAZ2 genetic maps using AXIOM genotyping data of the corresponding RILs.

III.4.6.12. QTL definition from the association results

On the three maps (INEDI, FUxPAZ2 and consensus), SNP associated to the same gene expression trait and less than 5 cM distant from the next associated SNP, were considered to form one single QTL. If associated SNPs were distant from more than 5cM on the consensus map but are part of the same QTL on INEDI or FUPAZ2 maps, they were considered to belong to the same QTL even on consensus map. Again, if two QTL, associated to different genes, were distant from less than 5cM those QTL were considered to be a single one.

End of the project of article: "Integration of the environment in gene regulatory networks: Identification of plastic regulations in the case of drought stress in sunflower via an association study on gene expression"

III.5 Conclusion and outlook concerning eQTL detection with gene correlated to drought responses

III.5.1 Genes grouped in regulatory pathways for two drought tolerance traits

Genes selected for the association study were described as having their expression correlated significantly to only one physiological trait of drought responses (Rengel *et al.*, 2012) using a sparse partial least square analysis. As several genes were actually found correlated to the same physiological trait, this suggests that they are involved in the same regulatory pathway and therefore share some genetic control. During our GWA study and the reconstruction of the underlying GRN, we characterized the genetic control of such genes. We observed that each group of genes correlated to Carbon Isotopic Discrimination (CID) in one hand and to the Evapotranspiration (ET) in another hand were associated to several common QTL. Therefore the hypothesis of a common regulatory pathway for the genes involved in ET and in CID respectively is consistent with the results of our GWA study on those gene expressions. On the contrary, no common QTL were found for genes correlated to Relative Water Content (RWC) and Osmotic Potential (OP). Thus, the hypothesis of a same regulatory pathway for those traits could not be confirmed by the present study.

This common regulation between CID and ET could be explained by their functional relationships. CID measures the ratio of incorporation of $^{13}C/^{12}C$ by the RUBISCO and varies according to stomatal closure. Then, discrimination against ^{13}C is proportional to plant water use efficiency (Farquhar *et al.*, 1989). Therefore, CID integrates the stress of the plant through its levels of regulation of transpiration over a long period of time (Araus *et al.*, 2003). On the opposite, ET reflects the transpiration of the plant at the specific time of harvest when gene expression levels were estimated as well. Our results characterize the genetic control of these temporally different measures of transpiration and allowed us to identify genetic variation that controls the stomatal closure threshold all over the plant life cycle. This simple genetic architecture and regulatory pathways for CID and ET make them maybe a more direct and easier target to breed for drought tolerance than OP and RWC.

III.5.2 Utilization of the Water Status Biomarker

During this GWA study we used WSB to estimate water status of each genotype of the association panel. As already mentioned, the WSB was built and validated for only four genotypes. Thus, it is very likely that the biomarker's genes are differentially expressed in a panel of genotypes with a larger genetic diversity. However, in this part of the project we took as a hypothesis that the WSB model was valid for all the genotypes of the association panel. Indeed, all these genotypes are modern sunflower cultivars, despite the introgression of wild alleles in some of them. Then, we have to keep in mind that the correction for water status introduced thanks to WSB estimation is not

completely accurate for all genotypes of the panel. However, the WSB likely allowed us to eliminate at least a part of the bias due to the different water status of each genotype, even if a part of the genotype x environment (GxE) effect is very likely still "captured" at the same time of the genotypic effect for the G model.

In this part of the project, the utilization and comparison of the two models (correcting or not for the plant water status) enabled us to evaluate the plastic and the constitutive parts of the drought gene expression. However, we have to keep in mind that the ratio between plastic and constitutive parts is dependent to the environmental conditions of the experiment and also to the genetic diversity present in the association panel. As the genotypes of the panel are chosen to represent a large variability of modern cultivars, results found in this study concerning genotypic and genotype x environment effect could probably be generalized for cultivated sunflowers.

Moreover, WSB utilization allowed comparing genotypes in the exact same environment regarding soil depth and composition, climate, and crop management. Traditionally, the evaluation of the QTL effects through various environments is performed thanks to multi-environment trials. This method is expensive in particular for the acquisition of the phenotyping data. Moreover, in addition to the water status, several other components of the environment can change from a trial to the other. Therefore, even if the correction with the WSB is not optimal, its cost and accuracy have to be compared with multi-environment trials. Another possibility would be to combine these two approaches. Indeed it would allow a better characterization of the multi-environment trials.

III.5.3 Association study with an association panel using hybrids: advantages and drawbacks

In this study, the lines of the association panel were crossed with testers and thus, the genotypes used for the association study were hybrids. Utilization of hybrids instead of inbred lines has both advantages and drawbacks.

Sunflower lines are more susceptible to diseases or other environmental stresses. In the panel, those stresses could be very important due to the presence of lines with wild introgression into modern cultivars more sensitive to diseases. Moreover, since the discovery of the cytoplasmic male sterility (CMS) (Leclercq, 1969) and of the fertility restoration genes (Kinman, 1970), sunflower breeding is based on hybrids. Therefore, the hybrids utilization in this study is more realistic in a context of a breeding program.

However, as genotypes were hybrids, we could not know if the allelic effect was due to the line or to the tester, even if the structure of the panel introduced in the GWAS took into account this distinction between male and female testers.. Indeed, the loci associated to the phenotype are

heterozygous with one allele from the tester and one allele from the line. Moreover, as female (B) and male (R) lines are comprised in the panel, two different testers were used (one for the B-line and one for the R-line). Therefore, testing this panel with other testers could help to determine which genotype brings which allelic effect and to test the stability of the QTL.

III.5.4 Expanding the study to the whole sunflower transcriptome

The genome-wide association mapping described here studies the genetic control of 86 expression genes with finally 33 genes involved in the reconstructed GRN. To obtain a more complete view of the GRN for drought responses it would be interesting to take into account the whole transcriptomes of the sunflower genotypes. A similar approach has been conducted by Gan *et al.* (2011) on 19 parents of the Multiparent Advanced Generation Inter-Cross (MAGIC) population in *Arabidopsis*. The sequencing, assembly and annotation of the genomes of these 19 lines were part of the *A.thaliana* 1001 Genomes Project (Weigel & Mott, 2009). Similar information for the sunflower association panel could be useful in order to map associated SNP in a more accurate way and to make hypotheses about the function of the candidate genes under those QTL.

In our study, only three genes appeared to have a significant plastic part in their response to drought. Expanding the study to the whole sunflower transcriptome would certainly lead to the identification of genes with more important GxE effect and complete what we have found in our GWAS.

III.6 Discussion about drought responsive genes correlated to traits of drought stress tolerance

III.6.1 Attempt in the distinction between the genetically-variable part of plasticity and the genotype-constitutive response to drought

The utilization of both models, with and without correction for the water status, allowed the distinction and the quantification of the plastic (GxE effect) and the constitutive (G effect) parts in the effect of some eQTL detected. As already discussed, the GxE part is likely to exist for a greater number of genes than found in our study (three eQTL). It is certainly due to the limited number of genes studied in this work (other genes not selected for this study may be controlled by a GxE interaction effect) but also to the relatively limited range of drought conditions in our experiment and the limited number of repetitions. Distinction between these two parts of the genetic control of genes involved in drought stress responses can be useful and help in the choice of breeding strategy. Considering genes and their effects independently, genes with an important known plastic part in the genetic control of their regulation should be favored in a strategy where genotypes are bred for a specific environment (if the GxE effect is advantageous in this environment). On the contrary, if the

breeding strategy is to identify genotypes adapted to a large range of environment, genes with an important constitutive part and a small plastic part in the genetic control of their regulation will be researched. Another strategy to build an ideotype, adapted to an environment with variable drought stress would be the combination of various genes with GxE effect in order to adapt to different stress intensity.

Figure III.12 illustrates the different parts of genetic control of genes involved in drought stress responses.

III.6.2 Genotypic control of the plants micro-environment

Water status of the plant estimated via the WSB has also been used in the genome-wide association study as a classical phenotypic trait. One QTL was found associated to this trait. If we assume that this QTL is not the sign of the fact that the WSB is not completely genotype-independent in our association panel, then, we can consider that it highlights regions in the sunflower genome that are associated to the water status (estimated through the WSB). This implies that a feedback loop exists and allows the adjustment of plant water status in function to the drought responses developed by the plant. This feedback loop is represented in the Figure III.12. It can be interpreted as the plant's control of its hydric micro-environment. There is a permanent dialogue between the plant and its micro-environment in order to adjust the response to water deficit. We can hypothesize that genes underlying the QTL associated to the water status might likely be involved in mechanisms that harness soil water and in the regulation of the water losses.

This result gives us some details about the relationships between genes involved in drought responses and other genes involved in the regulatory cascade of water deficit responses such as genes involved in the environmental signal perception (Figure III.12).

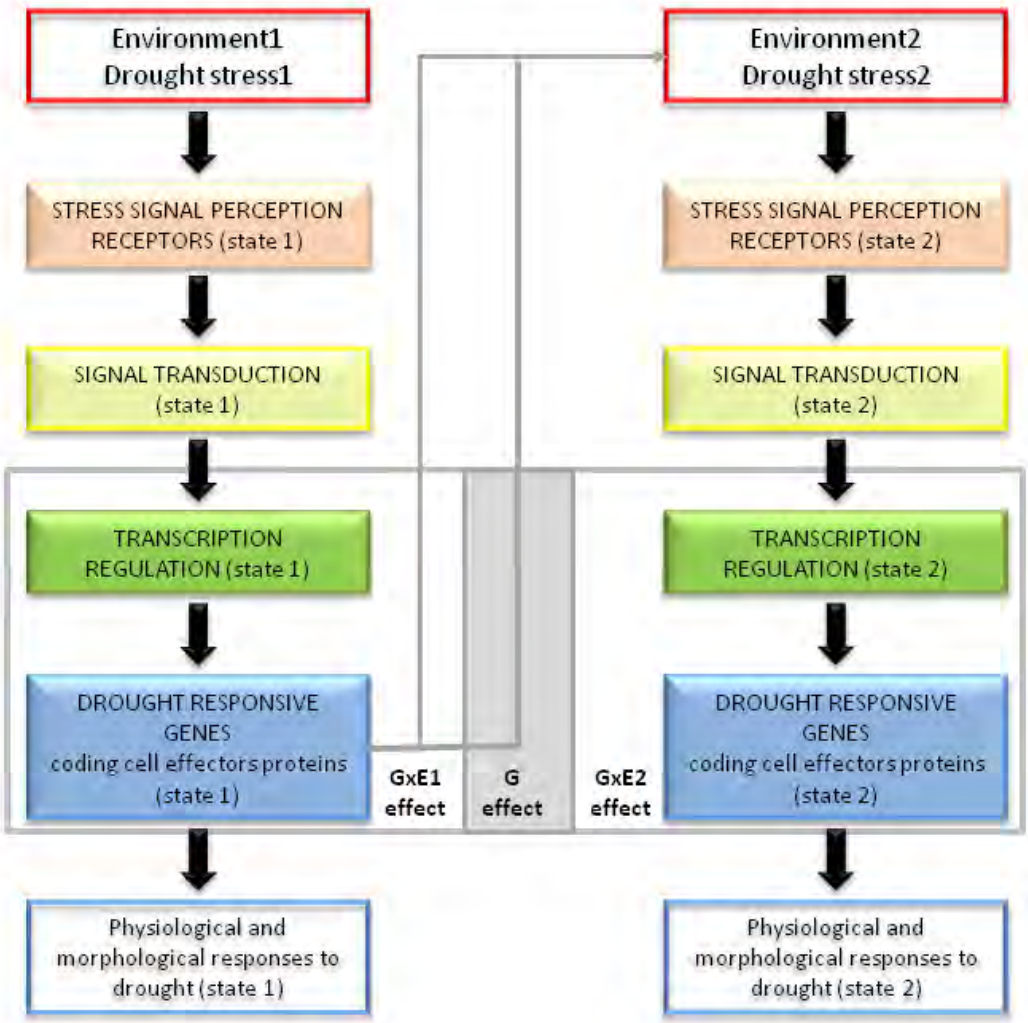


Figure III.12: Results about genes correlated to morphological and physiological traits for drought tolerance.

Identification of the plastic and constitutive parts in the genetic control of drought responsive genes and hypothesis of a feedback loop implying a control of its hydric micro-environment by the plant.

Chapter IV: Drought Gene Regulatory Network and implication in the evolution of *Helianthus annuus* and its relatives, a systems biology approach.

In this fourth chapter, we propose to focus on regulatory genes involved in transcription control and genes involved in functional responses to drought in order to understand how they interact between them (Figure IV.1). We propose to study those genes through a system biology approach in order to obtain a global view of the relationships between genes involved in the drought gene regulatory network and relate it to the genetic variability in *Helianthus* genus.

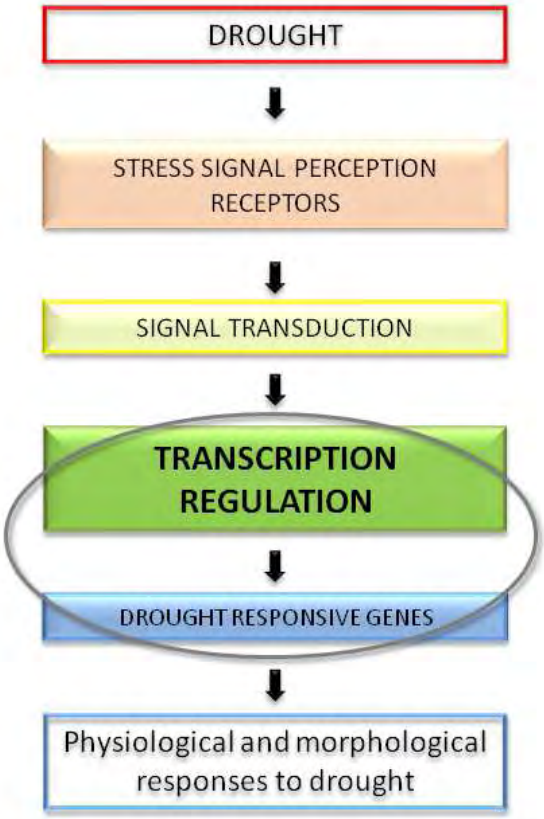


Figure IV.1: Genes studied are involved in transcription regulation and functional responses in the generic cascade induced by water deficit.

IV.1 Brief overview of systems biology approach

The goal of biology has evolved during the past centuries. At its beginning, its main purposes were the description and classification of living objects. These objectives have then evolved and nowadays the final goal appears as the understanding of every details and principles that govern the functioning of biological systems (such as plants or animals). Since the emergence of the field of molecular biology, substantial progress has been made that enables us to identify essential parts of the biological systems: genes and their products. The next major challenge is to understand at the system level the rules that govern the molecular components (genes, proteins, etc …) that have been revealed and studied individually by the molecular biology (Kitano, 2000). A system is a concept that basically refers to an assembly of components in a particular pattern. The term *systems biology* can be used in different contexts. It can mean a dynamic modeling or it can be used to refer to a multidimensional data analysis (Yuan *et al.*, 2008). An exhaustive definition could be stated with respect to its main objectives (Kitano, 2002) which consist in the four following points:

- *Identification of system structures*: Genes, proteins, metabolic pathways but also physical structures of organisms, cells, organelles or chromatin can be involved in any system description. Regulatory relationships that connect those components have to be identified as well. Identification of gene regulatory network for multi-cellular organisms is more complex as the cell-cell communication has to be taken into account. Yeast and *C.elegans* are examples of organisms in which considerable efforts have been made to obtain spatiotemporal data for gene expression and protein level (The *C.elegans* Sequencing Consortium, 1998; Ito *et al.*, 2000). Gathering of similar data is in progress for other biological systems such as *Arabidopsis thaliana*. Even if the first efforts are, up-to-now, limited to understanding components of the system and the local relationships between them, the results of these researches would be a first step for systems biology.
- *Analysis of system behavior:* In addition to the system structure, the dynamics of the system has to be understood. This could help, for example, to find an answer to the following questions. How does the system behave over time under various external stimuli or perturbations? How quickly does the system go back to its initial state? Moreover, having knowledge about how a system respond to stimuli can help in its definition. For example, understanding the leaves expansion rate under drought conditions can give some insight about the minimal and maximal possible sizes of a sunflower leaf.

- *Control of the system:* This objective is an application of the insights provided by the identification of the system's structures and its dynamics. This can provide the means to control the state of a system.
- *System design*: This is the ultimate goal of systems biology that reaches engineering field. Once the knowledge of the biological system functioning will be assimilated, the next step would be the prediction and the design of biological systems. Examples of such knowledge are the construction of synthetic organisms (announced in a close future) or the construction of "ideotypes" for breeding programs.

To achieve the four goals of a systems biology study, important efforts are required to obtain comprehensive, quantitatively accurate and systematic data sets. Such data sets are now possible to produce thanks to the important progress in the sequencing and transcriptomic technologies as for example the utilization of micro-fluidic systems or nano-technologies. Those technologies are the most exhaustive and affordable to date and facilitate the study of gene regulation.

Inference of gene regulatory network (GRN) is one of the aspects of systems biology. To identify GRN components two approaches can be used. The bottom-up approach tries to construct a gene regulatory network based on the compilation of independent experimental data, mostly through literature. Extensive databases are now available for gene expression and protein in various conditions, in particular for model species such as *Arabidopsis* (Zimmermann *et al.*, 2008). The top-down approach uses high-throughput data from expression arrays design for the network inference. Hybrid methods combining the bottom-up and the top-down approaches have also been experimented (Kitano, 2002). As the large datasets can provide information about various genes and network components, in many cases, it is interesting to begin with a focus on small networks (Middleton *et al.*, 2012)in order to make their understanding and utilization in future studies or research works easier.

The work presented in this chapter aims to reconstruct a GRN involved in drought tolerance. Gene expression data were retrieved from experiments designed specifically on sunflower and slightly completed using model species information. Therefore our gene selection combined a bottom-up and a top-down strategy.

IV.2 Main goals in the study of gene regulatory network

Inference of the network that connects genes differentially expressed during drought stress should highlight the main regulatory pathways in which those genes are involved. Indeed, the

reconstruction of the network through a systems biology approach provides us insights into the topology of this network and the evolutionary history it resulted from.

Therefore, a first aim of this work is the identification of key genes, such as hubs, that could explain robustness of the drought GRN across various drought stress scenarios and adaptation of the biological systems (here the plant) at an individual time-scale (that we will also call the physiological time-scale). This would reveal the relationships between regulatory genes and other genes involved in the generic cascade for drought responses.

Another interesting question is how patterns of the regulatory network have been conserved through evolution. The particular topology of a network leads to different constraints on the genes that form this network. For example, we can hypothesize that genes which are highly connected do not have to cope with the same evolutionary forces as peripheral genes. Therefore we can use network topology as a way to investigate selection pressure that shapes the evolution of the sunflower and the *Helianthus* genus in dry environments. Due to its history of domestication and the range of various habitats that sunflower and its relatives occupy, *Helianthus annuus* appears to be a good model to investigate this question.

IV.3 Article: Bridging physiological and evolutionary time scales in a gene regulatory network

Article status: **Accepted on New Phytologist, March 2014**

Authors:

Gwenaëlle Marchand[1,2]; Vân Anh Huynh-Thu[3]; Nolan C. Kane[4]; Sandrine Arribat[5]; Didier Varès[1,2]; David Rengel[1,2]; Sandrine Balzergue[5]; Loren H. Rieseberg[6,7]; Patrick Vincourt[1,2]; Pierre Geurts[3]; Matthieu Vignes[8]; Nicolas B. Langlade[1,2]

Corresponding author e-mail: nicolas.langlade@toulouse.inra.fr

Addresses:

1: INRA, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441

2: CNRS, Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, F-31326 Castanet-Tolosan, France

3: Department of Electrical Engineering and Computer Science and GIGA-R, Systems and Modeling, University of Liège, Liège, Belgium

4: Department of Ecology and Evolutionary Biology, University of Colorado at Boulder, Boulder, CO, USA

5: INRA, Unité de Recherche en Génomique Végétale (URGV), UMR1165 – Université d'Evry Val d'Essonne – ERL CNRS 8196, CP 5708, F-91057 Evry Cedex, France
6: Department of Botany, University of British Columbia, Vancouver, British Colombia, Canada
7: Department of Biology, Indiana University, Bloomington, IN, USA
8: INRA, Mathématiques et Informatique Appliquées (MIA), UPR875, F-31326 Castanet-Tolosan, France

IV.3.1. Summary

Gene regulatory networks (GRN) govern phenotypic adaptations and reflect the trade-offs between physiological responses and evolutionary adaptation that act at different time scales. To identify patterns of molecular function and genetic diversity in GRNs, we studied the drought response of the common sunflower, *Helianthus annuus*, and how the underlying GRN is related to its evolution.

We examined the responses of 32,423 expressed sequences to drought and to abscisic acid and selected 145 co-expressed transcripts. We characterized their regulatory relationships in nine kinetic studies based on different hormones. From this, we inferred a GRN by meta-analyses of a Gaussian graphical model and a random forest algorithm and studied the genetic differentiation among populations (F_{ST}) at nodes.

We identified two main hubs in the network that transport nitrate in guard cells. This suggests that nitrate transport is a critical aspect of sunflower physiological response to drought. We observed that differentiation of the network genes in elite sunflower cultivars is correlated with their position and connectivity.

This systems biology approach combined molecular data at different time scales and identified important physiological processes. At the evolutionary level, we propose that network topology could influence responses to human selection and possibly adaptation to dry environments.

IV.3.2. Keywords

ABA, abscisic acid; CLC-A chloride channel protein; drought; F_{ST}; genetic differentiation; network inference; NRT1.1; nitrate transporter 1

IV.3.3. Introduction

Phenotype is shaped during an organism's life by its physiological and developmental responses to environmental conditions and across generations through evolutionary genetic adjustments to new environments. On the time scale of individual organisms, the phenotype can change rapidly due to gene regulatory networks (GRNs), which translate environmental and internal

signals into physiological and developmental modifications. On an evolutionary time scale, such phenotypic modifications are based on changes in the genes composing the network that may alter this network at the structural or functional level.

Relating phenotypic modifications occurring at physiological and evolutionary time scales has been a major focus of evolutionary biologists for more than a century (Osborn, (1896) and Waddington, (1942) as well as more recently Queitsch *et al.*, (2002); Milo *et al.*, (2007)). Researchers have theorized (and later demonstrated) that physiological adaptation (for example, via regulation of gene expression or biochemical characteristics) can be replaced by an evolutionary change that becomes constitutive and alleviates the fitness costs associated with plasticity. This paradigm can be revisited in the context of a gene network. While gene regulatory networks are products of evolution, similar to other biological objects, GRNs also shape and constrain the evolvability of phenotypic responses to the environment.

Systems biology approaches, such as GRN inference, provide a global view of the different pathways that respond to environmental variation. A GRN is a genetic network based on gene expression levels (Wilkins, 2005). It describes transcriptional interactions and dynamics in response to environmental stressors, and therefore the GRN is key to understanding how organisms such as plants adapt to their environment.

Responses to environmental signals are often mediated through hormones. For example, in plants, abscisic acid (ABA) is produced during water stress in the vasculature and in the guard cells of the vegetative part of the plant (Boursiac *et al.*, 2013). Accordingly, the application of ABA induces the expression of genes involved in the response to dehydration and mimics drought stress. This interpretation has been confirmed by promoter analyses, which have demonstrated that these pathways share many targets (Shinozaki & Yamaguchi-Shinozaki, 1997). The signals of different hormones interact and are integrated to convey environmental signals through the plant *(*Wilkinson *et al., 2012),* suggesting that hormones should share transcriptomic targets.

Drought stress is a major abiotic factor that drives dramatic phenotypic changes in plants, including *Helianthus*, in which drought stress appears to constrain the colonization of new environments in the arid regions of the southwestern USA (Seiler & Rieseberg, 1997). Therefore, the drought-stress GRN represents a tool for studying the interactions between organismal acclimation on the physiological time scale and population adaptation on the evolutionary time scale.

Several hormones mediate drought-stress responses; thus, the utilization of multiple hormonal treatments can elucidate the underlying GRN and highlight possible relationships between the genes involved. However, there are practical difficulties associated with the study of genetic networks. For example, the GRN identified could be biased toward interactions that have been previously detected in model species (Wilkins, 2005). To date, systems biology approaches, such as GRN inference, have

been mostly restricted to model species, such as yeast (Dikicioglu *et al.*, 2011), *Drosophila (*Crombach *et al., 2012),* or *Arabidopsis* (Ma *et al.*, 2007), and are typically performed under laboratory conditions. However, modeling dynamic biological processes requires time-series gene expression data that are relevant to both the biological process of interest and to the species targeted by the study. To understand genome function and evolutionary processes in an organism such as the sunflower, it is important to infer the GRN for the gene sets that are actually involved in the responses to a given environmental stress and to avoid the pitfall of using non-adapted model species data.

In this study, we used inference methods on sunflower data complemented with knowledge from *Arabidopsis*. These methods were specifically designed for time-series gene expression data and allowed us to reconstruct a sunflower GRN. The inferred GRN provides us a global view of the main physiological functions involved in the drought-stress responses occurring in the leaf, as well as their chronology.

On the evolutionary time scale, studying the underlying GRN for responses to environmental stresses such as drought can help explain how plants evolved to become better suited to their environments. Knowledge of gene's position in the GRN and its topological characteristics provides useful information about likely evolutionary constraints. For example, a highly connected gene is likely to be subject to many trade-offs, which would limit the accumulation of genetic diversity. Here, we identify correlations between network topology and genetic divergence between elite lines and landraces of sunflower and propose a mechanism to explain how sunflower genetic differentiation could be constrained in response to selective forces.

IV.3.4. Material and methods

IV.3.4.1. Plant Material and growth conditions

Transcriptome interactions and dynamics were studied using the sunflower (*Helianthus annuus*) genotype XRQ. Plantlets were grown under hydroponic conditions in the previously described growth medium (Neumann *et al.*, 2000) in a growth chamber. After 14 days, the plantlets were treated by adding either mock solution (DMSO only in controls) or one of the following hormonal solutions : auxine (IAA); ethylene (ACC), gibberellic acid (GA3), salicylic acid (SA), methyl-jasmonate (MeJA), kinetin, ABA strigolactone (Stri) or Brassinol (Bras) Details about hormonal solutions are provided in Appendix IV.1. First pairs of leaves was harvested at 0 (just before treatment), 1, 3, 6, 9, 24, and 48 hours after treatment, immediately frozen in liquid nitrogen, and stored at -80°C. The whole procedure was repeated three times for ACC, Bras, GA3, IAA, kinetin, SA, and Stri and four times for ABA and MeJA.

IV.3.4.2. Gene selection

To identify genes which likely play a role in the drought GRN, a global transcriptomic approach was employed using an Affymetrix chip containing 32,423 probesets corresponding to sequences expressed in *Helianthus annuus (*Rengel *et al., 2012b)*. Three different global transcriptomic datasets were analyzed and used to select genes. We selected genes that responded to at least two of the following conditions: (1) drought stress under field conditions; (2) drought stress under greenhouse conditions; and (3) 10 µM ABA application under hydroponic conditions.

The microarray data and analyses of the field and greenhouse conditions were previously reported by (Rengel *et al.*, 2012b). Under field conditions, plants of the Melody genotype were harvested at the post-flowering stage at a stress intensity level of 0.63 and 0.22 (ratio between evapotranspiration and maximal evapotranspiration) for irrigated and non-irrigated plants, respectively. Under greenhouse conditions, we recorded data from Melody pre-flowering plants at a fraction of transpirable soil water (FTSW) of 0.83 and 0.03 for the irrigated and non-irrigated plants, respectively.

The global transcriptomic data for the application of 10 µM ABA are new results and were obtained using the 6-hour treatment with ABA in the hydroponic experiment on the genotype XRQ (CATdb: AFFY_ABA_Sunflower or GEO accession: GSE22519). RNA quality verification, cDNA synthesis, and chip hybridization and washing were all performed using the Affymetrix platform at the INRA-URGV in Evry, France, following the protocol described in (Rengel *et al.*, 2012). To identify the sunflower transcripts that were differentially regulated by ABA under our hydroponic conditions, the Affymetrix data were treated as previously described in (Bazin *et al.*, 2011).

This list was extended to 181 genes with genes known to respond to the application of ABA or other hormones (literature (Boudsocq & Lauriere, 2005; Kawaguchi *et al.*, 2004; Miller *et al.*, 2009; Seki *et al.*, 2007; Umezawa *et al.*, 2010; Shinozaki & Yamaguchi-Shinozaki, 2007; Wang *et al.*, 2003; Wasilewska *et al.*, 2008; Rook *et al., 2006)*; Sirichandra *et al.*, 2009; Pastori & Foyer, 2002; Hirayama & Shinozaki, 2010; Li, S *et al.*, 2006; Bray, 2004; Valliyodan & Nguyen, 2006) or GO analysis).

IV.3.4.3. Molecular analysis

The extraction of total RNA and cDNA synthesis were performed as described in (Rengel *et al.*, 2012). The expression levels of the 181 selected genes were analyzed in all samples by q-RT-PCR using the BioMark system (Fluidigm Corporation, San Francisco, CA, USA) as previously described (Spurgeon *et al.*, 2008). The q-RT-PCR results were analyzed following the 2^{ddCt} method (Livak & Schmittgen, 2001). Gene expression levels were normalized to the mean of previously validated reference genes (Rengel *et al.*, 2012) and to the corresponding control sample with the mock treatment. Detailed description of expression levels calculation is provided in the Appendix IV.1.

IV.3.4.4. Genetic differentiation among populations

Genetic polymorphisms of drought GRN genes were characterized in five different *Helianthus* populations, as described in a previous study (Renaut *et al.*, 2013): *H. argophyllus* (N=28), *H. petiolaris* (N=25), *H annuus* elite lines (N=9), *H. annuus* landrace lines (N=11), and wild *H. annuus* (N=11). Briefly, transcript sequences were obtained from young leave tissues with two RNAseq technologies (Roche 454 FLX and GAII Illumina pair-end sequencing 2x 100 bp). The transcript sequences were then aligned to the reference transcriptome using the Burros Wheeler Aligner (Li & Durbin, 2009). SNPs were called using the program SAMtools *(*Li *et al., 2009)* with a minimum with Phred scaled genotype likelihoods of 30, corresponding to a genotyping accuracy of at least 99.9%. The population genetics statistic F_{ST} was calculated between these populations for 89 of the 181 candidate genes using the R package HIERSTAT (Goudet, 2005). F_{ST} is a widely used measure of genetic differentiation among populations.

IV.3.4.5. GRN reconstruction

Missing values of gene expression (expressed as ΔΔCt) at time t=0 were imputed as values of 1. Other missing values (less than 1% of the values) were imputed with the R package IMPUTE by 10-nearest neighboring genes (Troyanskaya *et al.*, 2001).

After log transformation of the data, we performed an arithmetic mean over replicates to obtain a robust ΔΔCt expression value for each gene under each condition (time x treatment). We obtained nine datasets corresponding to the nine hormonal treatments and containing expression values for 145 genes with robust expression data at 7 different time points. From these nine datasets, we inferred 10 GRNs: one GRN from each hormonal treatment and a global GRN taking into account all treatments. Two complementary inference methods were used to achieve GRN predictions.

The first method represents an extension of GENIE3 (Huynh-Thu *et al.*, 2010) and was based on the random forest method (RF, (Breiman, 2001)). In summary, each gene expression at time t+1 was successively considered as a target, and the method sought regulators of that gene via their expression at time t. Several regulator inclusion steps were successively performed: according to a variance reduction criterion in a regression tree framework, each step resulted in the inclusion in the model of the best regulator. The process was repeated on a randomized ensemble of trees, which made up the so-called random forest. This method allowed us to derive a ranking of the importance of all regulator expressions for the target by averaging the scores over all the trees of the random forest. The randomized subset of regulators allowed us to avoid the local minima of the global score, and the random subsample of the data used for each tree avoided over-fitting of the data and hence permitted more robust estimates. We tested on simulated data whether including auto-loops in the

model improved the performance. Results are presented in Appendix IV.1 and they show that no gain was obtained with such modified version of our RF algorithm. Compared to previously developed tree ensemble methods, our method is novel because our modeling explicitly accounted for the dynamical and multi-condition aspects of the data.

The second method used a Gaussian graphical modeling (GGM) approach. In the GGM paradigm, an edge was inferred when a significant partial correlation was detected between the expressions profiles of two genes. Namely, the partial correlation between two genes is the correlation between the residuals of the expressions of these two genes after accounting for all other gene expressions patterns. A unique aspect of our approach is the combination of a temporal approach with a multiple graph structure inference scheme. The dynamic nature of the data allowed us to obtain directed edges between two genes (i.e., changes in the expression of gene *p* induced changes in the expression of gene *q* and not the converse). In addition, the multiple graph framework drove the inference of condition-specific networks. However, each of these hormonal networks took into account information from the others and therefore accounted for a coupled functioning of the biological mechanisms that they encoded. The details of the RF and GGM approaches are provided in the Appendix IV.1. For each of the ten GRNs, we selected only edges confirmed by both methods. The union of the nine hormonal consensus networks and the global consensus network formed a final unified network with hormone-specific edges and global edges.

IV.3.4.6. Topological parameters

The topology of a GRN depicts the relative positions of the genes in the network and their importance in the structure of the network. The topological parameters for each node therefore represent quantitative measures of gene connectivity and network position; these parameters are calculated from the oriented edges that connect one gene with another. The edge count, the indegree and the outdegree are three correlated parameters indicating the total number of edges (in and out) and the number of outgoing and ingoing edges respectively. The average shortest path length of a node *p* is the average length of the shortest path between *p* and any other node. The closeness centrality is the reciprocal of the average shortest path length. The eccentricity is the maximum non-infinite length of the shortest path between *p* and another node in the network. As the network is directed, if *p* is a node without outgoing edges, the values of the average shortest path length, the closeness centrality, and the eccentricity could not be calculated. The betweeness centrality of a node *p* is the number of shortest path from a node *q* to a node *r* (differents from *p*) divided by the number of shortest paths from *q* to *r* that pass through *p*. It reflects the amount of control that the node *p* exerts over the interactions of other nodes in the network. The stress centrality of a node *p* is the number of shortest paths passing through *p*. Finally, the neighborhood

connectivity of a node *p* is the average connectivity of all neighbors of *p*. These different metrics were calculated for all genes with the NetworkAnalyzer plugin for Cytoscape (Assenov *et al.*, 2008).

IV.3.4.7. Correlation between topological parameters and genetic differentiation

First, we performed firstly a principal component analysis (PCA) on the topological parameters of the GRN to study the dependency of those parameters, with the function *princomp*. This allowed us to identify the components explaining the most parameters variability. From these PCA results, we selected the most representative topological parameters in order to avoid redundancy. The F_{ST} values were grouped into 5 subsets, each of them expressing the F_{ST} between one *Helianthus* population (Wild *H.annuus*, Landraces, Elite, *H.argophyllus*, *H.petiolaris*) and the other populations. We performed a canonical correlation analysis (R function *cancor*) in order to identify the canonical correlations between the selected topological parameters on one side and each F_{ST} subset on the other side. We tested their significance with the test of Wilks as provided by the function *p.perm* of the R package CCP with 10 000 permutations.

IV.3.5. Results

IV.3.5.1. Gene selection to infer the drought GRN

Gene identification using a global transcriptomic approach

To identify genes that play a role in the drought GRN, a global transcriptomic approach was employed using an Affymetrix chip containing 32,423 probesets, which corresponded to sequences expressed in *H. annuus*. The differential analysis identified 337 genes that responded to drought stress under field conditions and 447 genes that responded to drought stress under greenhouse conditions (Rengel *et al.*, 2012). Because ABA is the major plant hormone involved in the drought-stress response, we also identified genes displaying differential expression 6 hours after ABA treatment at the plantlet stage under hydroponic conditions, using a similar global transcriptomic analysis. A total of 463 sunflower transcripts were found to be differentially expressed after ABA application (Appendix IV.2). The 463 ABA-regulated sunflower genes were validated by comparison with the expression of 226 homologues in *Arabidopsis* based on expression data from the Bio-Array Resource database or in projects from the AtGenExpress Consortium retrieved on the website http://www.weigelworld.org/resources/microarray/AtGenExpress/AtGe_Abiostress_gcRMA.zip.

The authors employed a kinetic analysis of three time points to assess the transcriptomic response to abiotic stresses such as cold, osmotic, salt, drought or heat stress in leaves using the *Arabidopsis* Affymetrix ATH1 microarray. This study was of particular interest because its kinetic approach imparts greater statistical power and avoids the issue of differences in kinetic parameters

between sunflower and *Arabidopsis*. The *Arabidopsis* homologs of the sunflower genes in this study are all BLAST reciprocal best hits between *Helianthus* ESTs and *Arabidopsis*. The covariance analysis (ANCOVA) showed that the expression modulation over time of 27% of these *Arabidopsis* homologues (60 genes) exhibited a treatment effect or a treatment x time interaction effect when exposed to abiotic stresses. This proportion of *Arabidopsis* genes homologous to *Helianthus* genes responding to ABA corresponds to a significant enrichment in *Arabidopsis* genes responding to abiotic stresses (hypergeometric test giving $p=1.10^{-4}$). The ANCOVA analysis, hypergeometric test and results are described in detail in the Appendix IV.1 and Appendix IV.3, respectively. This finding confirms that at the transcriptomic level, ABA regulation and its role in abiotic stress responses are globally conserved between *Arabidopsis* and *H.annuus*, as it has been documented in many plants; this conservation has occurred even though sunflowers are a very distantly related lineage separated by more than 90 million years of evolution (Chinnusamy *et al.*, 2004).

These three lists contain gene groups that respond to two drought stress intensities and ABA application (mimicking a third drought stress condition) at different developmental stages. Together, they provide complementary views of the drought-regulated genes in sunflower.

For inclusion in the GRN for drought stress, we stipulated that the genes must respond to at least two of the following conditions: (1) drought stress under field conditions, (2) drought stress under controlled greenhouse conditions, and/or (3) ABA under hydroponic conditions (FigureIV.2). As expected from the large variability of the biological material used to select the genes, the selected intersection was robust and should comprise the genes composing the core GRN for drought stress.
In addition to these groups of genes, we selected 56 genes that are known from the literature or gene ontology (GO) analysis to be regulated in response to ABA or one of the other main plant hormones used for the treatment in our hydroponic experiment.

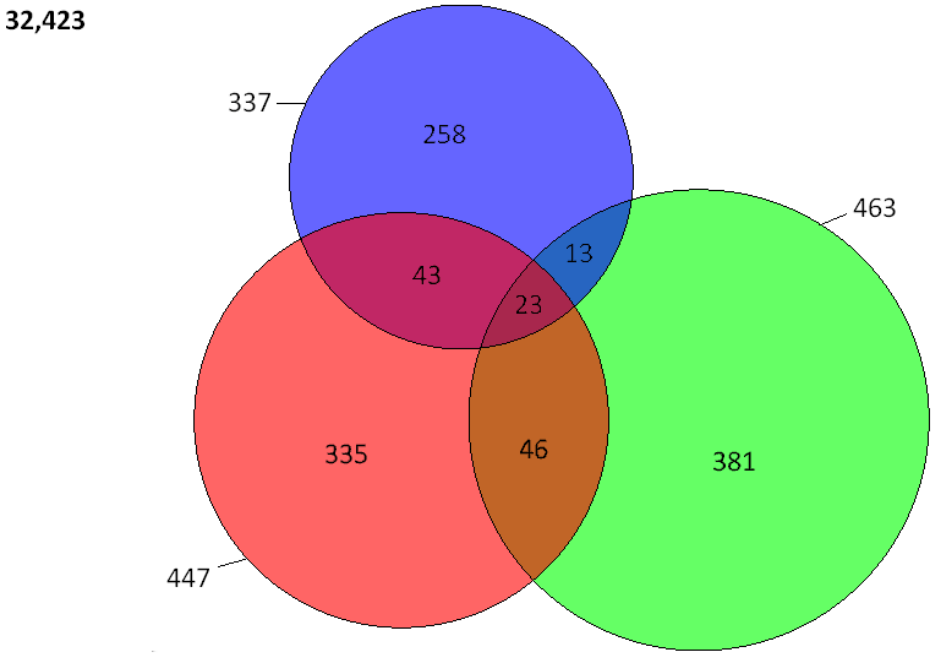


Figure IV.2: Selection of genes likely to be involved in the drought GRN.

Genes that responded to drought stress under field conditions, drought stress under greenhouse conditions, and ABA application under hydroponic conditions are indicated in blue, red, and green, respectively. The genes that were responsive under at least two of the different conditions were selected as part of the inferred GRN for drought-stress responses.

In all, 181 genes were selected (see complete list of sunflower transcripts, *Arabidopsis* homologs and annotations in Appendix IV.4).

Dissection of transcriptional regulation in the drought GRN by application of hormonal treatments

The GRN of these drought-regulated genes was reconstructed from their expression levels measured by q-RT-PCR. To perturb the network and identify regulatory relationships, leaf samples were harvested at seven different times after hormone treatment from hydroponically grown plants. A total of nine different hormones representing the main plant hormone groups were used. From the 181 selected candidate genes, we retained 145 robust genes based on technical filtering (efficiency, imputable missing data). The expression levels (expressed as ΔΔCt in reference to 5 control genes and the mock control) before and after imputation of missing data are shown in Appendix IV.5 and Appendix IV.6, respectively.

IV.3.5.2. Inference of the drought GRN from the GGM and RF methods

Inferences of a global GRN and nine hormonal GRNs lead to the identification of a robust unified drought GRN

To identify the final regulatory network between the 145 genes shown to be co-expressed during drought stress, we studied their regulation after several hormonal applications. This strategy was chosen because the environmental signal is transduced by different hormones whose regulatory pathways are very connected. The application of different hormones can reveal hormone-specific and global regulatory connections. Because we selected genes shown to respond to drought, the revealed regulatory connections are likely involved in drought-stress responses. We generated nine datasets corresponding to the nine hormonal treatments and containing expression values for the 145 robust genes at seven different time points. From these nine datasets, we established 10 GRNs: one GRN from each hormonal treatment and one global GRN, which represents a consensus array of all hormonal treatments. The GRNs were inferred using two different inference methods: Gaussian graphical modeling (GGM) and random forest (RF). These two approaches produce complementary predictions (Allouche *et al.*, 2013), and merging their results was shown to yield more reliable predictions than predictions obtained by any single method (Marbach *et al.*, 2012).

With the GGM method, we obtained between 112 and 158 edges for each hormonal network and a global network with 95 edges (Figure IV.3).

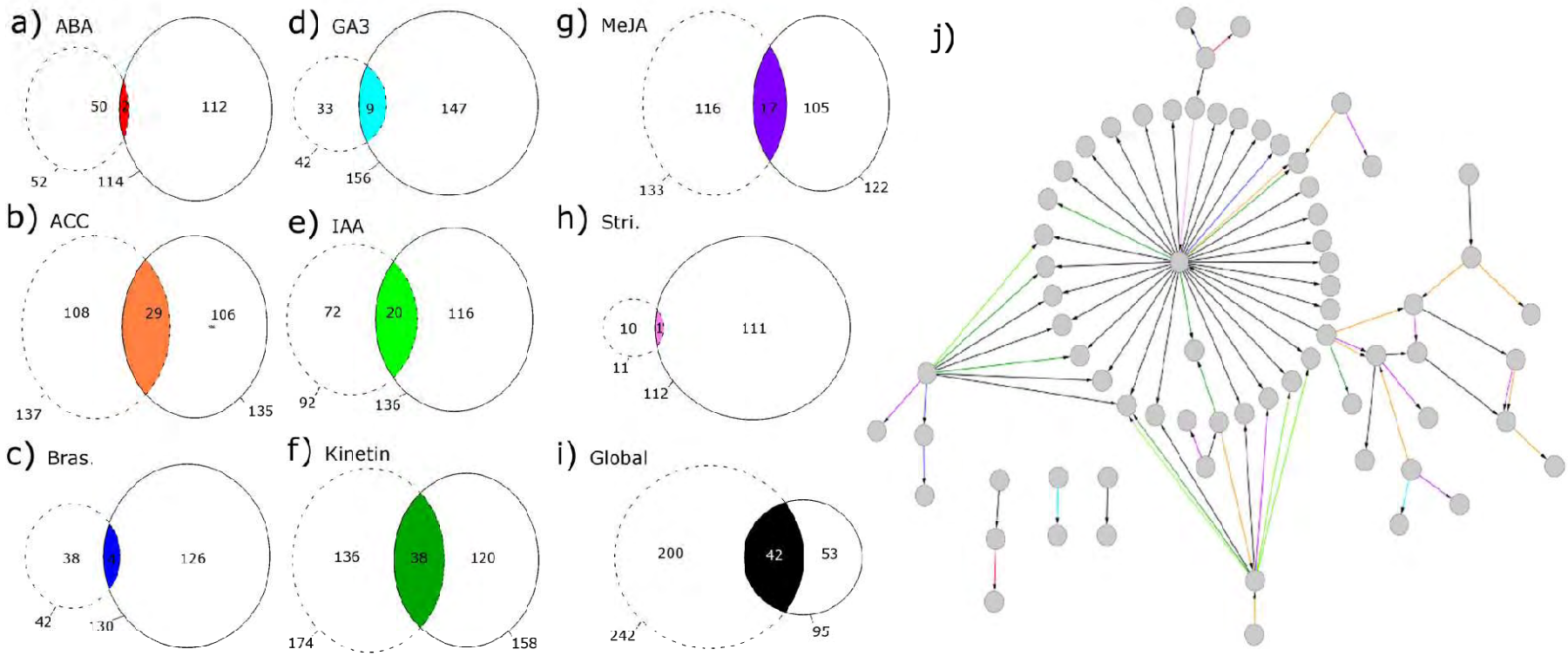


Figure IV.3: Drought GRN and selection of its edges.

a-i: The Venn diagrams for each hormonal GRN and global GRN represent the edges selected by the RF method (dotted line) and the GGM method (solid line). a) ABA. b) Ethylene. c) Brassinosteroid. d) Gibberellin. e) IAA. f) Kinetin. g) Methyl-jasmonate. h) Strigolactone. i) Global. j) Unified drought GRN representation. Grey circles represent the genes. Arrows represent the relationships between two genes (oriented edges), and their color represents the hormonal treatment that led to their identification: Red = ABA; Orange = Ethylene; Dark blue = Brassinosteroid; Light blue = Gibberellin; Light green = IAA; Dark green = Kinetin; Violet = Methyl-jasmonate; Pink = Strigolactone; and Black = Global or non-hormone-specific edges.

With the RF method, the number of edges for each hormonal network was very different and varied from 11 to 174 edges. The global GRN with the RF inference was composed of 242 edges (Figure IV.3).

Given the diversity in the inferred edges, we employed a very stringent approach to retain the core, most robust GRN. First, we discarded the results of SA treatment because the RF method inferred 629 edges. This number was far higher than that for the other hormones (49, 115, 38, 36, 94, 134, 147, and 16 when including SA). We chose not to take into account the SA edges in the final GRN to avoid an over-representation (more than 25%) of specific edges for this hormone instead of drought edges. Second, for each GRN (hormonal or global), we considered an edge to be robust if it was selected by both the GGM and RF methods. This is a conservative approach that leads to high-quality edges; we chose to focus on a network with very reliable edges at the expense of potentially missing some weaker associations that might be relevant. This trade-off was confirmed in very different scenarios based on both simulated and real data sets (Vignes *et al.*, 2011; Marbach *et al.*, 2012). We validated both our models using simulated data that had the specific features of the data being studied (see the Appendix IV.1). Note that the numbers of robust edges were very different depending on the focal GRN. The final unified network, hereafter called the drought GRN, was formed by the union of all these robust edges (Figure IV.3) and comprised 69 connected nodes, representing the genes linked by 79 unique edges. Among the 69 genes, 49 were differentially expressed in one of the three global transcriptomic experiments using the *Helianthus* Affymetrix chip, and only 20 came from the literature or GO analyses using BLAST reciprocal best hits to infer homology. Figure IV.4 summarizes the origins of the 69 final genes of the network.

The number of shared edges between the hormonal GRNs varied from 0 to 18 (Appendix IV.7 and Table IV.1). The ethylene, cytokinin, and auxin networks shared the largest number of edges, whereas the ABA, brassinosteroid, and strigolactone networks had no edges in common with the other hormonal networks.

	ABA	**ACC**	**Bras**	**GA3**	**IAA**	**Kine**	**MeJA**	**SA**	**Global**
ABA									
ACC	0								
Bras	0	0							
GA3	0	4	0						
IAA	0	9	0	5					
Kine	0	18	0	6	15				
MeJA	0	6	0	3	4	5			
SA	0	0	0	0	0	0	0		
Global	0	18	0	7	16	30	8	0	
Specific	2	8	4	2	3	6	7	1	6
Total	**2**	**29**	**4**	**9**	**20**	**38**	**17**	**1**	**42**

Table IV.1: Number of edges detected for each hormone

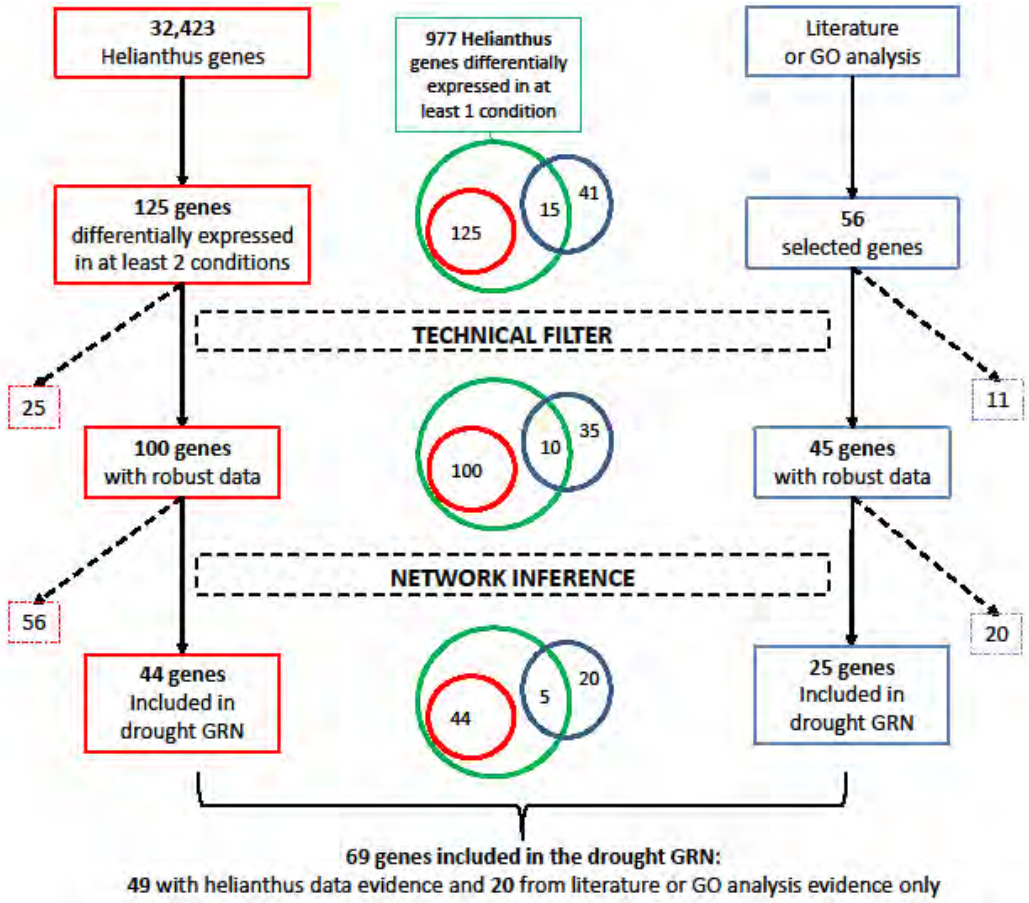


Figure IV.4: Origin of the selection for the inferred genes of the drought GRN.

Comparison of the drought GRN to *Arabidopsis* data and prior knowledge of biological networks

We compared our sunflower drought GRN to the model plant *Arabidopsis thaliana* using expression data from the AtGenExpress Consortium (Goda *et al.*, 2008) (GEO accession: GSE39384 from AtGenExpress Consortium). This *Arabidopsis* data set was similar to the *Helianthus* data and includes seven hormonal treatments but is limited to only three time points. Due to this difference in the sampling frequency, we were unable to define a network from these data using the inference methods described above. Therefore, we searched for gene expression correlations that were consistent (or inconsistent) with the sunflower data. Among the 116 *Arabidopsis* genes that were homologous to the 145 sunflower genes that were initially used to develop the consensus drought GRN, significant correlations between gene pairs were more frequent for pairs corresponding to the network edges, according to an exact hypergeometric test (p=0.005). The correlation analysis and hypergeometric test are described in the Appendix IV.1. This result demonstrated that the gene expression correlations identified from the *Arabidopsis* data were similar to the correlations identified in our sunflower drought GRN.

The topology of the drought GRN is consistent with what is known about biological networks. The degree distribution of the sunflower drought GRN followed a power law $y = 20.57x^{-1.98}$ with an R^2 of 0.72 (Figure IV.5). This means that a few nodes had many connections and that the majority of the nodes had few edges, a finding that is a typical feature of the scale-free topology of biological networks (Barabasi & Oltvai, 2004).

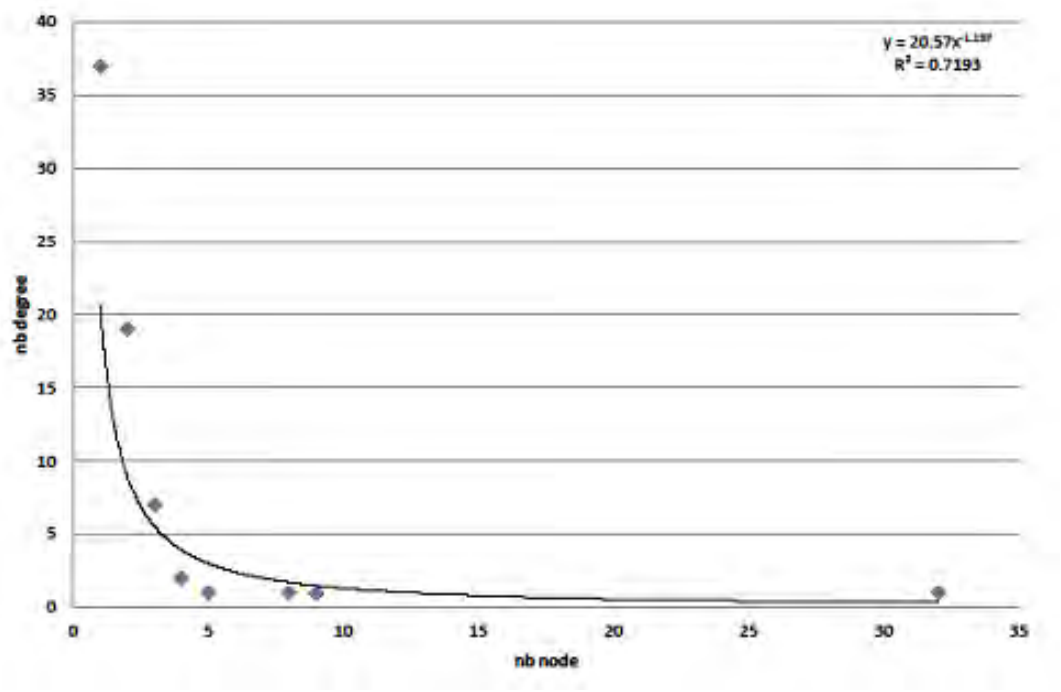


Figure IV.5:Degree distribution

IV.3.5.3.Node connectivity defines different gene classes

Identification of two hubs sharing common targets

The average value for the connectivity of a node (i.e., the number of outgoing or ingoing edges connecting a node to the others) in the inferred drought GRN was 2.3. However, we identified nodes with important connectivity; in particular, two nodes had the highest number of outgoing edges: 8 and 32 (with a connectivity of 9 and 32 respectively). These two genes were identified as important hubs in the inferred GRN. In addition, these genes shared 7 common targets, while no common sources (i.e., a gene *q* that targets the studied gene *p*) between these genes were identified.

Relation between connectivity and gene function

Gene ontology annotations of the *Arabidopsis* genes homologous to the 69 *Helianthus* genes connected in the unified drought GRN were retrieved from TAIR based on protein homology using the sunflower transcriptome web portal (www.heliagene.org/HaT13l).

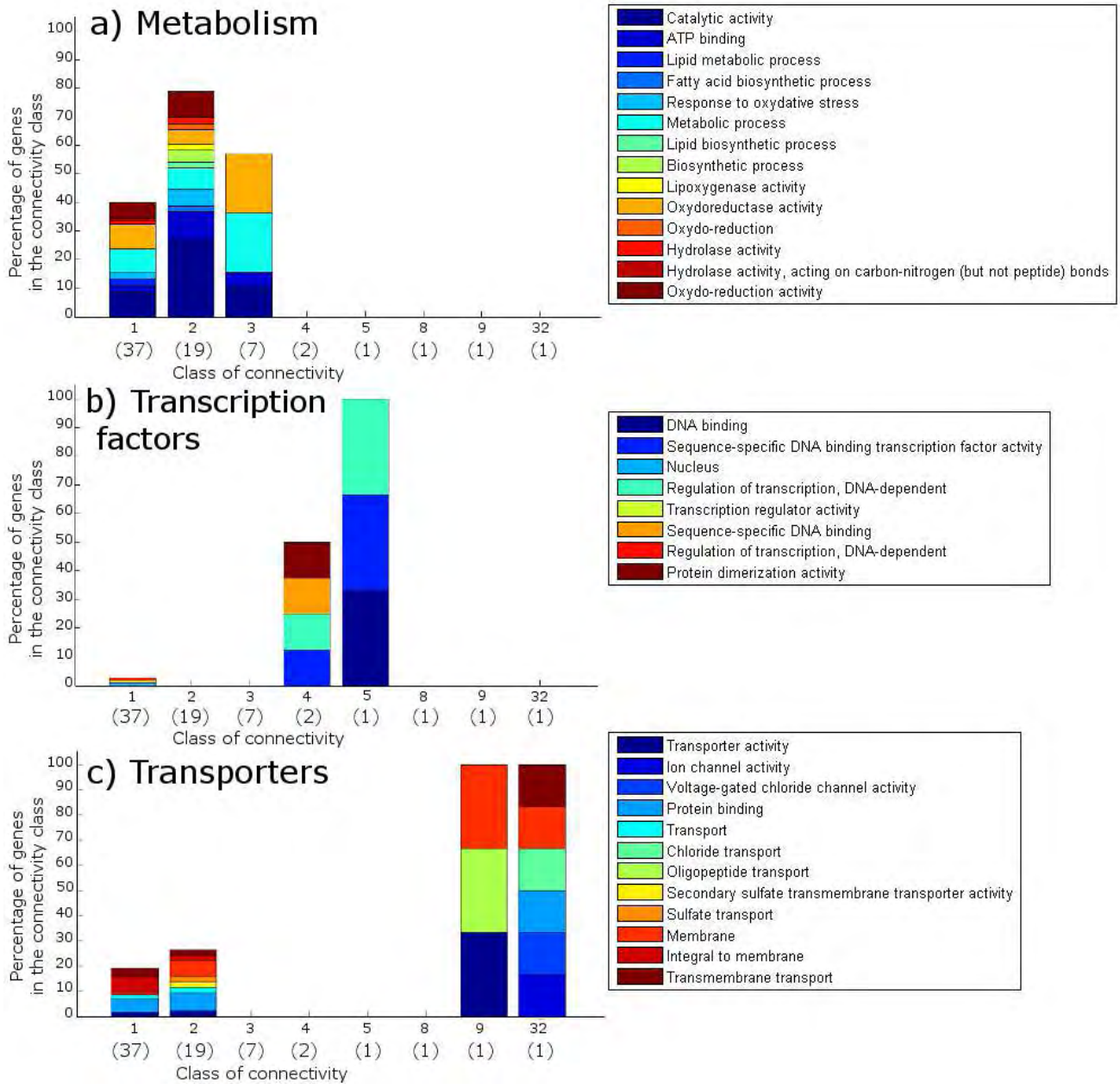


Figure IV.6:Percentage of genes within the drought GRN in each class of gene connectivity, with the GO representation in each class indicated by different colored bars.

a) Metabolism. b) Transcription factor or DNA binding. c) Transporters. Number of genes in each connectivity class is indicated between brackets. Note that the connectivity classes of 8 is represented by a unique gene which does not belong to any of the three main classes of GO represented here (metabolism, transcription factor and transporters).

We observed that genes in the GO metabolism category accounted for the majority of the genes with low connectivity values: 40%, 80% and 60% of the genes with a connectivity of one, two and three respectively, however there was no significant enrichment using a hypergeometric test (p=0.190). More interestingly, genes annotated as transcription factors and as having DNA-binding properties exhibited medium connectivity (i.e., four to five edges, p=0.002), with the exception of one gene that had a single edge, possibly because its targets were filtered out during our analysis. Finally, the most highly connected genes were anion transporters. While the GO transporter included 20-30% of the genes with low connectivity, it also contained all the genes with high connectivity, including both hubs, which had 9 and 32 edges (Figure IV.6). The test showed that despite the very low number of highly connected genes, this trend was significant (p=0.059).

IV.3.5.4.Canonical correlations between the topological parameters of the drought GRN and genetic differentiation statistics

To examine how the drought GRN might be related to the evolution of wild and domesticated sunflower populations, we looked for canonical correlations between non redundant network topology parameters and the genetic differentiation statistics of the drought GRN nodes or genes. The topological parameters for each node represent quantitative measures of the gene position and relationships to others in the network. They are calculated from the number of oriented edges that connect one gene with another and are not independent by construction. In our GRN, edges are oriented, thus, we only considered genes with outgoing edges to compare the predictive value of the topological parameters. In addition, we were able to calculate F_{ST} for 15 of these genes among five populations of *Helianthus*: wild *H. annuus*, landrace lines of *H. annuus,* elite lines of *H. annuus*, *H. petiolaris*, and *H. argophyllus*.

In a first step we used results from the PCA (cf Table IV.2.a and Figure IV.7) with topological parameters to reduce dimensionality and to obtain independent variables. The first and second components explained 67% of the variance. Regarding their loadings on the first two principal components, we selected ASPL and EdgeCount (cf Table IV.2.b).

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation of components	1.771	1.484	0.973	0.902	0.862	0.017	0.002
Proportion of cumulative variance	0.4	0.681	0.801	0.905	1	1	1

Table IV.2.a: Results of the Principal Component Analysis on the topological parameters for the drought GRN: standard deviation and proportion of cumulative variance of components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
AverageShortestPathLength	0.517	-0.129			0.336	0.487	0.603
BetweennessCentrality	-0.257	-0.179	-0.249	0.785	0.466		
ClosenessCentrality	-0.542	0.111			-0.190	-0.222	0.774
Eccentricity	0.530			0.118	0.110	-0.813	0.157
EdgeCount		0.659			0.221		
NeighborhoodConnectivity	0.303	0.159	-0.251	0.485	-0.733	0.210	
Outdegree		0.621	-0.336	-0.130	0.191		
Stress		0.307	0.862	0.332			

Table IV.2.b: Results of the Principal Component Analysis on the topological parameters for the drought GRN: loadings of the topological parameters on each components

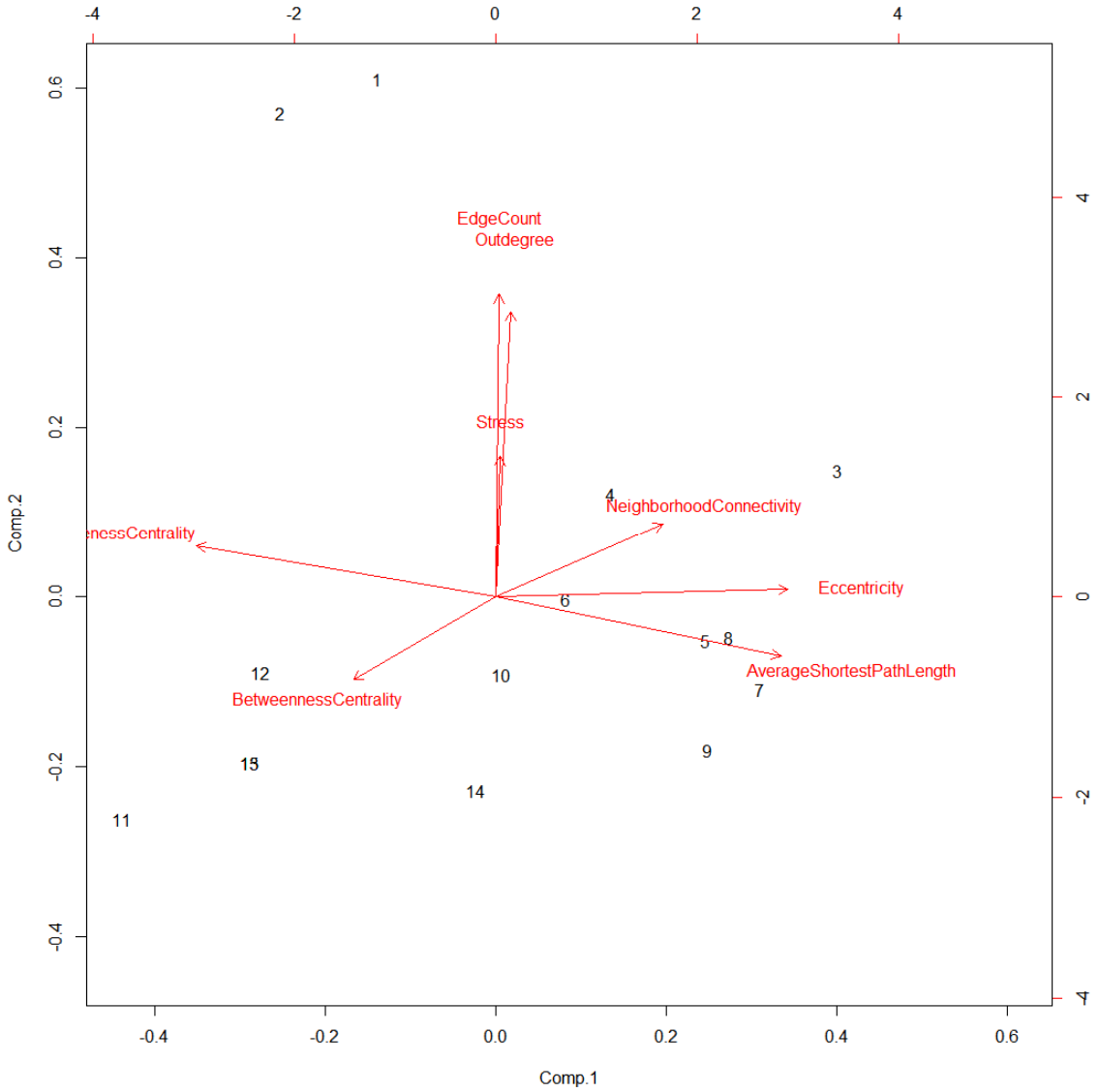


Figure IV.7: Bi-plot of the effects of the topological parameters in a Principal Component Analysis. Components 1 and 2 are shown.

Genetic differentiation was analyzed using five distinct F_{ST} subsets each of them expressing the F_{ST} between one *Helianthus* population and the other populations. Canonical correlation analysis (Table IV.3 and Appendix IV.8) between each of these five F_{ST} subsets on one side, and the two topological variables selected on the other side allowed to detect significant canonical correlations only for the Elite F_{ST} subset (Wilks's test p= 2.00 x 10^{-3}) and for the Landrace F_{ST} subset (p=1.00 x 10^{-4}). As the intersection between these two subsets was F_{ST} between Elite and Landrace, this suggests that this variable in particular is correlated to the topological properties of the GRN. It was confirmed by the comparison of the canonical correlation analyses including only the F_{ST} value between Landraces and Elite lines (Wilks's test p =1.90 x 10^{-3}) or the F_{ST} value between Landraces and Wild (Wilks's test p = 0.26). More specifically, we found a significant correlation of Pearson between F_{ST} value between Landraces and Elite lines and ASPL (R = 0.74, p=0.003).

	Rho Correlation coefficient 1	Rho Correlation coefficient 2
F_{ST} subset of *H. argophyllus*	0.672 (p-value= 0.299)	0.524 (p-value: NS)
F_{ST} subset of *H. petiolaris*	0.493 (p-value=0.818)	0.369 (p-value: NS)
F_{ST} subset of *H. annuus* Wild	0.728 (p-value= 0.362)	0.292 (p-value: NS)
F_{ST} subset of *H. annuus* Landraces	0.976 (p-value= 1x10^{-4})	0.299 (p-value: NS)
F_{ST} subset of *H. annuus* Elite lines	0.946 (p-value=0.002)	0.280 (p-value: NS)

Table IV.3: Coefficients of canonical correlations between in one hand, topological parameters values of the drought GRN nodes and in another hand, their genetic differentiation measured as FST and grouped in five subsets.

Each subset of F_{ST} compares genetic differentiation of one population *Helianthus* in comparison to the four other populations of *Helianthus*. Correlation superior to 50% were tested for significance with Wilks's test. P-value of Wilks's test are shown.

IV.3.6. Discussion

In this study, we reconstructed a GRN based on gene expression that portrays the transcriptional regulations that occur within a plant organ in response to environmental cues. As such, this drought GRN is not based on physical interactions between gene products and promoters and thus is not a molecular cell biology model. Instead, this GRN provides a more physiological view based on transcriptional events involved in drought stress responses similarly to the study of Hannah *et al.*, (2006) on freezing tolerance in *Arabidopsis*. In addition, due to the temporal approach, the network edges are oriented and can be interpreted as dependent relationships. Together, these characteristics produce a network based on molecular regulations that also integrates physiological

processes with their chronology at the organ level. This provides a representation of plant physiological responses to dry conditions and therefore of the fitness in such an environment.

IV.3.6.1. Network inference highlights the importance of nitrate transport in guard cells

Drought GRN hubs are nitrate transporters and drive transcriptional regulation

In the inferred network, two genes had many outgoing connections compared with other genes and could therefore be considered hubs. The first hub (HaT13l030730) is homologous to the transcript of the *Arabidopsis* gene chloride channel A (CLC-A, AT5G40890). CLC family members are involved in anion compartmentalization in intracellular organelles and in stomatal guard cell vacuoles (Jossier *et al.*, 2010). More precisely, CLC-A and CLC-C are expressed in stomata and control their opening through translocation of NO^-_3 and Cl^-, respectively. This difference in anion selectivity among the CLC family members is due to an amino acid change in the selectivity filter (Wege *et al.*, 2010). The sunflower transcript HaT13l030730, which is homologous to *Arabidopsis* CLC-A, possesses the same amino acid conferring nitrate specificity. This suggests that the main hub identified in the drought GRN is likely a nitrate channel involved in stomatal aperture control and, therefore, transpiration.

The second hub (HaT13l003541) is homologous to the transcript of the *Arabidopsis* gene NRT1.1 (AT1G12110), which encodes a dual-affinity nitrate transporter in *Arabidopsis*. Guo et al. (2003) demonstrated that this gene is expressed in guard cells of stomata and that transpiration is affected in mutants in an ABA-independent manner. The reduction of the stomatal aperture in mutants appeared to be due to nitrate uptake in guard cells. The control of stomatal transpiration by anion channels and transporters in guard cells was further confirmed (De Angeli *et al.,* 2013) in *Arabidopsis*.

Our approach identified the key role of two sunflower homologues of *Arabidopsis* anion transporters. This strongly suggests that this process is important for the regulation of the sunflower drought response. However, the two hubs do not directly regulate the expression of their target as transcription factors do; instead, the hubs drive downstream signaling cascades through indirect physiological and distant regulations.

The drought GRN identifies connections between ABA-dependent and ABA-independent pathways

In the inferred network, both hubs had seven common targets but no common source. This suggests that the NRT1.1 and CLC-A sunflower homologues could represent two pathways controlling drought stress responses. However, we could not exclude a cross-talk between NRT1.1 and CLC-A with an upstream regulator absent from our initial dataset. By inferring sunflower gene function based on *Arabidopsis* homology and the analogous expression response to drought, we could

tentatively investigate the molecular pathways characterized in the sunflower drought GRN. Functional annotation of the targets of the two hubs revealed genes that are directly involved in cell protection and stress tolerance, such as the ROS scavenger (APX1) and two enzymes involved in synthesis of an osmo-protectant, choline (PMEAMT and CCT2). Interestingly, we also identified genes involved in signal transduction, such as kinases (HaT13l074901 and emb1075), phosphatases (HAB1), calmodulin-binding proteins (CPK5), and transcriptional regulators (MYC2, ARIA), downstream of the anion transporters, as described in Figure IV.8.

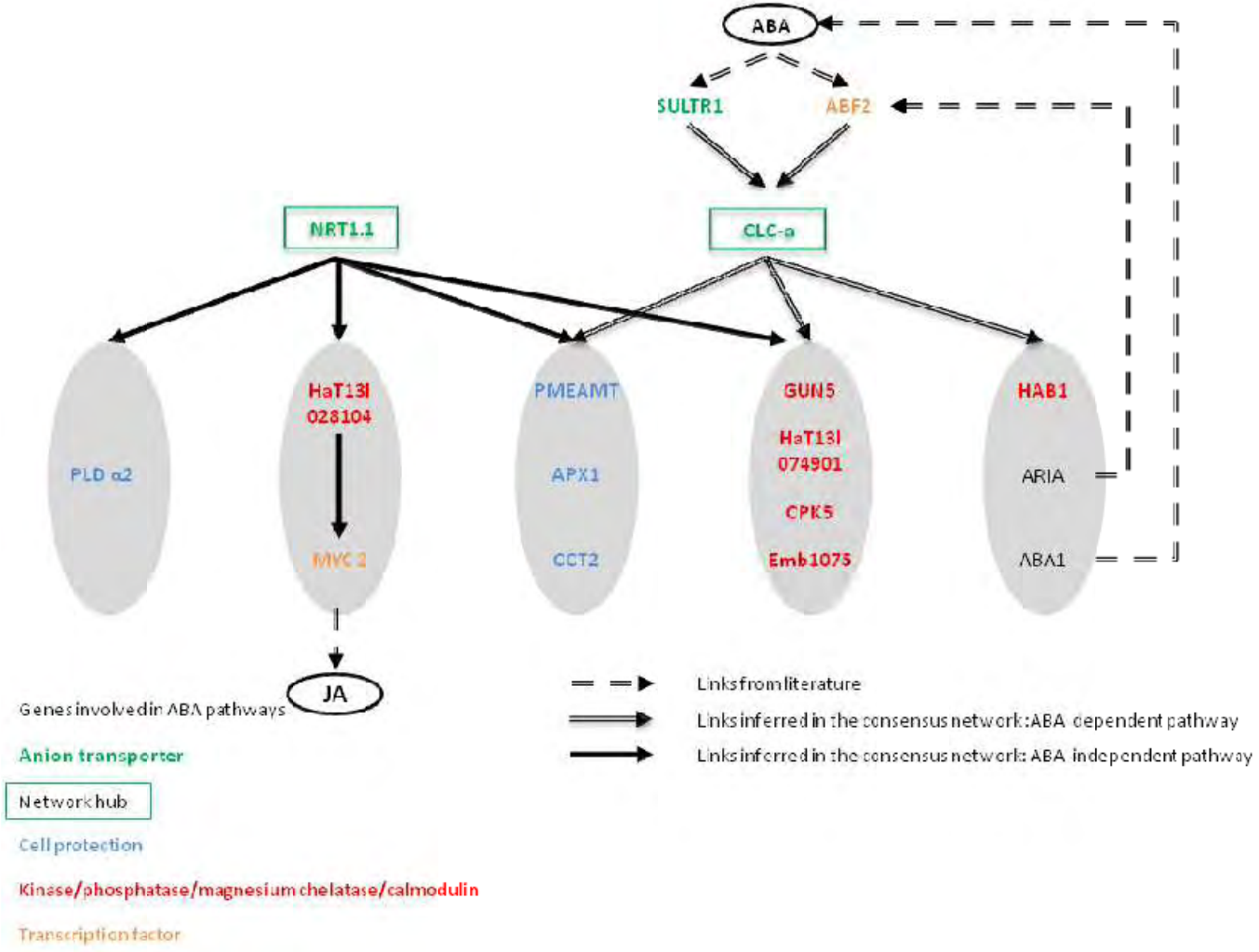


Figure IV.8: Functional network involving the two hubs of the inferred drought GRN, their sources, and their targets.

Blank edges represent the ABA-dependent pathway, including the CLC-A. Solid edges represent the ABA-independent pathway, including NRT1.1. Common targets involved in signal transduction are indicated in red, those involved in transcriptional regulation are shown in orange, and those involved in cell protection are shown in blue.

CLC-A and NRT1.1 define an ABA-dependent and an ABA-independent, respectively, pathway in our experimental results, as well as in *Arabidopsis* (Guo *et al.*, 2003; Jossier *et al.*, 2010). Both sources of CLC-A, SULTR1 and ABF2, are regulated by ABA in *Arabidopsis* (Fujita *et al.*, 2005; Ernst *et al.*, 2010) and also in our experiment for ABF2. In addition, specific targets of CLC-A are part of the ABA signaling cascade in *Arabidopsis*. HAB1 is a protein phosphatase that is strongly up-regulated by ABA (Rodriguez, 1998) and functions in ABA signaling. ABA1 is known to catalyze the first step of ABA synthesis (Rock & Zeevaart, 1991), and ARIA is an armadillo repeat protein that is known to interact with the transcription factor ABF2 (Kim *et al.*, 2004). Together, these regulatory connections identified in *Arabidopsis* form a loop involving ABA synthesis (in vascular cells) (Boursiac *et al.*, 2013) and a signaling pathway across the different cell types (including guard cells) throughout the leaf (Figure IV.8). In the drought GRN, we were able to partially identify the corresponding regulatory loop between sunflower homologues. These results suggest that the same ABA regulatory loop exists in the sunflower drought GRN and therefore could be largely shared across the plant kingdom.

Similar to the shared targets of CLC-A and NRT1.1, specific targets of NRT1.1 are also involved in cell protection (PLDα2) and signal transduction (HaT13l028104). An interesting downstream target is MYC2, which is a central regulator of the hormone jasmonate, which is mostly involved in plant defense and the development and integration of many hormonal signals (Kazan & Manners, 2013).
Across the sunflower drought GRN, several different pathways show some conservation across plant species, such as *Arabidopsis*. Therefore, the GRN inference approach developed in this study appears to be robust, and we can make the reasonable hypothesis that the main regulatory pathways and hubs identified in the drought GRN are likely conserved among distant plant species and therefore also across the *Helianthus* genus. Although, from our data we were not able to demonstrate the network conservation across *Helianthus* population (it would require inferring the network for each one which would be too laborious with the present technologies), this hypothesis allows us to explore new questions about how the GRN could constrain plant adaptation to dry environments.

IV.3.6.2. Drought GRN topology and Helianthus evolution

Network topology constrains genetic variation of the gene network

Gene networks are the products of evolution, similarly to other biological objects, but gene network relationships can also constrain evolutionary changes, such as adaptations to new environments and responses to selective pressure during domestication or breeding. For example, (Rausher *et al.*, 1999) demonstrated different evolutionary histories for upstream and downstream genes in the anthocyanin biosynthetic pathway.

The evolution of the GRN architecture can lead to new nodes, potentially introducing new functions and new edges between these nodes. Previous researchers (Hinman *et al.*, 2003) examined GRN evolution in echinoderms and demonstrated that some features of developmental GRNs were conserved and that others were specific to each taxa. Network architecture is known to affect evolutionary rates (Ramsay *et al.*, 2009), and we expect evolutionary changes to the nodes to be constrained by their connectivity and the number of neighbors. A hub in the network is involved in several pathways. The functional trade-offs for such genes are higher than those for peripheral genes that are neither involved in regulatory processes nor in the interaction with partners.

To understand how populations and species evolve and adapt to a new environment, we examined the putative constraints of the network architecture on the genetic differentiation between populations of *H. annuus,* and two wild species that are cross-compatible with *H. annuus*: *H. argophyllus* and *H. petiolaris*.

No evidence of network topology constraints during the divergence of H. argophyllus and H. petiolaris

Helianthus argophyllus is native to the dry, sandy soils of southern Texas, an arid environment that imposes strong selection for tolerance to drought stress. Indeed, *H. argophyllus* is considered the most drought-tolerant sunflower species because its pubescent leaves reflect sunlight, reduce water loss, and exhibit low transpiration (Seiler & Rieseberg, 1997). However, network topology and F_{ST} values between *H. argophyllus* and other populations were not significantly correlated. This could be because the adaptation of *H. argophyllus* to dry environments involved physiological mechanisms that are not captured in our GRN or because the network topology has itself evolved and the topological parameters in H. *argophyllus* are too dissimilar to those in *H. annuus*. Interestingly, the highest value of F_{ST} between *H. argophyllus* and other populations was for the network hub, NRT1.1, which is involved in transpiration. This result is consistent with positive selection acting on NRT1.1 during adaptation of the *H. argophyllus* to dry environments. Keeping in mind the overall non-significant correlation, it suggests that NRT1.1 could be an example of the fore-mentioned hypothesis.

In *H. petiolaris,* we observed no correlation between the GRN topology and F_{ST} for comparisons with other populations. Because *H. petiolaris* has a large geographic range that overlaps with that of *H. annuus* in the Great Plains of the USA, drought stress might not be the major selective force separating these species. This could explain the similar divergence patterns within the drought network genes between these two populations as illustrated in Figure IV.9.b.

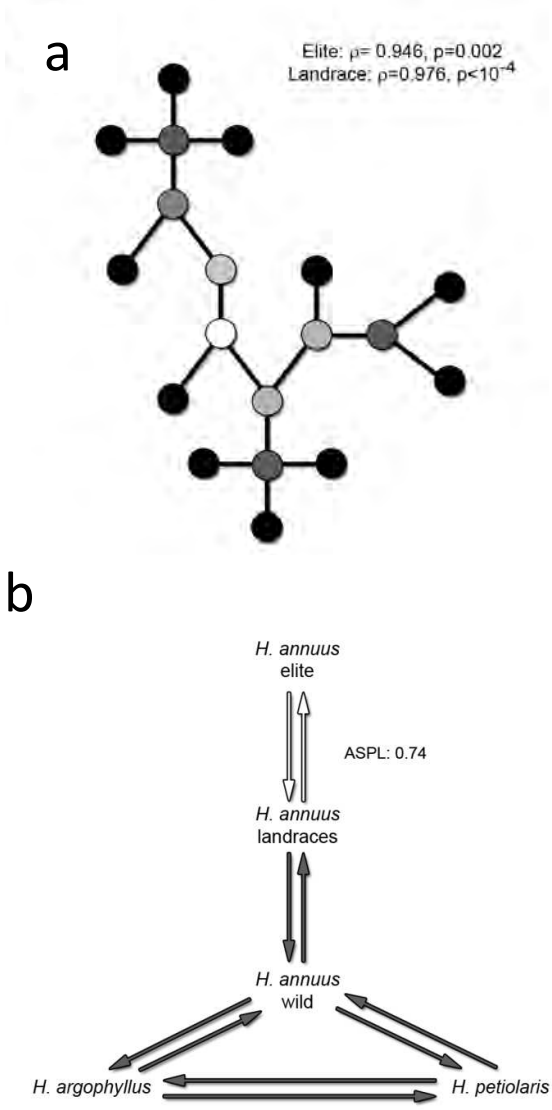


Figure IV.9: (a) Representation of genetic differentiation between *H. annuus* Landraces and *H. annuus* Elite Lines in function of the gene positions in a schematic gene regulatory network.

Colors of the genes in the schematic GRN represent the difference of heterozygocity between the two populations for the considered gene. Node color represents differentiation between elite and landrace sunflowers: darker nodes appeared more differentiated compared to lighter nodes. Canonical coefficients (ρ) and p-value of the Wilks's test for the correlations between network topological parameters and F_{ST} values of one population compared to the others are indicated for Elite lines and landraces. **(b). Hypothesis about differences of genetic differentiation between the five *Helianthus* populations.** Note that only the five comparisons representing the selective history of the sunflower are shown. Black edges indicate no variability in the genetic differentiation within genes network between the two populations. White edges indicate changes in the genetic differentiation between populations as observed for the 15 genes in the drought GRN analyzed in the CCA. The coefficient of the Person's correlation between the topological parameter Average Shortest Path Length (ASPL) and F_{ST} between *H. annuus* Elite lines and Landraces is indicated.

Genetic diversity within the GRN was modified during modern breeding

The network topological parameters and the F_{ST} between the landraces and elite lines of *Helianthus annuus* were correlated (Figure IV.9.a). This reflects a difference of genetic differentiation between these two populations between the center and the periphery of the network. We did not observe this correlation for F_{ST} between wild *H. annuus* and landraces. This suggests that the position and connectivity of genes in the drought GRN influenced the response to selection during the last century of genetic improvement but not during the initial domestication of *H. annuus*. This difference in selective responses could be due to the fact that highly connected genes are subjected to more trade-offs as they are master regulators with involvement in several genetic pathways in contrast to less connected terminal genes (Figure IV.9.a). Drought tolerance is considered to be a long standing goal of sunflower breeders. We would expect that the selection they exert had led to a global reduction of genetic diversity in the drought GRN. However, we observed a higher divergence of terminal genes compared to central ones, which implies a stabilizing selection acting on the network hubs. Interestingly, our F_{ST} studies in *H. argophyllus*, suggest that a different selective pressure acted on one of the network hub (Figure IV.9.b). This highlights our lack of global understanding on how evolutionary forces and functional relationships interacted to produce contemporary phenotypic diversity and suggests a potentially important way of improving the breeders' methods, through the integration of regulatory networks in quantitative genetics models such as genomic selection.

In conclusion, this work investigates the interaction between physiological and evolutionary processes in the context of a genetic network for the drought-stress response. Interactions between physiological and evolutionary time scales could be revealed in the future through global transcriptomic studies, although some limitations of network inference methods remain to be overcome. This type of work will facilitate the study of responses to other environmental factors and clarify whether physiological mechanisms and evolutionary adaptation, which are reciprocally constrained in the gene regulatory network, are similar in abiotic and biotic interactions.

End of article "Bridging physiological and evolutionary time scales in a gene regulatory network", accepted on March 2014 in New Phytologist

IV.4 Main conclusions about the drought GRN and transcription regulation

In this work we have been able to build a drought GRN thanks to a systems biology approach. In the inferred network, main groups of regulatory pathways already identified in literature have been highlighted as phosphatase cascades or the induction of genes involved in cell protection (see previous section). The gene regulatory network also distinguishes an ABA-dependent pathway and an ABA-independent pathway as documented in numerous studies about responses to drought stress (Bray, 2004; Shinozaki & Yamaguchi-Shinozaki, 2007; Todaka *et al.*, 2012).

Several transcription factors in this inferred GRN have numerous connections with other genes, suggesting that they directly or indirectly regulate their transcription levels. This result is consistent with the concept of transcription factors that are considered to be the main actors of transcription regulation at the molecular level. Surprisingly, however, the most highly connected genes in the inferred regulatory network are two anions transporters. As already detailed in the previous section, anion transporters have been demonstrated to be involved in stomatal movement, in the model plant *Arabidopsis*. We can draw new hypotheses to explain the high connectivity of those anions transporters in the network. Changes in stomatal aperture and in cell osmotic potential might have a dramatic impact of the cell state and produce a major physiological reprogramming that would indirectly induce changes in transcription level of other genes involved in drought stress responses. Therefore, genes originally classified in the group of the effectors genes, as the anion transporters described in our work, also play an indirect but important role in the transcription regulation of drought responsive genes. It highlights that feedback loops between effectors genes and transcription factor exists and might have a major role. Then, the two distinct groups identified in the generic cascade for drought stress (transcription factor for transcription regulation and effectors genes involved in drought tolerance mechanisms) are likely involved in a same GRN with permanent feedback loops between them.

IV.5 Outlooks for the drought GRN study

In order to carry on the study of the drought GRN several complementary researches could be conducted with different objectives.

IV.5.1 Functional characterization of the inferred drought GRN

The gene regulatory network is inferred thanks to partial correlation between gene expressions. The connections between two genes are not completely demonstrated adopting a point of view of molecular biology. Therefore, for some important regulatory pathways, demonstration of

the functional link between two genes could be envisaged using mutant and knock-out strategies. In particular, this complementary approach could be used for the functional characterization of the regulatory pathways involving the two anion transporters. Indeed anions transport seems to be an important mechanism in drought tolerance and the detailed characterization of the genes involved in this pathway should be interesting and examined more closely in order to improve the tolerance of sunflower to water deficit.

IV.5.2 Toward a more complete systems biology study

Systems biology is the understanding of a biological systems (here the sunflower) at the system level i.e. in a holistic perspective. The study that we have conducted is only the first step towards this challenging goal and many other research projects will be necessary to achieve this aim. Here, I would like to present some example of studies that would help to gain a global understanding of the sunflower system under drought stress.

First, sunflower is a multi-cellular organism. Therefore, one important question that needs to be answered is the spatial characterization of gene regulation. For the GRN inference presented previously, we used the entire leaf tissue. However, it would be interesting to know in which type of cells the different genes, which are involved in this network, are expressed. For example, we could verify that the anions transporters are indeed expressed in guard cells. This supplementary knowledge about the GRN can now be obtained for example thanks to single-cell RNA-seq experiments (Tang *et al.*, 2009; Brennecke *et al.*, 2013). Still in the same objective of a better understanding of the drought gene regulatory pathways in a multi-cellular organism, the inference of drought GRN could be conducted in other plant tissue such as, for example, the root system, which plays a major role in drought tolerance. Comparison of the leaf and root GRNs would allow us to gather information on the communication between the different plant organs and to define a unified drought GRN which would be even more relevant in predictive biology.

A second enhancement of the inferred GRN would be to improve our understanding of the edges linking two genes. In the work described in this chapter, connection between two genes gives no indication about the relationship between them and how expression level of the first regulate expression of the second. The detection of one edge indicates that a change in expression of the first at one time accounts for a change in the expression of the second at the following time. It is a Boolean relationship. Instead of this Boolean look at edges, a more powerful insight would be achieved by understanding how the expression of a target gene is dependent of the expression of its source genes in a quantitative way. In this new model, edges would represent a function of gene expression with expression level of source genes, stress intensity and time as parameters. This would be the first step in the study of the system dynamics while for the moment only system structures

and their relationships have been described. To achieve this goal, a method such as the bifurcation analysis can be used. Bifurcation analysis traces time-varying changes in the state of the system in a multidimensional space where each dimension represents a particular intensity of the perturbation involved or level of gene expressions (Kitano, 2002).

Such modeling and level of system understanding could be interesting in order to compare gene regulatory network for various environmental stresses or to compare GRN between species. Actually, several abiotic stresses, such as for example drought, cold or salt stresses, share pathways and molecular responses. Therefore, it is very likely that several genes would be found in common in their regulatory pathways using these modeling techniques. Distinction between two pathways that involve similar genes would be accessible with an accurate and detailed knowledge of the mathematical function linking gene expressions in a time and dose-dependent manner.

Finally, as already suggested above, a gene regulatory network should be inferred for the different stresses that are likely to perturb the biological system. This goal can be achieve using whole transcriptomic studies with different datasets for each stress. This will give a complete picture of the system dynamics and lead to the next steps of a system biology approach i.e the learning of the system control and finally of system design.

IV.6 Conclusions and outlooks about *Helianthus* evolution study thanks to GRN

In the work presented in this chapter, we inferred a drought GRN accounting for physiological adaptation for water deficit tolerance. These modifications reflect biochemical, morphological, and phenological changes occurring at the time-scale of an organism life. However, the ambition of this study was also to understand how the particular topology of the gene regulatory network could constrain the adaptation on a longer time-scale such as the evolutionary time-scale. Then it could help to understand how phenotypic plasticity produces phenotypes that can become constitutive in order to adapt species in a new constrained environment. Figure IV.10 presents a schematic view of the generic cascade involved in drought responses and integrates results of this chapter concerning evolutionary constraints. To investigate this process, we made the strong hypothesis that the GRN, that we inferred from cultivated *H.annuus,* is conserved across *Helianthus annuus* and its relatives. An important improvement of this work would be to demonstrate this assumption. It could be verified if the same whole transcriptomic strategy and system biology approach were set up in order to infer a new GRN for each species. Nevertheless, this strategy would not be easy to set up because it is very expensive and time-consuming. Actually, a lot of samples have to be harvested in order to have important dataset for each species. It would also be difficult to obtain the exact same

experimental condition for each *Helianthus* species since they develop very differently and the correspondence of developmental stages to best compare them would be delicate. Indeed the wild accessions need usually more time to reach the same developmental stage than cultivated sunflowers due to seed dormancy

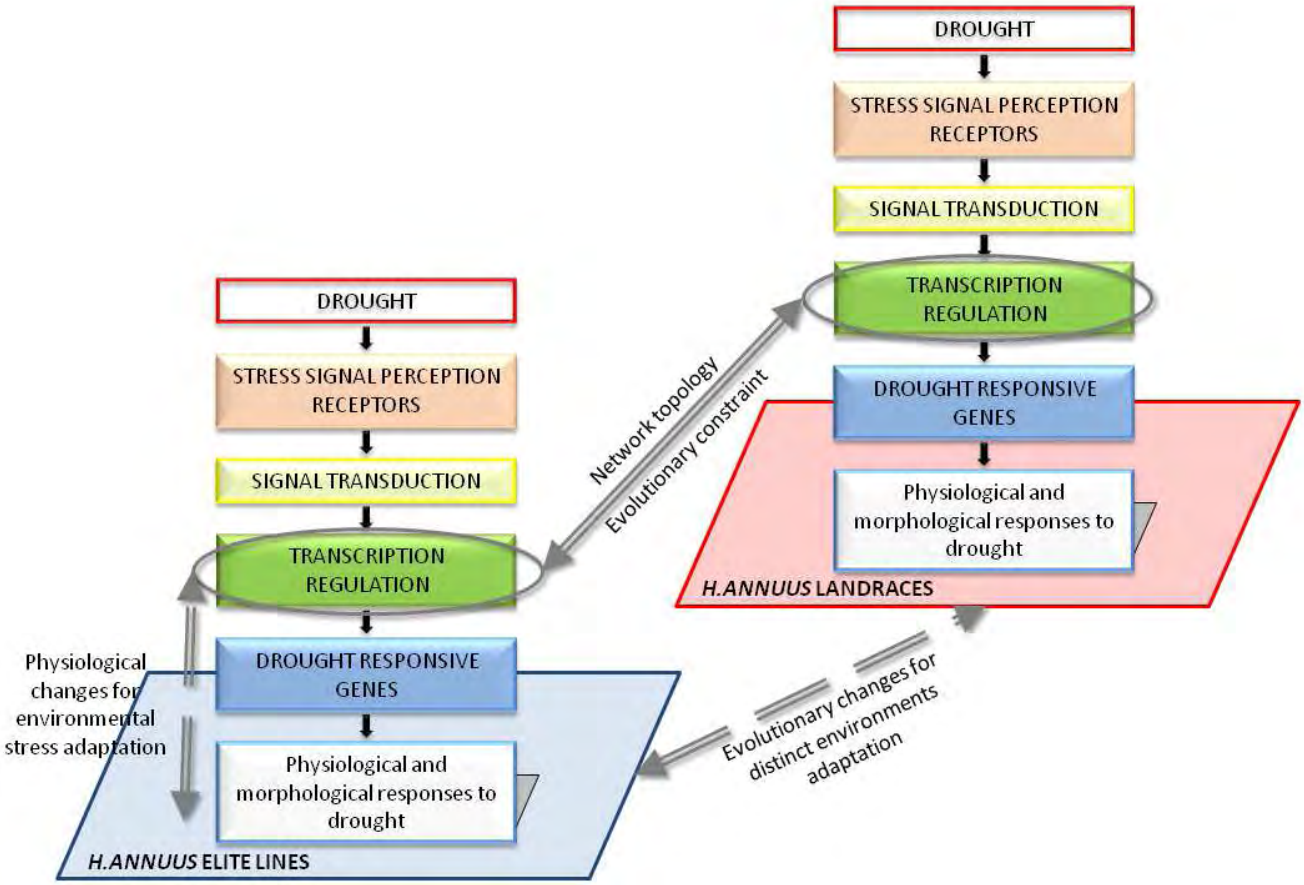


Figure IV.10: Two gene regulatory cascades for drought stress responses in different *Helianthus* annuus species.

This figure shows a schematic representation of the drought GRN explaining the link between physiological changes to cope with environment stress and the evolutionary adaptation of the two species to distinct environments. The network topology gives some insights about how phenological changes could become constitutive and contribute to species evolution history.

Chapter V: Conclusions and perspectives

Drought stress is one of the main abiotic stresses that occur in the field and constrain plants' growth, development, fitness, and yield (concerning species with agronomic interest). To cope with this environmental pressure, plants have developed a large range of mechanisms that take place from the whole plant scale to the molecular level. Moreover, these drought tolerance mechanisms impact various plant functions such as phenology, biomass allocation, transpiration or development. Thanks to numerous studies, major morphological and physiological phenomena that allow drought tolerance begin to be well known. However, their genetic control is far from being completely understood. Although some major regulatory pathways and key genes have been identified, mainly for model plants, a global view of the genetic control of drought stress responses is still lacking. Knowledge about gene regulatory network (GRN) underlying traits involved in drought tolerance could be very useful to answer various questions about drought tolerance: What is the genetic control of the perception of the environmental signal by the plants? How is controlled the phenotypic plasticity for traits involved in drought tolerance? How does the physiological adaptation to drought relate to species evolution? For the particular case of the sunflower, bringing to light these points could be interesting for breeders since one of their major challenges for the coming years is the improvement of drought tolerance.

Throughout this PhD work, I have attempted to answer those questions. The strategy was to study the main classes of drought responsive genes involved in the different steps of the response to the environmental signal in order to obtain more insight into the gene regulatory network(s) underlying morphological and physiological traits that confer drought tolerance.

V.1. A more complex picture of the genetic control of drought stress responses

V.1.1. Genes involved in the perception of the drought signal and cross-talk between the plant and its environment

The first class of genes that we studied comprised genes involved in the drought perception. The goal of this first part of the work was to highlight some genes whose expressions only depend on the intensity of the water stress. Response to drought stress differs between two genotypes. It can be explained by the genotype-dependent expression of genes involved in the cascade for water stress responses. Several hypotheses can be constructed concerning which class of genes is

genotype-dependent or genotype-independent. The two extreme hypotheses would be (1) all genes are genotype-dependent and (2) no genes are genotype-dependent (this last one not being able to explain the difference of response for drought between genotypes). A transitional hypothesis, our original one, would be that the genes first involved in the perception of the environmental signal could be genotype-independent and then genes involved in transcription regulation and coding for effector proteins could be variably expressed according to the genotypes. Therefore genes involved in the environmental perception are assumed to be candidates for genotype-independent genes. In our results we identified three genes with expression correlated to plant water status (genes used to build the water status biomarker described in chapter II) and with no genotype effect in the range of the genetic diversity observed (four genotypes in our study and eight in the previous study of Rengel *et al.,* 2012). Those genes were either transcription factor or genes coding for effectors involved in the molecular responses to drought. Therefore a reasonable hypothesis would be that (at least within an operational range of genetic variability and of environmental conditions, both of them being of interest for sunflower breeding) likely two cascades for drought stress responses are involved and probably with cross-talk between them: one with only genotype-independent genes and the second with a mix between genotype-dependent and genotype-independent genes. Therefore genes used for the WSB construction would be either at the beginning of the mix cascade or anywhere in the genotype-independent cascade (figure V.1). It is also important to keep in mind that genotype-independent expression of the genes was evaluated on a limited genetic variability taking into account only eight (Rengel *et al.*, 2012) and then four (Marchand *et al.*, 2013) genotypes. Widening the genetic diversity of the study could lead to restrict the number of genotype-independent genes and make a stronger hypothesis about where those genes are located in the generic cascade controlling drought stress responses.

The first part of the PhD work shows the existence of an oriented link between expression of genes involved in environmental signal perception and genes directly supporting responses to water deficit as, for example, genes coding for effectors proteins involved in physiological responses to drought. During the association study (see chapter III), genetic loci responsible for water status variations (estimated through the use of the WSB) were identified. This genetic control is the sign of a link between effectors or regulatory genes involved in water deficit responses and the genes involved in the water status perception (figure V.1). A possible interpretation of this result would be the existence of a system where the plant controls its water status and therefore its micro-environment. This adjustment toward a new water status is genotype-dependent, as it is the consequence of the genotype-dependent strategy of the plant to tolerate drought stress. It can therefore be interpreted as a cross-talk between the plant and its environment, each influencing the other.

In conclusion about the genes receptor or indicator of the water status, we can retain that at least a part of those genes could be genotype-independent observing a narrow genetic diversity (i.e. few genotypes express those genes in a same way that depends only on their water status). They are located at different levels along the generic cascade for drought stress responses. Moreover, the very likely existence of a feedback loop between these classes of genes allows adjustment between the plant, the water status, and its micro-environment. This adjustment is proper to the genotype and reflects its strategy for drought tolerance.

V.1.2. Existence of feedback loops between regulatory genes and effectors genes

The second class of genes that we studied was involved in the signal transduction and the control of the transcription regulation for down-stream genes. Selecting a part of those down-stream genes, we reconstructed the GRN that links them through network inference in a systems biology approach (see chapter IV). It allowed us to identify the hubs of this network. Those highly connected genes induce many downstream genes and therefore are important factors in the transcriptional regulation. Surprisingly, the main hubs of our re-constructed GRN are two anion transporters. The transcription factors, known for their action in transcription regulation, are found only in the second rank of gene connectivity. Anion transporters in the model plant *Arabidopsis* are involved in stomatal closure. The functional protein domains are likely conserved in the *Helianthus* homologues. Hence, anion transporters may occupy a major role in drought tolerance regulation in sunflower. We can discuss the fact that the most highly connected genes were not transcription factors as we could originally supposed. The first hypothesis would be that in our genes selection we missed major sunflower transcription factors that regulate the transcription of downstream genes or failed, in the GRN inference, to link them with their target genes. However, the connections between anion transporters and their target genes could not be false positives and therefore those hubs are very likely important genes for the sunflower GRN. That leads us to our second hypothesis where genes such as anion transporters, originally classified in effectors genes for drought stress responses also play a role in the regulation of the drought responsive genes. This indicates that the genes involved in the transcriptional regulation as transcriptional factors and effectors genes are involved in a same GRN with permanent feedback loops between them. The links between those genes are represented in Figure V.1.

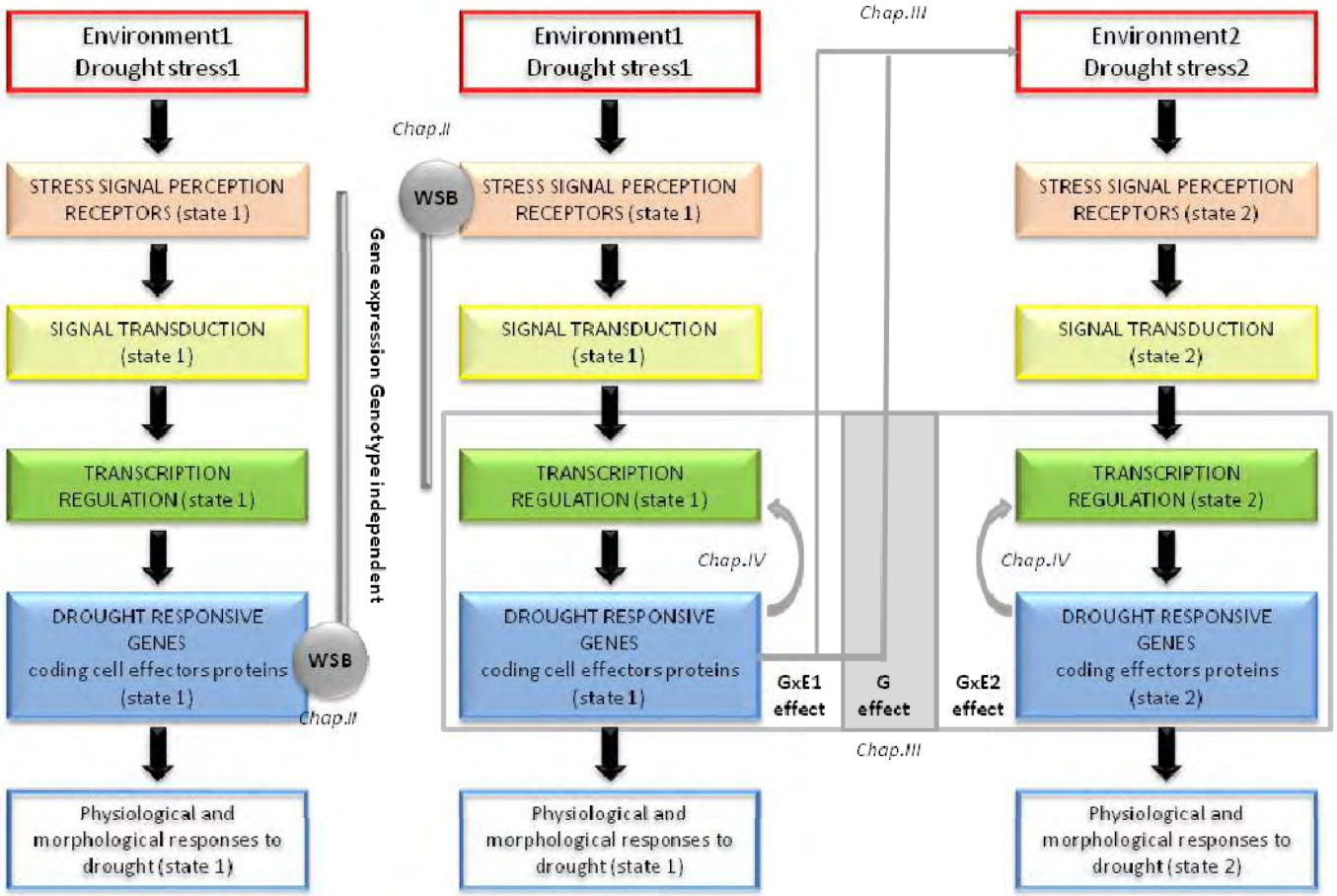


Figure V.1: Generic cascade of genes involved in drought stress responses with a summary of all results pointed out throughout the PhD work.

Additions in comparison to the cascade adapted from Huang et al., 2011 are in grey and related to the respective chapter. Genes with expression independent of the genotype are either at the end of a whole pathway of genotype independent genes or at the beginning of a mixed pathway with expression of genes genotype-dependent and independent (Chapter II). There are feedback loops between effectors genes and regulatory genes and genes preceptor of the environment. Expression of effectors genes can be governed by genotypic (G) effects and Genotype x environment (GxE) effects.

V.1.3. Genetic architecture of genes underlying morphological and physiological traits conferring drought tolerance

The last class of genes that we studied in this PhD project concerned the genes involved in the physiological and morphological traits that drive drought stress tolerance in the plant. Genes with expression correlated to those traits (Rengel *et al.*, 2012) were used as phenotypic traits in the association study (see chapter III). This allowed us to link the physiological level to the genetic level using the intermediate molecular level in drought stress responses. Therefore the results give us a first look of the genetic control of those genes whose expression is genotype-dependent and also water status-dependent. This study was also a first attempt at understanding the genetic basis of the phenotypic plasticity of traits involved in drought tolerance and at identifying the underlying mechanisms such as the genotype x environment interaction. Associations between gene expression levels and SNPs found in the vicinity of the genes (cis-regulations) had only genotype-dependent effects, whereas the distant (trans-) regulations were more likely subjected to a genotype x environment interaction effect that can modulate the genotype-dependent effect. In our study, we found only few genes under the control of a significant genotype x environment effect. This could be due to different factors: choice of genes, not enough stressful conditions in our experiment (see chapter III). However, it can open the way for a more important study comprising, for example, a wider selection of genes. Differences between genetic controls of the drought responsive genes are summarized in Figure V.1.

V.1.4. From a "simple" gene cascade to a more complex picture of the drought gene regulatory network

All these different parts of the PhD work permit to detail the generic pathway for drought stress responses. They gave supplementary information about the location of the genes and the interaction between them that allow the complex regulation of the drought responses. This work also yielded some evidence about the genetic architecture of the regulation of these genes, with the distinction between genes whose expression is genotype-dependent/independent and the intensity of the genotype x environment interactions effects involved in their regulations. Our original and naive vision of the generic pathway (presented in chapter I) was linear with different classes of genes that take part successively in the responses to drought stress. This background picture of the drought tolerance genetic control has been very useful to conduct the first approaches to study water deficit stress and gain a first understanding of the drought tolerance phenomenon at the genetic level. However, from the results obtained during this PhD work and several other research projects using systems biology approaches, this first version is obviously too simple. A more appropriate model for sunflower drought regulatory pathways implying several feedback loops can be drawn as shown in Figure V.1.

V.1.5. Limits of our approach and future perspectives

Results provided by the different sections of this work provide some insights into genetic control of drought stress responses with a first attempt to obtain an overview of the regulatory processes involved in sunflower drought responses. Indeed, we used systems biology approaches and whole-genome association mapping strategy to explore the genetic architecture of drought tolerance traits. This led to the reconstruction of several drought GRNs based on different experimental data sets or approaches. If these representations of the water stress genetic regulation allow approaching a more complete view of their complexity than the previous linear cascade, we have to keep in mind that this model is still a largely incomplete vision. It only allows one to perceive the complexity of these regulatory mechanisms. In addition to that, at each step, only a small fraction of sunflower genes was studied. Therefore, the picture can still only be considered to be fragmentary.

For the future, as already discussed in a previous chapter of this work (see chapter IV), a more complete systems biology approach should be considered. The next studies should, hence, tend to take into account the whole sunflower gene set and the kinetics of their expression in different water stress intensity in order to achieve a global reconstruction of the drought GRN. This is in fact an ambitious work and a major challenge. Nowadays, obtaining expression levels of thousands of genes does not form an obstacle any more, but there are still other technical problems to overcome to obtain the final GRN model. The first one is the important phenotyping investment necessary to draw the kinetic curves of gene expression in several drought stress conditions. The second one is the computational problems that involved a GRN inference with thousand of genes in different conditions. Nevertheless, systems biology approaches seem a promising way to obtain a more complete picture of the drought regulatory pathway that govern responses to water deficit.

V.2. From physiological acclimation of a genotype to the species evolution and adaptation

In the last part of this work we adopted an even broader point of view. The GRN we built in this last part of the project is a way to aggregate different information (direct environmental signal, plant physiology and phenological stages, molecular levels, etc). This involves many genes in order to control physiological and morphological responses to adapt the phenotype of a genotype to its environment. Assuming that main regulatory pathways are conserved among *Helianthus* species, GRN can be a prism to interpret species evolution and how they can adapt to and colonize new

environments with a different drought constraint. According to the GRN topology, selection pressure should be different between genes (see chapter IV).

Once again, the results about *Helianthus* evolution described in this work are only preliminary findings that should be treated with caution. Indeed, the study is far from taking into account all genes involved in the mechanisms driving to evolution and adaptation of sunflower to dry environments. However, we believe that using GRN to draw a link between physiological adaptation that occurs at the organism time-scale and the species evolution could yield interesting results in specific contexts..

It could be used, for example, to study evolution of species due to an environmental constraint that could be biotic or abiotic. This could also help predict the liable target genes for heterozygosis losses and establish long-term scenarios about genetic diversity losses after selection for the specific trait regulated by the observed GRN.

V.3. Perspectives of utilization in a crop model

A crop model is a way to simulate the functioning of a plant i.e. development (including growth and phenology) and physiology (including environmental response and biomass allocation), in order to predict the yield of the crop. To achieve this goal, the complex system that the plant represents is simplified in order to keep only the major factors that impact the final output of the model i.e the crop yield. Therefore, a crop model has three main objectives. The first one is to gather interdisciplinary knowledge of plant and crop functioning. The second is to introduce enough complexity and knowledge about plant functioning in order to have a good estimate of the yield and an accurate view of the crop development. The third is to be flexible and open to further development, which means that knowledge has to be simplified in order to keep the model easy to manipulate. The challenge of crop modeling is, therefore, to compromise between these different goals.

A crop model for sunflower, called SUNFLO has been developed by Casadebaig *et al.*, (2011). This crop model has input parameters taking into account information about the environment on one hand (climate, soil, crop management, nitrogen availability) and information about plant development in non-constraining conditions and about abiotic stresses sensibility (which are considered genotype-dependent) on another hand. All this information is crossed in a stress module that adjusts the plant functioning according to the environmental parameters (figure V.2).

Among the different applications of this crop model, it can be used to make prediction about the yield under specific environmental conditions, or to help define new ideotypes. Up-to-now, plant parameters implemented in SUNFLO are physiological, phenological and morphological parameters

that explain the phenotype and are determined experimentally. Genotypic and molecular knowledge could be integrated upstream in SUNFLO and used to estimate the present phenotypic parameters. This implies the identification of the main regulatory genetic pathways that influence the phenotypic parameters already developed in SUNFLO such as, for example, the transpiration rate and the leaf expansion. An allelic combination could replace the phenotypic parameters by taking into account the genotypic effect and the genotype x environment interaction effect of each genetic variant. Therefore the preliminary results obtained in the present work and that aimed at identifying the genetic control of the main regulatory pathways of drought stress responses could be integrated in a crop model such as SUNFLO.

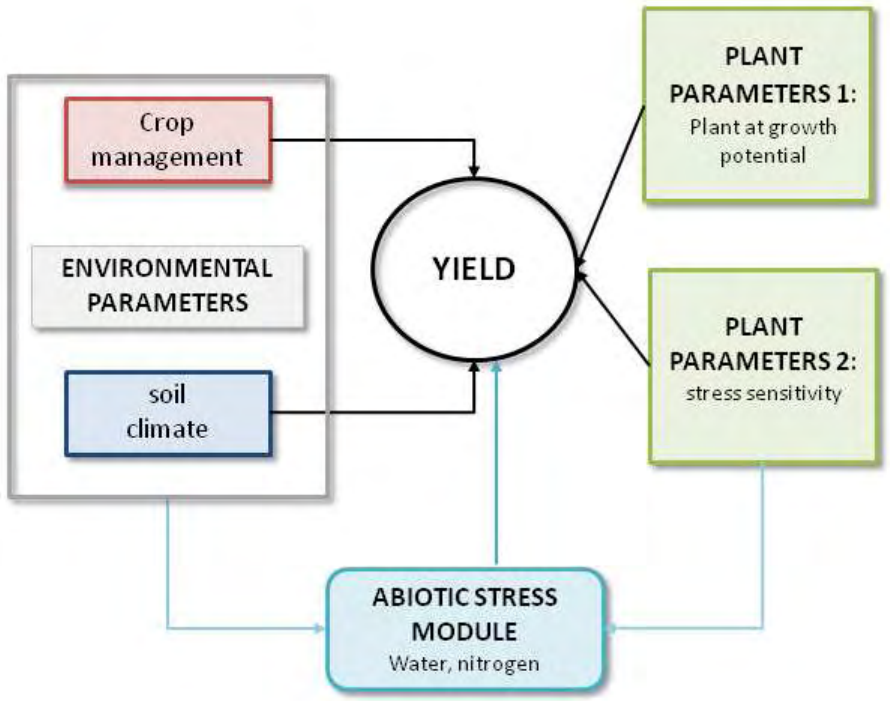


Figure V.2. Functional schema of the crop model SUNFLO.
Environmental and plant parameters relative to stress sensitivity are used in the stress module to adjust the standard yield estimation provided via plant parameters for growth potential (adapted from Casadebaig et al., 2010).

The GRNs that we have been able to reconstruct throughout this PhD work are only partial and the road toward an extensive view and a complete understanding of the sunflower drought GRN is still long. Moreover, once this knowledge obtained, the question of its integration in further research project will remain. However, the integration of the partial results in the SUNFLO crop model could be a first step in the further application of such knowledge.

References

The *C.elegans* Sequencing Consortium 1998. Genome sequence of the nematode C-elegans: A platform for investigating biology. *Science* **282**(5396): 2012-2018.

Aasamaa K, Sober A. 2011. Stomatal sensitivities to changes in leaf water potential, air humidity, CO2 concentration and light intensity, and the effect of abscisic acid on the sensitivities in six temperate deciduous tree species. *Environmental and Experimental Botany* **71**(1): 72-78.

Abe H, Urao T, Ito T, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. 2003. Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* **15**(1): 63-78.

Agrama H, Eizenga G, Yan W. 2007. Association mapping of yield and its components in rice cultivars. *Molecular Breeding* **19**(4): 341-356.

Aguirrezabal LAN, Lavaud Y, Dosio GAA, Izquierdo NG, Andrade FH, Gonzalez LM. 2003. Intercepted solar radiation during seed filling determines sunflower weight per seed and oil concentration. *Crop Science* **43**(1): 152-161.

Ahuja I, de Vos RCH, Bones AM, Hall RD. 2010. Plant molecular stress responses face climate change. *Trends in Plant Science* **15**(12): 664-674.

Alabadi D, Oyama T, Yanovsky M, Harmon F, Mas P, Kay S. 2001. Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science* **293**(5531): 880-883.

Allouche D, Cierco-Ayrolles C, de Givry S, Guillermin G, Mangin B, Schiex T, Vandel J, Vignes M 2013. A panel of learning methods for the reconstruction of gene regulatory networks in a systems genetics context. In: Fuente Adl ed. *Verification of methods for gene network inference from Systems Genetics data*: Springer.

Ameglio T, Archer P, Cohen M, Valancogne C, Daudet F, Dayau S, Cruiziat P. 1999. Significance and limits in the use of predawn leaf water potential for tree irrigation. *Plant and Soil* **207**(2): 155-167.

Anderson R, Tanaka D, Merrill S. 2003. Yield and water use of broadleaf crops in a semiarid climate. *Agricultural Water Management* **58**(3): 255-266.

Araus J, Villegas D, Aparicio N, del Moral L, El Hani S, Rharrabti Y, Ferrio J, Royo C. 2003. Environmental factors determining carbon isotope discrimination and yield in durum wheat under Mediterranean conditions. *Crop Science* **43**(1): 170-180.

Araus JL, Slafer GA, Reynolds MP, Royo C. 2002. Plant breeding and drought in C-3 cereals: What should we breed for? *Annals of Botany* **89**: 925-940.

Asch F, Dingkuhn M, Sow A, Audebert A. 2005. Drought-induced changes in rooting patterns and assimilate partitioning between root and shoot in upland rice. *Field Crops Research* **93**(2-3): 223-236.

Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M. 2008. Computing topological parameters of biological networks. *Bioinformatics* **24**(2): 282-284.

Bach F. 2008. Bolasso: model consistent lasso estimation through the bootstrap. *Proceedings of the 25th international conference on Machine learning*: 33-40.

Bajji M, Kinet JM, Lutts S. 2002. The use of the electrolyte leakage method for assessing cell membrane stability as a water stress tolerance test in durum wheat. *Plant Growth Regulation* **36**(1): 61-70.

Bandillo N, Raghavan C, Muyco PA, Sevilla MA, Lobina IT, Dilla-Ermita CJ, Tung CW, McCouch S, Thomson M, Mauleon R, Singh RK, Gregorio G, Redoña E, Leung H. 2013. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice (N Y)* **6**(1): 11.

Barabasi A, Albert R. 1999. Emergence of scaling in random networks. *Science* **286**(5439): 509-512.

Barabasi AL, Oltvai ZN. 2004. Network biology: Understanding the cell's functional organization.

Nature Reviews Genetics **5**(2): 101-U115.

Bartels D, Sunkar R. 2005. Drought and salt tolerance in plants. *Critical Reviews in Plant Sciences* **24**(1): 23-58.

Bazin J, Langlade N, Vincourt P, Arribat S, Balzergue S, El-Maarouf-Bouteau H, Bailly C. 2011. Targeted mRNA Oxidation Regulates Sunflower Seed Dormancy Alleviation during Dry After-Ripening. *Plant Cell* **23**(6): 2196-2208.

Benjamini Y, 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**(1): 289-300.

Bhatnagar-Mathur P, Vadez V, Sharma K. 2008. Transgenic approaches for abiotic stress tolerance in plants: retrospect and prospects. *Plant Cell Reports* **27**(3): 411-424.

Bing N, Hoeschele I. 2005. Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170**(2): 533-542.

Blackman B, Scascitelli M, Kane N, Luton H, Rasmussen D, Bye R, Lentz D, Rieseberg L. 2011. Sunflower domestication alleles support single domestication center in eastern North America. *Proceedings of the National Academy of Sciences of the United States of America* **108**(34): 14360-14365.

Blum A, Mayer J, Gozlan G. 1982. Infrared thermal sensing of plant canopies as a screening technique for dehydration avoidance in wheat. *Field Crops Research* **5**(2): 137-146.

Bota J, Medrano H, Flexas J. 2004. Is photosynthesis limited by decreased Rubisco activity and RuBP content under progressive water stress? *New Phytologist* **162**(3): 671-681.

Boudsocq M, Lauriere C. 2005. Osmotic signaling in plants. Multiple pathways mediated by emerging kinase families. *Plant Physiology* **138**(3): 1185-1194.

Boursiac Y, Léran S, Corratgé-Faillie C, Gojon A, Krouk G, Lacombe B. 2013. ABA transport and transporters. *Trends in plant science* **18**(6): 1878-4372.

Bradshaw AD. 1965. Evolutionary significance of phenotypic plasticity in plants. *Advances in genetics* **13**(1): 115-155.

Bray E. 2004. Genes commonly regulated by water-deficit stress in Arabidopsis thaliana. *Journal of Experimental Botany* **55**(407): 2331-2341.

Breiman L. 2001. Random forests. *Machine Learning* **45**(1): 5-32.

Brem R, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**(5568): 752-755.

Brennecke P, Anders S, Kim J, Kolodziejczyk A, Zhang X, Proserpio V, Baying B, Benes V, Teichmann S, Marioni J, Heisler M. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10**(11): 1093-1095.

Brugiere N, Jiao SP, Hantke S, Zinselmeier C, Roessler JA, Niu XM, Jones RJ, Habben JE. 2003. Cytokinin oxidase gene expression in maize is localized to the vasculature, and is induced by cytokinins, abscisic acid, and abiotic stress. *Plant Physiology* **132**(3): 1228-1240.

Cadic E, Coque M, Vear F, Grezes-Besset B, Pauquet J, Piquemal J, Lippi Y, Blanchard P, Romestant M, Pouilly N, Rengel D, Gouzy J, Langlade N, Mangin B, Vincourt P. 2013. Combined linkage and association mapping of flowering time in Sunflower (Helianthus annuus L.). *Theoretical and Applied Genetics* **126**(5): 1337-1356.

Casadebaig P, Debaeke P, Lecoeur J. 2008. Thresholds for leaf expansion and transpiration response to soil water deficit in a range of sunflower genotypes. *European Journal of Agronomy* **28**(4): 646-654.

Casadebaig P, Guilioni L, Lecoeur J, Christophe A, Champolivier L, Debaeke P. 2011. SUNFLO, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments. *Agricultural and Forest Meteorology* **151**(2): 163-178.

Cattivelli L, Rizza F, Badeck FW, Mazzucotelli E, Mastrangelo AM, Francia E, Mare C, Tondelli A, Stanca AM. 2008. Drought tolerance improvement in crop plants: An integrated view from breeding to genomics. *Field Crops Research* **105**(1-2): 1-14.

Cellier F, Conejero G, Casse F. 2000. Dehydrin transcript fluctuations during a day/night cycle in

drought-stressed sunflower. *Journal of Experimental Botany* **51**(343): 299-304.

Charbonnier C, Chiquet J, Ambroise C. 2010. Weighted-LASSO for Structured Network Inference from Time Course Data. *Statistical Applications in Genetics and Molecular Biology* **9**(1).

Chater CC, Oliver J, Casson S, Gray JE. 2014. Putting the brakes on: abscisic acid as a central environmental regulator of stomatal development. *New Phytologist*.

Chaves MM, Oliveira MM. 2004. Mechanisms underlying plant resilience to water deficits: prospects for water-saving agriculture. *Journal of Experimental Botany* **55**(407): 2365-2384.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**(7063): 1365-1369.

Chi W, Fung R, Liu H, Hsu C, Charng Y. 2009. Temperature-induced lipocalin is required for basal and acquired thermotolerance in Arabidopsis. *Plant Cell and Environment* **32**(7): 917-927.

Chinnusamy V, Schumaker K, Zhu JK. 2004. Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. *Journal of Experimental Botany* **55**(395): 225-236.

Chiquet J, Grandvalet Y, Ambroise C. 2011. Inferring multiple graphical structures. *Statistics and Computing* **21**(4): 537-553.

Chiquet J, Smith A, Grasseau G, Matias C, Ambroise C. 2009. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics* **25**(3): 417-418.

Choi HI, Hong JH, Ha JO, Kang JY, Kim SY. 2000. ABFs, a family of ABA-responsive element binding factors. *Journal of Biological Chemistry* **275**(3): 1723-1730.

Coca MA, Almoguera C, Jordano J. 1994. EXPRESSION OF SUNFLOWER LOW-MOLECULAR-WEIGHT HEAT-SHOCK PROTEINS DURING EMBRYOGENESIS AND PERSISTENCE AFTER GERMINATION - LOCALIZATION AND POSSIBLE FUNCTIONAL IMPLICATIONS. *Plant Molecular Biology* **25**(3): 479-492.

Cox D, Wermuth N. 1993. Linear dependencies represented by chain graphs. *Statistical Science* **8**(3): 204-218.

Crombach A, Wotton KR, Cicin-Sain D, Ashyraliyev M, Jaeger J. 2012. Efficient Reverse-Engineering of a Developmental Gene Regulatory Network. *Plos Computational Biology* **8**(7): 21.

Cubillos F, Coustham V, Loudet O. 2012a. Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. *Current Opinion in Plant Biology* **15**(2): 192-198.

Cubillos F, Yansouni J, Khalili H, Balzergue S, Elftieh S, Martin-Magniette M, Serrand Y, Lepiniec L, Baud S, Dubreucq B, Renou J, Camilleri C, Loudet O. 2012b. Expression variation in connected recombinant populations of Arabidopsis thaliana highlights distinct transcriptome architectures. *Bmc Genomics* **13**.

Davies WJ, Kudoyarova G, Hartung W. 2005. Long-distance ABA signaling and its relation to other signaling pathways in the detection of soil drying and the mediation of the plant's response to drought. *Journal of Plant Growth Regulation* **24**(4): 285-295.

De Angeli A, Zhang JB, Meyer S, Martinoia E. 2013. AtALMT9 is a malate-activated vacuolar chloride channel required for stomatal opening in Arabidopsis. *Nature Communications* **4**.

de Givry S, Bouchez M, Chabrier P, Milan D, Schiex T. 2005. CAR(H)(T)AGene: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* **21**(8): 1703-1704.

Des Marais DL, Hernandez KM, Juenger TE. 2013. Genotype-by-environment interaction and plasticity: exploring genomic responses of plants to the abiotic environment. *Annual Review of Ecology, Evolution, and Systematics* **44**(1).

Dezar CA, Fedrigo GV, Chan RL. 2005. The promoter of the sunflower HD-Zip protein gene Hahb4 directs tissue-specific expression and is inducible by water stress, high salt concentrations and ABA. *Plant Science* **169**(2): 447-456.

Dikicioglu D, Karabekmez E, Rash B, Pir P, Kirdar B, Oliver SG. 2011. How yeast re-programmes its transcriptional profile in response to different nutrient impulses. *Bmc Systems Biology* **5**.

Dimri GP, Lee XH, Basile G, Acosta M, Scott C, Roskelley C, Medrano EE, Linskens M, Rubelj I,

Pereirasmith O, Peacocke M, Campisi J. 1995. A BIOMARKER THAT IDENTIFIES SENESCENT HUMAN-CELLS IN CULTURE AND IN AGING SKIN IN-VIVO. *Proceedings of the National Academy of Sciences of the United States of America* **92**(20): 9363-9367.

Dixon A, Liang L, Moffatt M, Chen W, Heath S, Wong K, Taylor J, Burnett E, Gut I, Farrall M, Lathrop G, Abecasis G, Cookson W. 2007. A genome-wide association study of global gene expression. *Nature Genetics* **39**(10): 1202-1207.

Dobriyal P, Qureshi A, Badola R, Hussain S. 2012. A review of the methods available for estimating soil moisture and its implications for water resource management. *Journal of Hydrology* **458**: 110-117.

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *Annals of Statistics* **32**(2): 407-451.

El-Soda M, Malosetti M, Zwaan BJ, Koornneef M, Aarts MG. 2014. Genotype× environment interaction QTL mapping in plants: lessons from Arabidopsis. *Trends in Plant Science*.

Endo A, Sawada Y, Takahashi H, Okamoto M, Ikegami K, Koiwai H, Seo M, Toyomasu T, Mitsuhashi W, Shinozaki K, Nakazono M, Kamiya Y, Koshiba T, Nambara E. 2008. Drought induction of Arabidopsis 9-cis-epoxycarotenoid dioxygenase occurs in vascular parenchyma cells. *Plant Physiology* **147**(4): 1984-1993.

Ernst L, Goodger JQD, Alvarez S, Marsh EL, Berla B, Lockhart E, Jung J, Li PH, Bohnert HJ, Schachtman DP. 2010. Sulphate as a xylem-borne chemical signal precedes the expression of ABA biosynthetic genes in maize roots. *Journal of Experimental Botany* **61**(12): 3395-3405.

Ernst W, Peterson P. 1994. The role of biomarkers in environmental assessment .4. Terrestrial plants. *Ecotoxicology* **3**(3): 180-192.

Farooq M, Wahid A, Kobayashi N, Fujita D, Basra S. 2009. Plant drought stress: effects, mechanisms and management. *Agronomy For Sustainable Development* **29**(1): 185-212.

Farquhar G, Ehleringer J, Hubick K. 1989. Carbon isotope discrimination and photosynthesis. *Annual Review of Plant Physiology and Plant Molecular Biology* **40**: 503-537.

Ferreira MI, Valancogne C, Daudet FA, Ameglio T, Pacheco CA, Michaelsen J 1996. Evapotranspiration and crop-water relations in a peach orchard. In: Camp CR, Sadler EJ, Yoder RE eds. *Evapotranspiration and irrigation scheduling. Proceedings of the International Conference, San Antonio, Texas, USA, November 3-6 1996.* St Joseph,: American Society of Agricultural Engineers (ASAE), 61-68.

Flint-Garcia SA, Thornsberry JM, Buckler ES. 2003. Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**: 357-374.

Foyer CH, Fletcher JM. 2001. Plant antioxidants: colour me healthy. *Biologist* **48**(3): 115-120.

Fujita Y, Fujita M, Satoh R, Maruyama K, Parvez MM, Seki M, Hiratsu K, Ohme-Takagi M, Shinozaki K, Yamaguchi-Shinozaki K. 2005. AREB1 is a transcription activator of novel ABRE-dependent ABA signaling that enhances drought stress tolerance in Arabidopsis. *Plant Cell* **17**(12): 3470-3488.

Furihata T, Maruyama K, Fujita Y, Umezawa T, Yoshida R, Shinozaki K, Yamaguchi-Shinozaki K. 2006. Abscisic acid-dependent multisite phosphorylation regulates the activity of a transcription activator AREB1. *Proceedings of the National Academy of Sciences of the United States of America* **103**(6): 1988-1993.

Fusari CM, Di Rienzo JA, Troglia C, Nishinakamasu V, Moreno MV, Maringolo C, Quiroz F, Alvarez D, Escande A, Hopp E, Heinz R, Lia VV, Paniego NB. 2012. Association mapping in sunflower for sclerotinia head rot resistance. *Bmc Plant Biology* **12**.

Fusari CM, Lia VV, Hopp HE, Heinz RA, Paniego NB. 2008. Identification of Single Nucleotide Polymorphisms and analysis of Linkage Disequilibrium in sunflower elite inbred lines using the candidate gene approach. *Bmc Plant Biology* **8**.

Gan XC, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, Clark RM, Ratsch G, Mott R. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**(7365): 419-423.

Gardner CMK, Bell JP, Cooper JD, Dean TJ, Gardner N, Hodnett MG. 1991. Soil water content. *Soil analysis: physical methods.*: 1-73.

Garg BK, Burman U, Kathju S. 2004. The influence of phosphorus nutrition on the physiological response of moth bean genotypes to drought. *Journal of Plant Nutrition and Soil Science-Zeitschrift Fur Pflanzenernahrung Und Bodenkunde* **167**(4): 503-508.

Gavrilets S, Scheiner S. 1993. The genetics of phenotypic plasticity .5. Evolution of reaction norm shape. *Journal of Evolutionary Biology* **6**(1): 31-48.

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li WQ, Ogawa M, Yamauchi Y, Preston J, Aoki K, Kiba T, Takatsuto S, Fujioka S, Asami T, Nakano T, Kato H, Mizuno T, Sakakibara H, Yamaguchi S, Nambara E, Kamiya Y, Takahashi H, Hirai MY, Sakurai T, Shinozaki K, Saito K, Yoshida S, Shimada Y. 2008. The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. *Plant Journal* **55**(3): 526-542.

Goldhamer D, Fereres E. 2001. Irrigation scheduling protocols using continuously recorded trunk diameter measurements. *Irrigation Science* **20**(3): 115-125.

González I, Lê Cao KA, Déjean S 2011. MixOmics: Omics data integration project. URL http://www. math. univ-toulouse. fr/~ biostat/mixOmics.

Gorantla M, Babu PR, Lachagari VBR, Reddy AMM, Wusirika R, Bennetzen JL, Reddy AR. 2007. Identification of stress-responsive genes in an indica rice (Oryza sativa L.) using ESTs generated from drought-stressed seedlings. *Journal of Experimental Botany* **58**(2): 253-265.

Goudet J. 2005. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**(1): 184-186.

Granier C, Tardieu F. 1999. Water deficit and spatial pattern of leaf development. Variability in responses can be simulated using a simple model of leaf development. *Plant Physiology* **119**(2): 609-619.

Grishkevich V, Yanai I. 2013. The genomic determinants of genotype x environment interactions in gene expression. *Trends in Genetics* **29**(8): 479-487.

Guo FO, Young J, Crawford NM. 2003. The nitrate transporter AtNRT1.1 (CHL1) functions in stomatal opening and contributes to drought susceptibility in arabidopsis. *Plant Cell* **15**(1): 107-117.

Hall AJ, Connor DJ, Whitfield DM. 1990. ROOT RESPIRATION DURING GRAIN FILLING IN SUNFLOWER - THE EFFECTS OF WATER-STRESS. *Plant and Soil* **121**(1): 57-66.

Hannah MA, Wiese D, Freund S, Fiehn O, Heyer AG, Hincha DK. 2006. Natural genetic variation of freezing tolerance in arabidopsis. *Plant Physiology* **142**(1): 98-112.

Harb A, Krishnan A, Ambavaram M, Pereira A. 2010. Molecular and Physiological Analysis of Drought Stress in Arabidopsis Reveals Early Responses Leading to Acclimation in Plant Growth. *Plant Physiology* **154**(3): 1254-1271.

Hasegawa PM, Bressan RA, Zhu JK, Bohnert HJ. 2000. Plant cellular and molecular responses to high salinity. *Annual Review of Plant Physiology and Plant Molecular Biology* **51**: 463-499.

Herve D, Fabre F, Berrios EF, Leroux N, Al Chaarani G, Planchon C, Sarrafi A, Gentzbittel L. 2001. QTL analysis of photosynthesis and water status traits in sunflower (Helianthus annuus L.) under greenhouse conditions. *Journal of Experimental Botany* **52**(362): 1857-1864.

Hinman VF, Nguyen AT, Cameron RA, Davidson EH. 2003. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proceedings of the National Academy of Sciences of the United States of America* **100**(23): 13356-13361.

Hirayama T, Shinozaki K. 2010. Research on plant abiotic stress responses in the post-genome era: past, present and future. *Plant Journal* **61**(6): 1041-1052.

Hoekstra FA, Golovina EA, Buitink J. 2001. Mechanisms of plant desiccation tolerance. *Trends in Plant Science* **6**(9): 431-438.

Holloway B, Luck S, Beatty M, Rafalski JA, Li B. 2011. Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *Bmc Genomics* **12**.

Honsdorf N, Becker H, Ecke W. 2010. Association mapping for phenological, morphological, and quality traits in canola quality winter rapeseed (Brassica napus L.). *Genome* **53**(11): 899-907.

Huang G-T, Ma S-L, Bai L-P, Zhang L, Ma H, Jia P, Liu J, Zhong M, Guo Z-F. 2012. Signal transduction during cold, salt, and drought stresses in plants. *Molecular Biology Reports* **39**(2): 969-987.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *Plos One* **5**(9).

Ingram J, Bartels D. 1996. The molecular basis of dehydration tolerance in plants. *Annual Review of Plant Physiology and Plant Molecular Biology* **47**: 377-403.

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y. 2000. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences of the United States of America* **97**(3): 1143-1147.

Jansen R, Nap J. 2001. Genetical genomics: the added value from segregation. *Trends in Genetics* **17**(7): 388-391.

Jiménez-Gómez JM. 2011. Next generation quantitative genetics in plants. *Front Plant Sci* **2**: 77.

Jossier M, Kroniewicz L, Dalmas F, Le Thiec D, Ephritikhine G, Thomine S, Barbier-Brygoo H, Vavasseur A, Filleur S, Leonhardt N. 2010. The Arabidopsis vacuolar anion transporter, AtCLCc, is involved in the regulation of stomatal movements and contributes to salt tolerance. *Plant Journal* **64**(4): 563-576.

Jung M, Ching A, Bhattramakki D, Dolan M, Tingey S, Morgante M, Rafalski A. 2004. Linkage disequilibrium and sequence diversity in a 500-kbp region around the adh1 locus in elite maize germplasm. *Theoretical and Applied Genetics* **109**(4): 681-689.

Kane NC, Burke JM, Marek L, Seiler G, Vear F, Baute G, Knapp SJ, Vincourt P, Rieseberg LH. 2013. Sunflower genetic, genomic and ecological resources. *Molecular Ecology Resources* **13**(1): 10-20.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* **178**(3): 1709-1723.

Kawaguchi R, Girke T, Bray E, Bailey-Serres J. 2004. Differential mRNA translation contributes to gene regulation under non-stress and dehydration stress conditions in Arabidopsis thaliana. *Plant Journal* **38**(5): 823-839.

Kazan K, Manners JM. 2013. MYC2: the master in action. *Molecular plant* **6**(3): 686-703.

Keurentjes J, Fu J, Terpstra I, Garcia J, van den Ackerveken G, Snoek L, Peeters A, Vreugdenhil D, Koornneef M, Jansen R. 2007. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **104**(5): 1708-1713.

Kiani SP, Grieu P, Maury P, Hewezi T, Gentzbittel L, Sarrafi A. 2007a. Genetic variability for physiological traits under drought conditions and differential expression of water stress-associated genes in sunflower (Helianthus annuus L.). *Theoretical and Applied Genetics* **114**(2): 193-207.

Kiani SP, Talia P, Maury P, Grieu P, Heinz R, Perrault A, Nishinakamasu V, Hopp E, Gentzbittel L, Paniego N, Sarrafi A. 2007b. Genetic analysis of plant water status and osmotic adjustment in recombinant inbred lines of sunflower under two water treatments. *Plant Science* **172**(4): 773-787.

Kim S, Choi HI, Ryu HJ, Park JH, Kim MD, Kim SY. 2004. ARIA, an Arabidopsis arm repeat protein interacting with a transcriptional regulator of abscisic acid-responsive gene expression, is a novel abscisic acid signaling component. *Plant Physiology* **136**(3): 3639-3648.

Kinman ML 1970. New developments in the USDA and state experiment station sunflower breeding programs. *4th International Sunflower Conference*. Memphis, USA. 181-183.

Kitano H. 2000. Perspectives on systems biology. *New Generation Computing* **18**(3): 199-216.

Kitano H. 2002. Systems biology: A brief overview. *Science* **295**(5560): 1662-1664.

Klenke J, Flint A. 1991. Collimated neutron probe for soil-water content measurements. *Soil Science Society of America Journal* **55**(4): 916-923.

Kobayashi Y, Murata M, Minami H, Yamamoto S, Kagaya Y, Hobo T, Yamamoto A, Hattori T. 2005.

Abscisic acid-activated SNRK2 protein kinases function in the gene-regulation pathway of ABA signal transduction by phosphorylating ABA response element-binding factors. *Plant Journal* **44**(6): 939-949.

Kolkman JM, Berry ST, Leon AJ, Slabaugh MB, Tang S, Gao WX, Shintani DK, Burke JM, Knapp SJ. 2007. Single nucleotide polymorphisms and linkage disequilibrium in sunflower. *Genetics* **177**(1): 457-468.

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R. 2009. A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genet* **5**(7): e1000551.

Lacaze X, Hayes P, Korol A. 2009. Genetics of phenotypic plasticity: QTL analysis in barley, Hordeum vulgare. *Heredity* **102**(2): 163-173.

Lander E, Botstein D. 1989. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics* **121**(1): 185-199.

Leclercq P. 1969. Une sterilite male cytoplasmique chez le tournesol. *Ann. Amelior. Plant* **19**(2): 99-106.

Leport L, Turner NC, Davies SL, Siddique KHM. 2006. Variation in pod production and abortion among chickpea cultivars under terminal drought. *European Journal of Agronomy* **24**(3): 236-246.

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

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

Li S, Assmann S, Albert R. 2006. Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *Plos Biology* **4**(10): 1732-1748.

Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JAG, Hazendonk E, Prins P, Plasterk RHA, Jansen RC, Breitling R, Kammenga JE. 2006. Mapping determinants of gene expression plasticity by genetical genomics in C-elegans. *Plos Genetics* **2**(12): 2155-2161.

Liu Q, Kasuga M, Sakuma Y, Abe H, Miura S, Yamaguchi-Shinozaki K, Shinozaki K. 1998. Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in Arabidopsis. *Plant Cell* **10**(8): 1391-1406.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. *Methods* **25**(4): 402-408.

Loreto F, Tricoli D, DiMarco G. 1995. On the relationship between electron transport rate and photosynthesis in leaves of the C-4 plant Sorghum bicolor exposed to water stress, temperature changes and carbon metabolism inhibition. *Australian Journal of Plant Physiology* **22**(6): 885-892.

Lu B, Gong Z, Wang J, Zhang J, Liang J. 2007. Microtubule dynamics in relation to osmotic stress-induced ABA accumulation in Zea mays roots. *Journal of Experimental Botany* **58**(10): 2565-2572.

Lu G, Paul A, McCarty D, Ferl R. 1996. Transcription factor veracity: Is GBF3 responsible for ABA-regulated expression of Arabidopsis Adh? *Plant Cell* **8**(5): 847-857.

Ma SS, Gong QQ, Bohnert HJ. 2007. An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research* **17**(11): 1614-1625.

Mackay I, Powell W. 2007. Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* **12**(2): 57-63.

Mackay T. 2001. The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**: 303-339.

Mackay T, Stone E, Ayroles J. 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**(8): 565-577.

Manavella PA, Arce AL, Dezar CA, Bitton F, Renou JP, Crespi M, Chan RL. 2006. Cross-talk between

ethylene and drought signalling pathways is mediated by the sunflower Hahb-4 transcription factor. *Plant Journal* **48**(1): 125-137.

Mandel JR, Nambeesan S, Bowers JE, Marek LF, Ebert D, Rieseberg LH, Knapp SJ, Burke JM. 2013. Association Mapping and the Genomic Consequences of Selection in Sunflower. *Plos Genetics* **9**(3).

Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C. 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* **108**(3): 285-291.

Mansfield TA, Hetherington AM, Atkinson CJ. 1990. SOME CURRENT ASPECTS OF STOMATAL PHYSIOLOGY. *Annual Review of Plant Physiology and Plant Molecular Biology* **41**: 55-75.

Marbach D, Costello J, Kuffner R, Vega N, Prill R, Camacho D, Allison K, Kellis M, Collins J, Stolovitzky G, Consortium D. 2012. Wisdom of crowds for robust gene network inference. *Nature Methods* **9**(8): 796-+.

Marchand G, Mayjonade B, Vares D, Blanchet N, Boniface M-C, Maury P, Nambinina FA, Burger P, Debaeke P, Casadebaig P, Vincourt P, Langlade NB. 2013. A biomarker based on gene expression indicates plant water status in controlled and natural environments. *Plant Cell and Environment* **36**: 2175-2189.

Marcus A, Moore R, Cyr R. 2001. The role of microtubules in guard cell function. *Plant Physiology* **125**(1): 387-395.

Maurel C, Chrispeels MJ. 2001. Aquaporins. A molecular entry into plant water relations. *Plant Physiology* **125**(1): 135-138.

Mendes P, Sha W, Ye K. 2003. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* **19**: II122-II129.

Middleton A, Farcot E, Owen M, Vernoux T. 2012. Modeling Regulatory Networks to Understand Plant Development: Small Is Beautiful. *Plant Cell* **24**(10): 3876-3891.

Miller G, Suzuki N, Ciftci-Yilmaz S, Mittler R. 2009. oxygen species homeostasis and signalling during drought and sallinity stresses. *Plant, Cell and Environment* **33**(4): 453-467.

Milo R, Hou JH, Springer M, Brenner MP, Kirschner MW. 2007. The relationship between evolutionary and physiological variation in hemoglobin. *Proceedings of the National Academy of Sciences of the United States of America* **104**(43): 16998-17003.

Miyashita K, Tanakamaru S, Maitani T, Kimura K. 2005. Recovery responses of photosynthesis, transpiration, and stomatal conductance in kidney bean following drought stress. *Environmental and Experimental Botany* **53**(2): 205-214.

Monteith JL 1965. Evaporation and environment. 4.

Morgante M, Salamini F. 2003. From plant genomics to breeding practice. *Curr Opin Biotechnol* **14**(2): 214-219.

Moriondo M, Bindi M, Kundzewicz Z, Szwed M, Chorynski A, Matczak P, Radziejewski M, McEvoy D, Wreford A. 2010. Impact and adaptation opportunities for European agriculture in response to climatic change and variability. *Mitigation and Adaptation Strategies For Global Change* **15**(7): 657-679.

Moriondo M, Giannakopoulos C, Bindi M. 2011. Climate change impact assessment: the role of climate extremes in crop yield simulation. *Climatic Change* **104**(3-4): 679-701.

Mummey DL, Stahl PD, Buyer JS. 2002. Microbial biomarkers as an indicator of ecosystem recovery following surface mine reclamation. *Applied Soil Ecology* **21**(3): 251-259.

Neumann G, Massonneau A, Langlade N, Dinkelaker B, Hengeler C, Romheld V, Martinoia E. 2000. Physiological aspects of cluster root function and development in phosphorus-deficient white lupin (Lupinus albus L.). *Annals of Botany* **85**(6): 909-919.

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D. 2002. The extent of linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics* **30**(2): 190-193.

Nylander M, Svensson J, Palva ET, Welin BV. 2001. Stress-induced accumulation and tissue-specific localization of dehydrins in Arabidopsis thaliana. *Plant Molecular Biology* **45**(3): 263-279.

Osborn HF. 1896. A mode of evolution requiring neither natural selection nor the inheritance of acquired characters. *Transactions of the New York academy of sciences* **15**: 141-142.

O'Toole J, Cruz R, Singh T. 1979. Leaf rolling and transpiration. *Plant Science Letters* **16**(1): 111-114.

Pastori G, Foyer C. 2002. Common components, networks, and pathways of cross-tolerance to stress. The central role of "redox" and abscisic acid-mediated controls. *Plant Physiology* **129**(2): 460-468.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *Plos Genetics* **2**(12): 2074-2093.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**(2): 945-959.

Pustovoit VS. 1964. Conclusions of work on the selection and seed production of sunflowers. *Agrobiology* **5**: 672-697.

Queitsch C, Sangster TA, Lindquist S. 2002. Hsp90 as a capacitor of phenotypic variation. *Nature* **417**(6889): 618-624.

Rafalski JA. 2002. Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science* **162**(3): 329-333.

Ramanjulu S, Bartels D. 2002. Drought- and desiccation-induced modulation of gene expression in plants. *Plant Cell and Environment* **25**(2): 141-151.

Ramsay H, Rieseberg LH, Ritland K. 2009. The Correlation of Evolutionary Rate with Pathway Position in Plant Terpenoid Biosynthesis. *Molecular Biology and Evolution* **26**(5): 1045-1053.

Rausher M, Miller R, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Molecular Biology and Evolution* **16**(2): 266-274.

Remington D, Thornsberry J, Matsuoka Y, Wilson L, Whitt S, Doeblay J, Kresovich S, Goodman M, Buckler E. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America* **98**(20): 11479-11484.

Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, Bowers JE, Burke JM, Rieseberg LH. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature communications* **4**: 1827-1827.

Rengel D, Arribat S, Maury P, Martin-Magniette M, Hourlier T, Laporte M, Vares D, Carrere S, Grieu P, Balzergue S, Gouzy J, Vincourt P, Langlade N. 2012. A Gene-Phenotype Network Based on Genetic Variability for Drought Responses Reveals Key Physiological Processes in Controlled and Natural Environments. *Plos One* **7**(10).

Rieseberg LH, Whitton J, Gardner K. 1999. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* **152**(2): 713-727.

Rock CD, Zeevaart JAD. 1991. THE ABA MUTANT OF ARABIDOPSIS-THALIANA IS IMPAIRED IN EPOXY-CAROTENOID BIOSYNTHESIS. *Proceedings of the National Academy of Sciences of the United States of America* **88**(17): 7496-7499.

Rodriguez PL. 1998. Protein phosphatase 2C (PP2C) function in higher plants. *Plant Molecular Biology* **38**(6): 919-927.

Rook F, Hadingham S, Li Y, Bevan M. 2006. Sugar and ABA response pathways and the control of gene expression. *Plant Cell and Environment* **29**(3): 426-434.

Schadt E, Molony C, Chudin E, Hao K, Yang X, Lum P, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey J, Avila-Campillo I, Kruger M, Johnson J, Rohl C, van Nas A, Mehrabian M, Drake T, Lusis A, Smith R, Guengerich F, Strom S, Schuetz E, Rushmore T, Ulrich R. 2008. Mapping the genetic architecture of gene expression in human liver. *Plos Biology* **6**(5): 1020-1032.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**(6929): 297-302.

Scheiner S. 1993. Genetics and evolution of phenotypic plasticity. *Annual Review of Ecology and Systematics* **24**: 35-68.

Schmidhuber J, Tubiello F. 2007. Global food security under climate change. *Proceedings of the National Academy of Sciences of the United States of America* **104**(50): 19703-19708.

Seassau C, Dechamp-Guillaume G, Mestries E, Debaeke P. 2010. Nitrogen and water management can limit premature ripening of sunflower induced by Phoma macdonaldii. *Field Crops Research* **115**(1): 99-106.

Seiler, Rieseberg 1997. Systematics, origin, and germplasm resources of the wild and domesticated sunflower. In: Schneiter ed. *Sunflower Science and Technology*. Madison, Wisconsin: ASA, CSSA, and ASSA, 21-66.

Seki M, Umezawa T, Urano K, Shinozaki K. 2007. Regulatory metabolic networks in drought stress responses. *Current Opinion in Plant Biology* **10**(3): 296-302.

Shaffer JP. 1995. MULTIPLE HYPOTHESIS-TESTING. *Annual Review of Psychology* **46**: 561-584.

Shinozaki K, Yamaguchi-Shinozaki K. 2007. Gene networks involved in drought stress response and tolerance. *Journal of Experimental Botany* **58**(2): 221-227.

Shinozaki K, Yamaguchi-Shinozaki K, Seki M. 2003. Regulatory network of gene expression in the drought and cold stress responses. *Current Opinion in Plant Biology* **6**(5): 410-417.

Shinozaki K, Yamaguchi-Shinozaki K. 1997. Gene expression and signal transduction in water-stress response. *Plant Physiology* **115**(2): 327-334.

Simpson SD, Nakashima K, Narusaka Y, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. 2003. Two different novel cis-acting elements of erd1, a clpA homologous Arabidopsis gene function in induction by dehydration stress and dark-induced senescence. *Plant Journal* **33**(2): 259-270.

Sinclair T. 1986. Water and nitrogen limitations in soybean grain production .1. Model development. *Field Crops Research* **15**(2): 125-141.

Sinclair T. 2005. Theoretical analysis of soil and plant traits influencing daily plant water flux on drying soils. *Agronomy Journal* **97**(4): 1148-1152.

Sinclair T, Seligman N. 2000. Criteria for publishing papers on crop modeling. *Field Crops Research* **68**(3): 165-172.

Singh R, Bhat K, Bhatia V, Mohapatra T, Singh N. 2008. Association mapping for photoperiod insensitivity trait in soybean. *National Academy Science Letters-India* **31**(9-10): 281-283.

Sirichandra C, Wasilewska A, Vlad F, Valon C, Leung J. 2009. The guard cell as a single-cell model towards understanding drought tolerance and abscisic acid action. *Journal of Experimental Botany* **60**(5): 1439-1463.

Smith EN, Kruglyak L. 2008. Gene-environment interaction in yeast gene expression. *Plos Biology* **6**(4): 810-824.

Somerville C, Briscoe L. 2001. Genetic engineering and water. *Science* **292**(5525): 2217-2217.

Spurgeon SL, Jones RC, Ramakrishnan R. 2008. High Throughput Gene Expression Measurement with Real Time PCR in a Microfluidic Dynamic Array. *Plos One* **3**(2).

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET. 2005. Genome-wide associations of gene expression variation in humans. *Plos Genetics* **1**(6): 695-704.

Strimbu K, Tavel JA. 2010. What are biomarkers? *Curr Opin HIV AIDS* **5**(6): 463-466.

Stuessy T. 2010. The rise of sunflowers. *Science* **329**(5999): 1605-1606.

Swanson-Wagner RA, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, Nettleton D, Schnable PS. 2009. Paternal Dominance of Trans-eQTL Influences Gene Expression Patterns in Maize Hybrids. *Science* **326**(5956): 1118-1120.

Taiz L, Zeiger E. 2006. *Plant Physiology*. Massachusetts: Sinauer Associates Inc. Publishers.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch B, Siddiqui A, Lao K, Surani M. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**(5): 377-U386.

Tardieu F. 2013. Plant response to environmental conditions: assessing potential production, water demand, and negative effects of water deficit. *Frontiers in physiology* **4**.

Tardieu F, Granier C, Muller B. 2011. Water deficit and growth. Co-ordinating processes without an orchestrator? *Current Opinion in Plant Biology* **14**(3): 283-289.

Tardieu F, Tuberosa R. 2010. Dissection and modelling of abiotic stress tolerance in plants. *Current Opinion in Plant Biology* **13**(2): 206-212.

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp mays L.). *Proceedings of the National Academy of Sciences of the United States of America* **98**(16): 9161-9166.

Thompson D, Wilkinson S, Bacon M, William J. 1997. Multiple signals and mechanisms that regulate leaf growth and stomatal behaviour during water deficit. *Physiologia Plantarum* **100**(2): 303-313.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES. 2001. Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* **28**(3): 286-289.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* **58**(1): 267-288.

Timme RE, Simpson BB, Linder CR. 2007. High-resolution phylogeny for Helianthus (Asteraceae) using the 18s-26s ribosomal DNA external transcribed spacer. *American Journal of Botany* **94**(11): 1837-1852.

Todaka D, Nakashima K, Shinozaki K, Yamaguchi-Shinozaki K. 2012. Toward understanding transcriptional regulatory networks in abiotic stress responses and tolerance in rice. *Rice* **5**.

Topp G, Davis J. 1985. Measurement of soil-water content using time-domain reflectrometry (TDR) - a field-evaluation. *Soil Science Society of America Journal* **49**(1): 19-24.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**(6): 520-525.

Turner NC, Wright GC, Siddique KHM. 2001. Adaptation of grain legumes (pulses) to water-limited environments. *Advances in Agronomy* **71**: 194-233.

Umezawa T, Nakashima K, Miyakawa T, Kuromori T, Tanokura M, Shinozaki K, Yamaguchi-Shinozaki K. 2010. Molecular Basis of the Core Regulatory Network in ABA Responses: Sensing, Signaling and Transport. *Plant and Cell Physiology* **51**(11): 1821-1839.

Uno Y, Furihata T, Abe H, Yoshida R, Shinozaki K, Yamaguchi-Shinozaki K. 2000. Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proceedings of the National Academy of Sciences of the United States of America* **97**(21): 11632-11637.

Valliyodan B, Nguyen H. 2006. Understanding regulatory networks and engineering for enhanced drought tolerance in plants. *Current Opinion in Plant Biology* **9**(2): 189-195.

Van Sandt V, Suslov D, Verbelen J, Vissenberg K. 2007. Xyloglucan endotransglucosylase activity loosens a plant cell wall. *Annals of Botany* **100**(7): 1467-1473.

Vear F, Bony H, Joubert G, de Labrouhe D, Pauchet I, Pinochet X. 2003. 30 years of sunflower breeding in France. *Ocl-Oleagineux Corps Gras Lipides* **10**(1): 66-73.

Verzelen N. 2012. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics* **6**: 38-90.

Vignes M, Vandel J, Allouche D, Ramadan-Alban N, Cierco-Ayrolles C, Schiex T, Mangin B, de Givry S. 2011. Gene Regulatory Network Reconstruction Using Bayesian Networks, the Dantzig Selector, the Lasso and Their Meta-Analysis. *Plos One* **6**(12).

Waddington CH. 1942. Canalization of development and the inheritance of acquired characters. *Nature* **150**(3811): 563-565.

Wang J, Yu H, Xie W, Xing Y, Yu S, Xu C, Li X, Xiao J, Zhang Q. 2010. A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. *Plant Journal* **63**(6): 1063-1074.

Wang W, Vinocur B, Altman A. 2003. Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. *Planta* **218**(1): 1-14.

Wasilewska A, Vlad F, Sirichandra C, Redko Y, Jammes F, Valon C, Frey N, Leung J. 2008. An update on abscisic acid signaling in plants and more ... *Molecular Plant* **1**(2): 198-217.

Wege S, Jossier M, Filleur S, Thomine S, Barbier-Brygoo H, Gambale F, De Angeli A. 2010. The proline 160 in the selectivity filter of the Arabidopsis NO3-/H+ exchanger AtCLCa is essential

for nitrate accumulation in planta. *Plant Journal* **63**(5): 861-869.

Weigel D, Mott R. 2009. The 1001 Genomes Project for Arabidopsis thaliana. *Genome Biology* **10**(5).

West M, Kim K, Kliebenstein D, van Leeuwen H, Michelmore R, Doerge R, Clair D. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* **175**(3): 1441-1450.

Whalley W, Dean T, Izzard P. 1992. Evaluation of the capacitance technique as a method for dynamically measuring soil-water content. *Journal of Agricultural Engineering Research* **52**(2): 147-155.

Whittaker. 1990. *Graphical Models in Applied Multivariate Statistics*. NY: Wiley.

Wilkins AS. 2005. Recasting developmental evolution in terms of genetic pathway and network evolution ... and the implications for comparative biology. *Brain Research Bulletin* **66**(4-6): 495-509.

Wilkins O, Brautigam K, Campbell M. 2010. Time of day shapes Arabidopsis drought transcriptomes. *Plant Journal* **63**(5): 715-727.

Wilkinson S, Kudoyarova GR, Veselov DS, Arkhipova TN, Davies WJ. 2012. Plant hormone interactions: innovative targets for crop breeding and management. *Journal of Experimental Botany* **63**(9): 3499-3509.

Yamaguchi-Shinozaki K, Shinozaki K. 2006. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annual Review of Plant Biology* **57**: 781-803.

Yang X, Wu J, Ziegler T, Yang X, Zayed A, Rajani M, Zhou D, Basra A, Schachtman D, Peng M, Armstrong C, Caldo R, Morrell J, Lacy M, Staub J. 2011. Gene Expression Biomarkers Provide Sensitive Indicators of in Planta Nitrogen Status in Maize. *Plant Physiology* **157**(4): 1841-1852.

Yu JM, Buckler ES. 2006. Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology* **17**(2): 155-160.

Yu JM, Holland JB, McMullen MD, Buckler ES. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**(1): 539-551.

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**(2): 203-208.

Yuan J, Galbraith D, Dai S, Griffin P, Stewart C. 2008. Plant systems biology comes of age. *Trends in Plant Science* **13**(4): 165-171.

Zhiponova M, Vanhoutte I, Boudolf V, Betti C, Dhondt S, Coppens F, Mylle E, Maes S, Gonzalez-Garcia M, Cano-Delgado A, Inze D, Beemster G, De Veylder L, Russinova E. 2013. Brassinosteroid production and signaling differentially control cell division and expansion in the leaf. *New Phytologist* **197**(2): 490-502.

Zhu CS, Gore M, Buckler ES, Yu JM. 2008. Status and Prospects of Association Mapping in Plants. *Plant Genome* **1**(1): 5-20.

Zhu M, Dai S, Chen S. 2012. The stomata frontline of plant interaction with the environment-perspectives from hormone regulation. *Frontiers in Biology* ***7***(2): 96-112.

Zimmermann P, Laule O, Schmitz J, Hruz T, Bleuler S, Gruissem W. 2008. Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases. *Molecular Plant* **1**(5): 851-857.

Appendices

Appendices Chapter II

Appendix II.1: Model for soil evaporation

$E_{pd}=aP_{pd}+bH_d+cT_d+d$

Where

E_{pd} is the soil evaporation for the pot *p* at the day *d*,

P_{pd} is the wheight of the pot *p* at the day *d*

H_d is the average humidity in glasshouse at the day *d*

T_d is the average temperature in glasshouse at the day *d*

a, b, c and d are constants with the following values:

a=19.86902302

b=-3.425441698

c=-5.025490995

d=126.0254727

Appendix II.2: List of selected genes for the WSB construction and their functional annotations.

This appendix is available on line at the address:

https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents the genes selected from Rengel *et al.* 2012 for the WSB construction, circadian clock genes, sunflower dehydrins and reference genes (Excel file).

Appendix II.3: List of primers for candidate genes of the WSB construction.

This appendix is available on line at the address:

https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents the primers for the candidate genes for WSB construction (Excel file).

Appendix II.4: Supporting Materials and Methods of article describing the WSB construction

Transcriptomic analysis

Tissue harvest and RNA extraction

Leaves were harvested between 11:00 and 13:00, except during the diurnal variation study To harvest non-senescent and non-growing leaves, the total leaf number (N_{tot}) was estimated, and two thirds of the total leaves from the bottom were tagged and termed the n^{th} leaf. In the greenhouse experiment, we selected the n+1 leaf, and in the field experiment, we selected the N_{tot}-4 leaf of one plant. Each leaf was sampled and treated separately.

The leaves were cut without their petiole and immediately frozen in liquid nitrogen (greenhouse) or in dry ice (field). Grinding was performed using a ZM200 grinder (Retsch, Haan, Germany) with a 0.5-mm sieve. Total RNA was extracted using QIAzol Lysis Reagent following the manufacturer's instructions (Qiagen, Dusseldorf, Germany). The quantity of RNA was estimated using a ND-1000 spectrophotometer (Nanodrop, Wilmington, DE, USA). The RNA quality was checked by electrophoresis on an agarose gel. The cDNA was synthesized from 2 µg of total RNA using an anchored oligo dT (dT15-V) and the Transcriptor First Strand cDNA Synthesis Kit (Roche, Basel, Switzerland).

Estimation of gene expression by qRT-PCR

Gene expression was estimated using the BioMark™ HD System (Fluidigm, San Franscisco CA, USA) with a 96.96 Dynamic Array IFC and EvaGreen® (Bio-Rad, Hercules CA, USA) as the DNA binding dye (Spurgeon et al., 2008).

Primers and cDNA samples were adjusted to a concentration of 20 µM and 5 ng/µl, respectively.

A specific target amplification (STA) was performed for each sample using TaqMan® PreAmp Master Mix (Applied Biosystems/Life Technologies, PN 4361128, Carlsbad CA, USA). The reaction mixture and thermal cycling were performed according to the Fluidigm protocol PN100-1208 B1.

After the STA step, an exonuclease treatment (M0293S, New England Biolabs, Ipswich MA, USA) was performed following the manufacturer's instructions to remove any primers still present in the reaction mixture.

Finally, all samples were diluted 1:5 in water (taking into account the dilution of the exonuclease treatment). Further steps concerning the loading chip and thermal cycling were performed following the Fluidigm Protocol PN100-1208 B1.

The expression levels of gene *i* expressed as the cycle threshold Ct were normalized according to the

amplification efficiency (noted eff_i below) and the expression levels of seven reference genes (noted *r* below) identified in (Rengel et al., 2012) were estimated as follows:

$dCt_i = ((1+eff_i)^{Cti}) / mean((1+eff_r)^{Ctr})$.

Statistical analysis

Statistical construction of the water status biomarker (WSB)

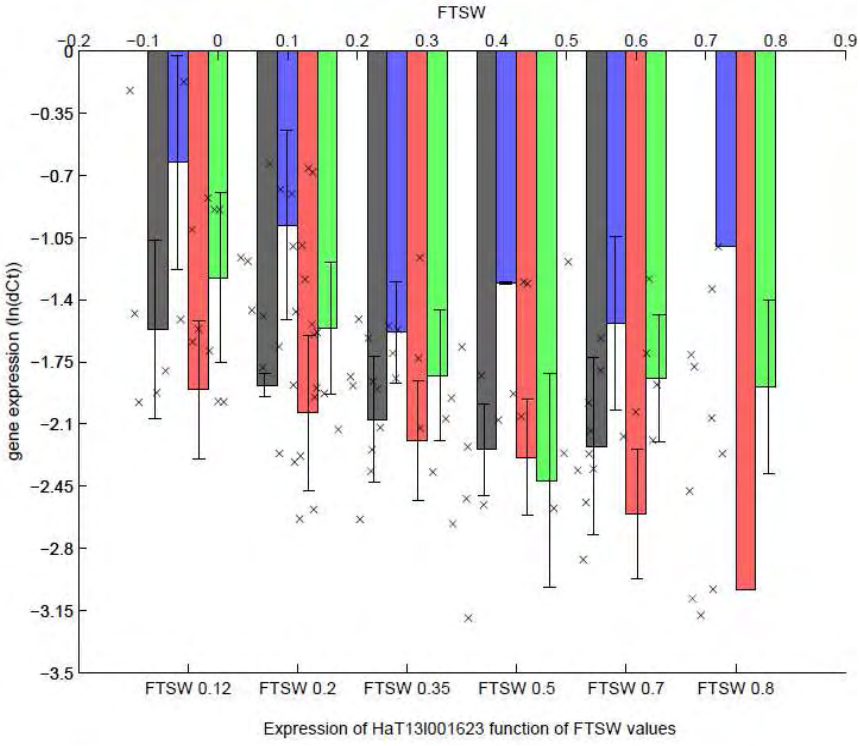
The pre-dawn leaf water potentials were not measured at the same time in field and in greenhouse experiments. Ψ_{PD}' in the greenhouse experiment measured the water status at the time of the transcriptomic harvest and Ψ_{PD} in field measured the water status at the end of the day before the transcriptomic harvest. These two measurements allowed us to access to water available for the plant at two different time of the experiment (Figure II.2). The WSB was calibrated with greenhouse data to estimate Ψ_{PD}'. Using Ψ_{PD} in the field experiments, that did not account for WS variation in the morning between the WS measurement and the transcriptomic harvest, and introduced a bias for field validation. To correct it and compare equivalent WS in greenhouse and in field experiments, we estimated Ψ_{PD}' from the pre-dawn leaf water potential measured in the field by subtracting the mean difference between $WSB_{\psi PD}$ and Ψ_{PD} as follows:

$\Psi_{PDi}' = \Psi_{PDi} - mean\ (WSB_{\psi PDim} - \Psi_{PDi})$,

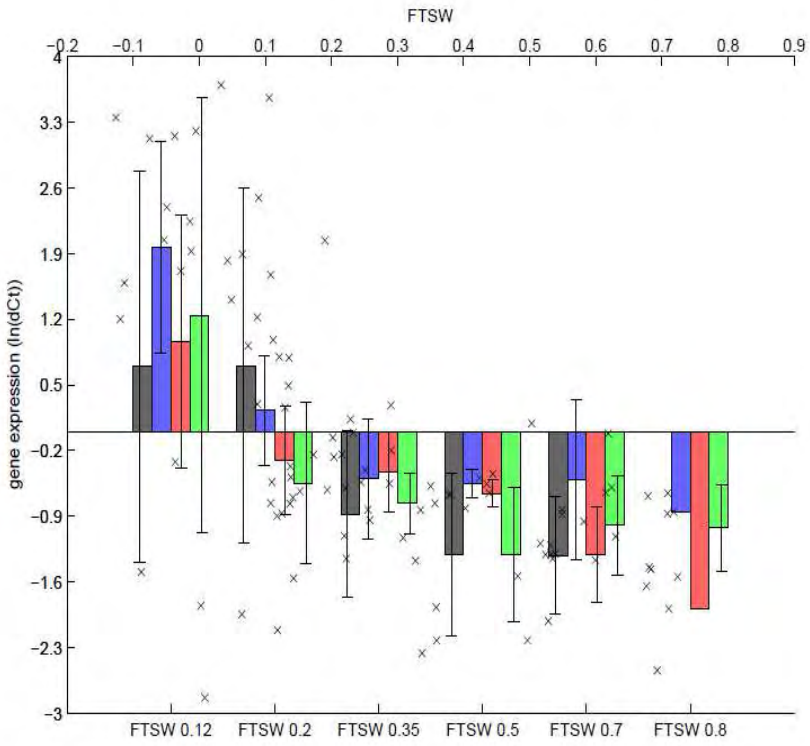
where Ψ_{PDi} is the pre-dawn leaf water potential of plant *i* measured in the field between 4:00 and 5:30 and $WSB_{\psi PDim}$ is the pre-dawn leaf water potential value for plant *i* predicted by the model *m* calibrated in the greenhouse. This transformation allowed us to choose the best model with observed data from field equivalent those from greenhouse (used to calibrate the model). In fact this correction accounted for the over-estimation by Ψ_{PD} of the WS at the time of the transcriptomic harvest, it didn't modify the model ranking based on R^2 of the correlation between predicted and observed data, but only reduced the RMSE (Figure II.3).

Appendix II.5: Raw data of gene expression level function of FTSW values.

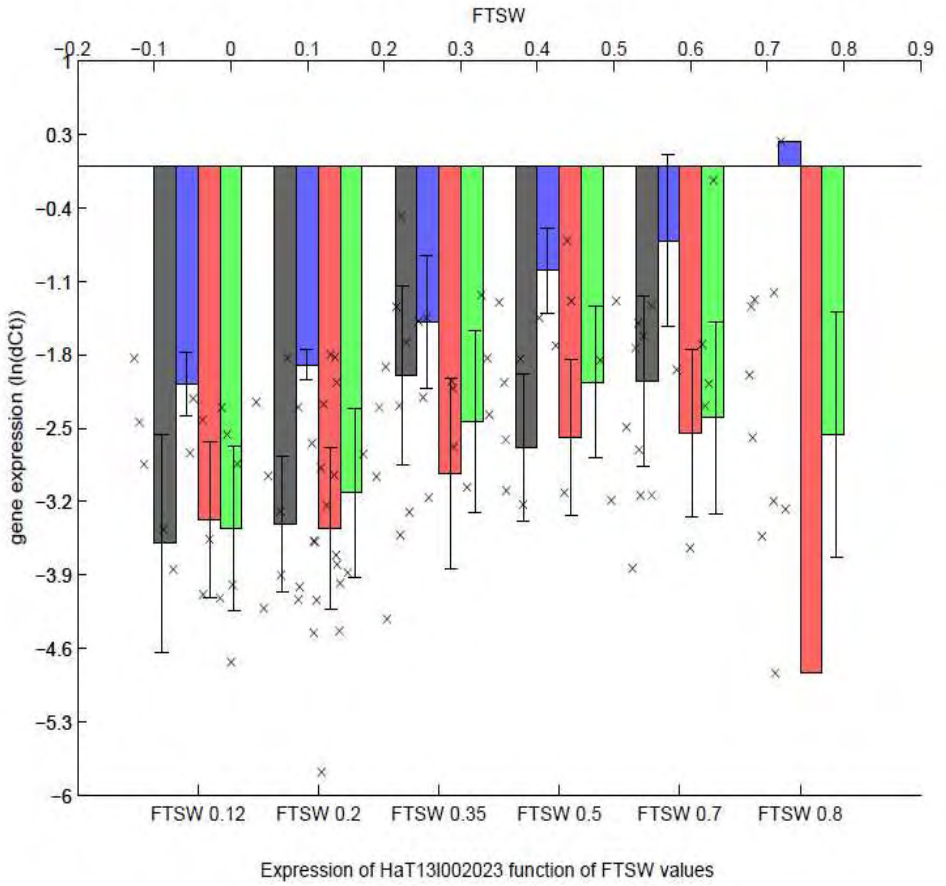
The histogram shows average gene expression level for the corresponding FTSW level. Inedi is represented in black, PSC8 in blue, XRQ in red and Melody in green. The error bars represent the standard deviation for each genotype and each FTSW condition. The scatter plot shows gene expression level in function of FTSW value. One point represents one individual plant.



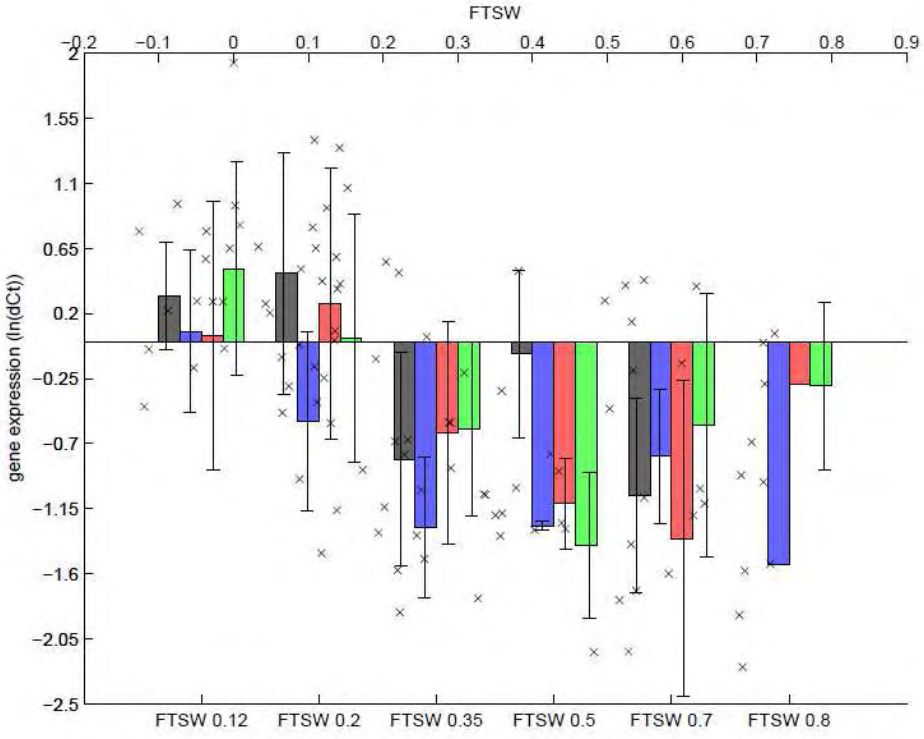
Expression of HaT13I001623 function of FTSW values



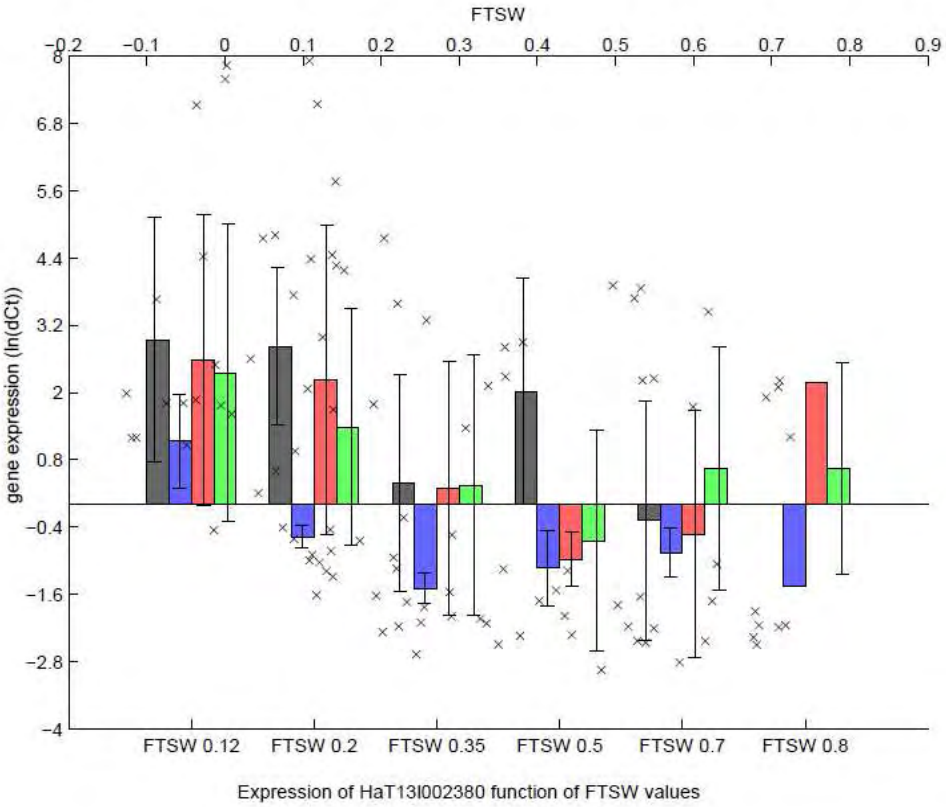
Expression of HaT13I001684 function of FTSW values

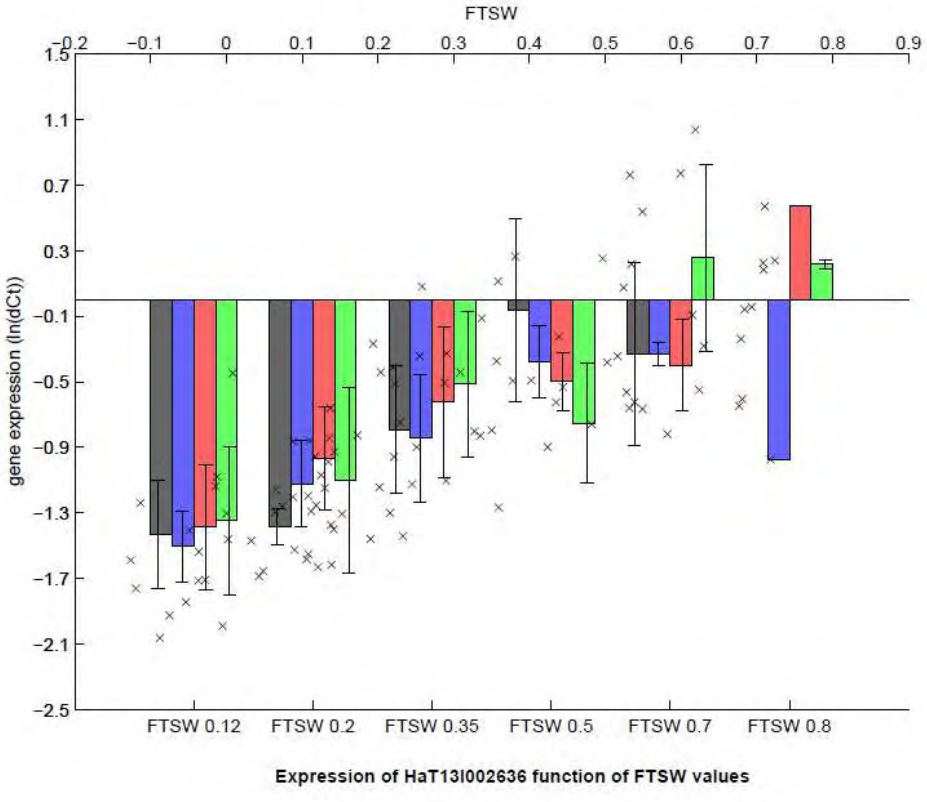


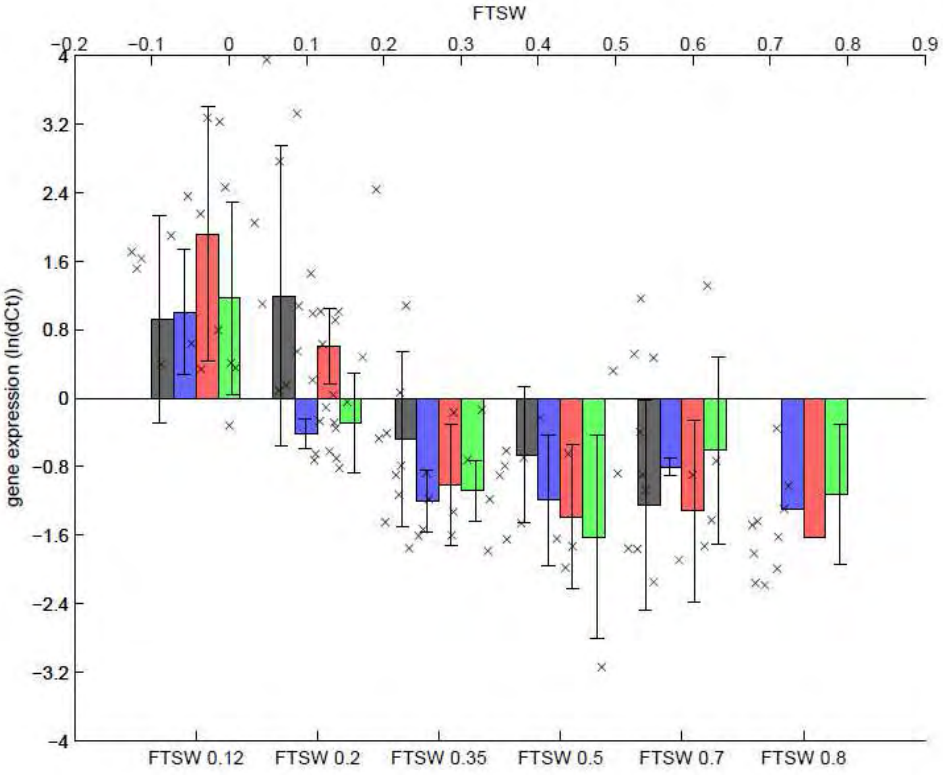
Expression of HaT13I002023 function of FTSW values



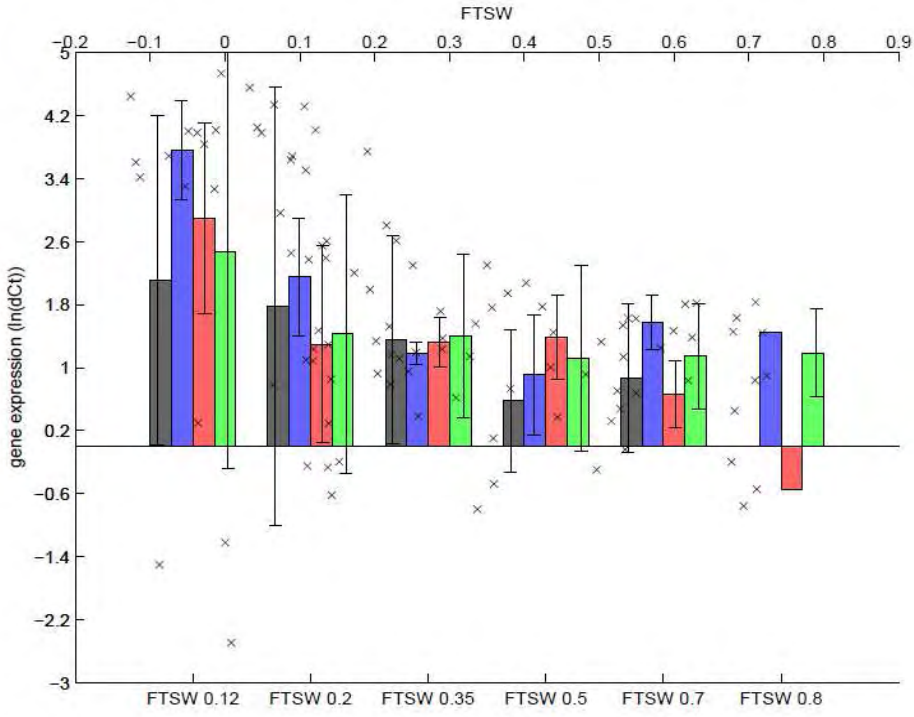
Expression of HaT13I002207 function of FTSW values


FTSW
−0.2 −0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
gene expression (ln(dCt))
8 6.8 5.6 4.4 3.2 2 0.8 −0.4 −1.6 −2.8 −4
FTSW 0.12 FTSW 0.2 FTSW 0.35 FTSW 0.5 FTSW 0.7 FTSW 0.8
Expression of HaT13I002380 function of FTSW values

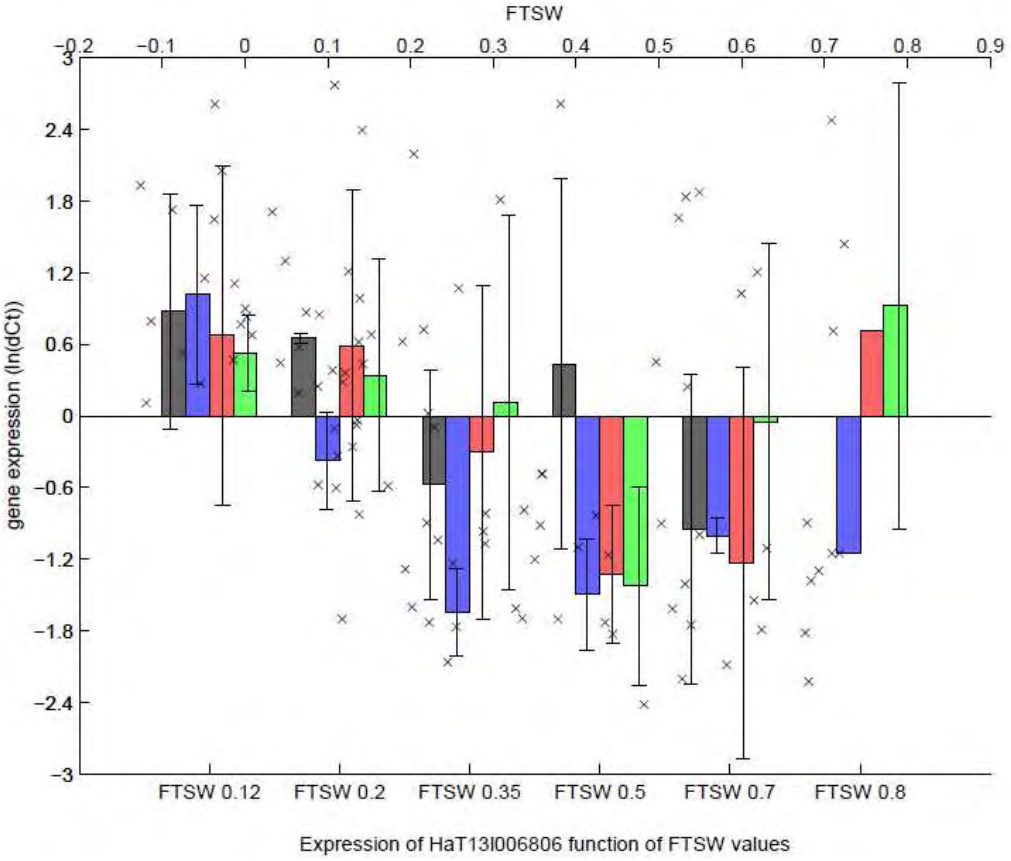

FTSW
−0.2 −0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
gene expression (ln(dCt))
1.5 1.1 0.7 0.3 −0.1 −0.5 −0.9 −1.3 −1.7 −2.1 −2.5
FTSW 0.12 FTSW 0.2 FTSW 0.35 FTSW 0.5 FTSW 0.7 FTSW 0.8
Expression of HaT13I002636 function of FTSW values



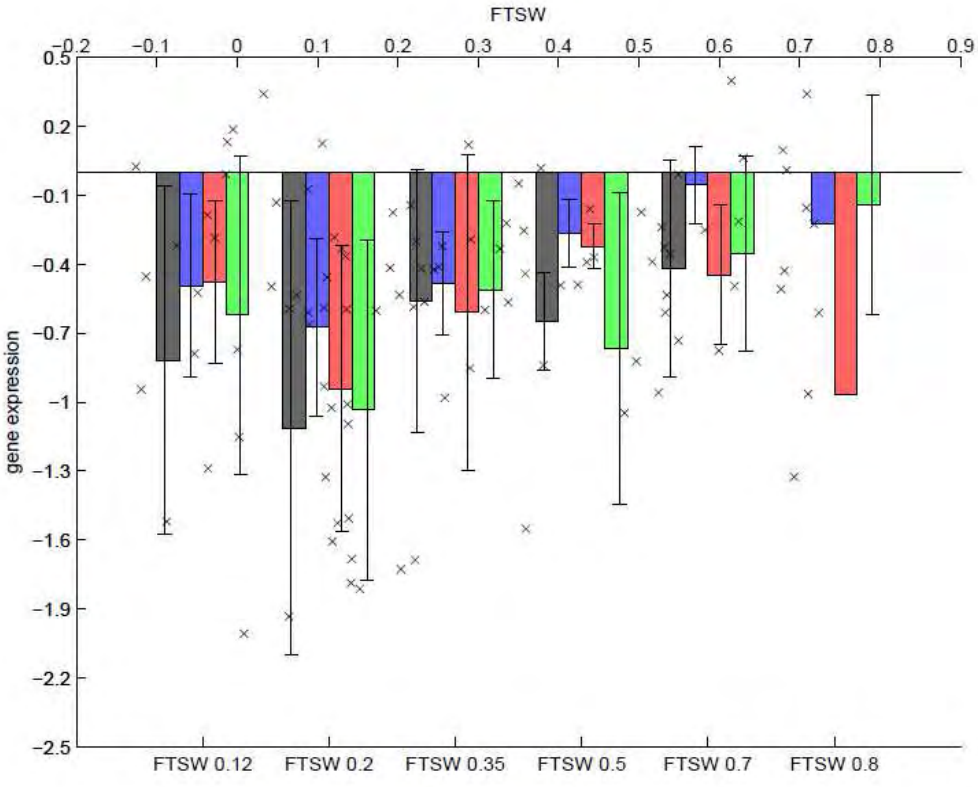
Expression of HaT13I005199 function of FTSW values



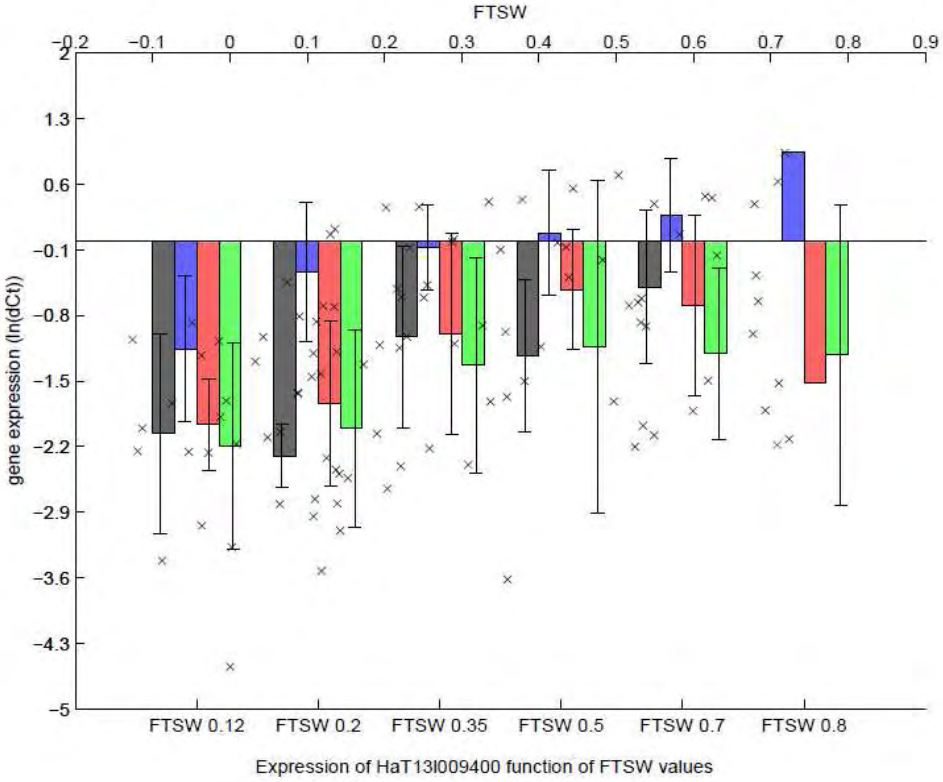
Expression of HaT13I006216 function of FTSW values



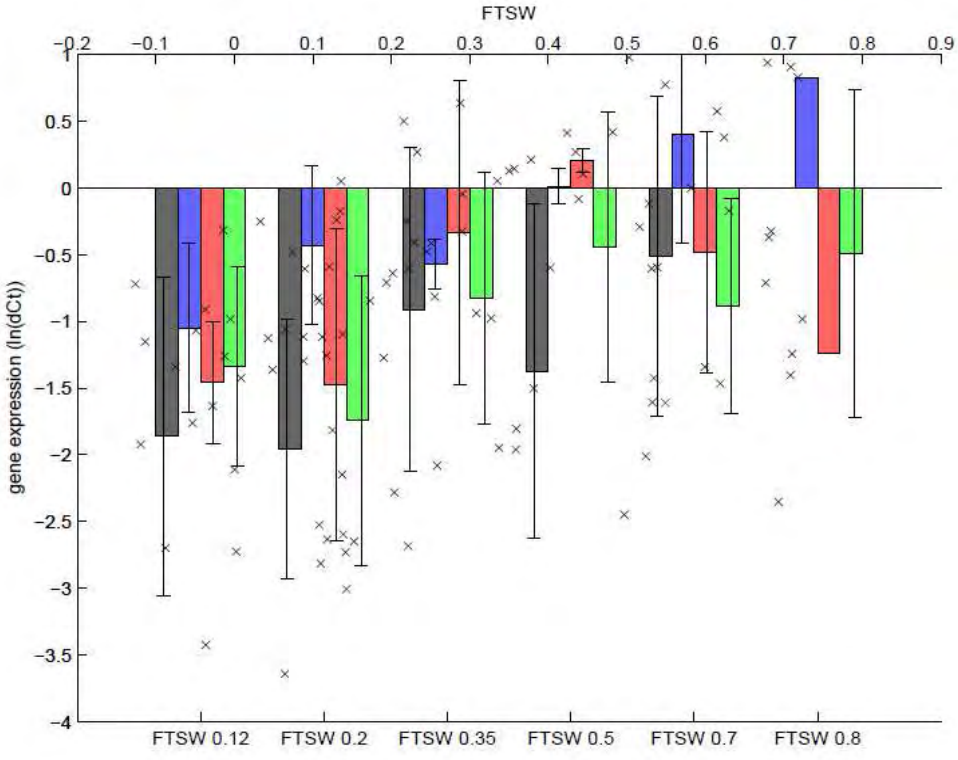
Expression of HaT13I006806 function of FTSW values



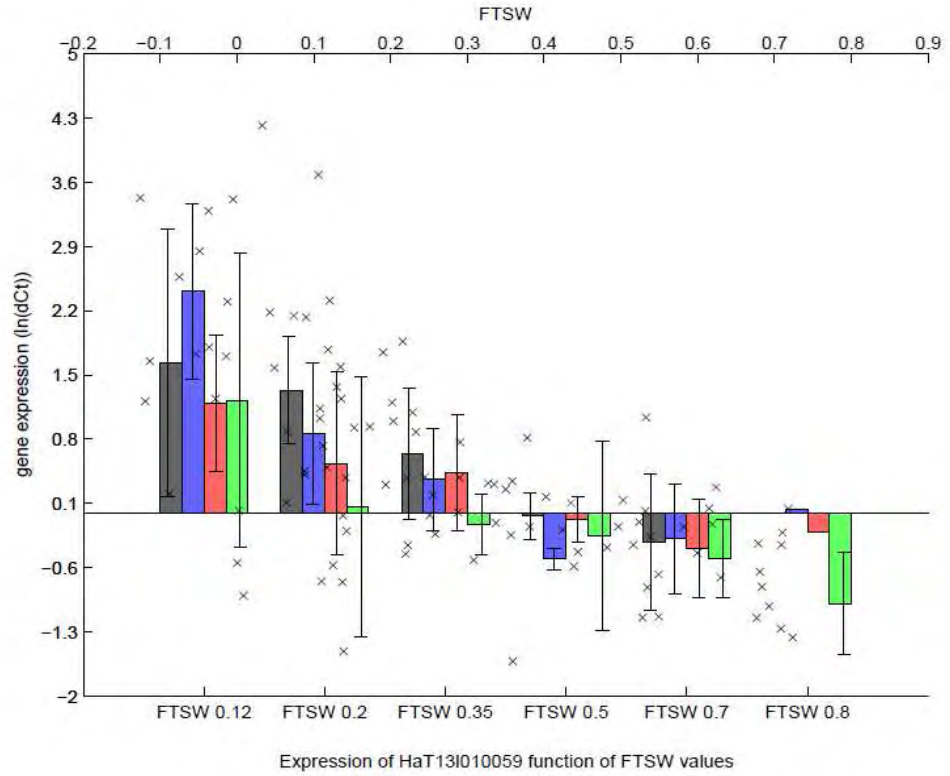
Expression of HaT13I008375 function of FTSW values



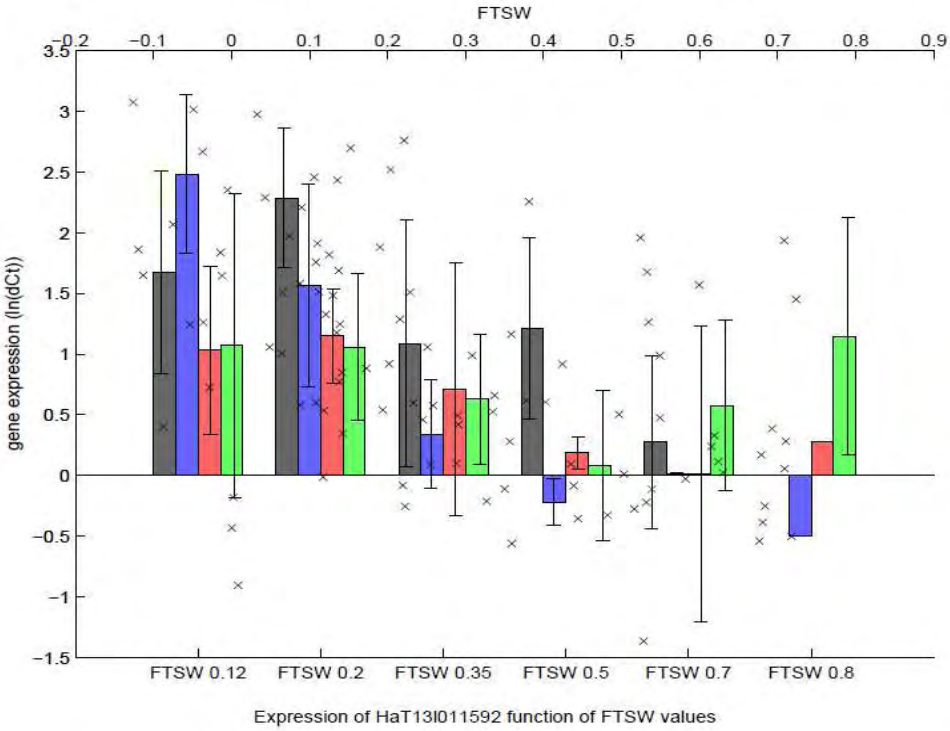
Expression of HaT13I009400 function of FTSW values



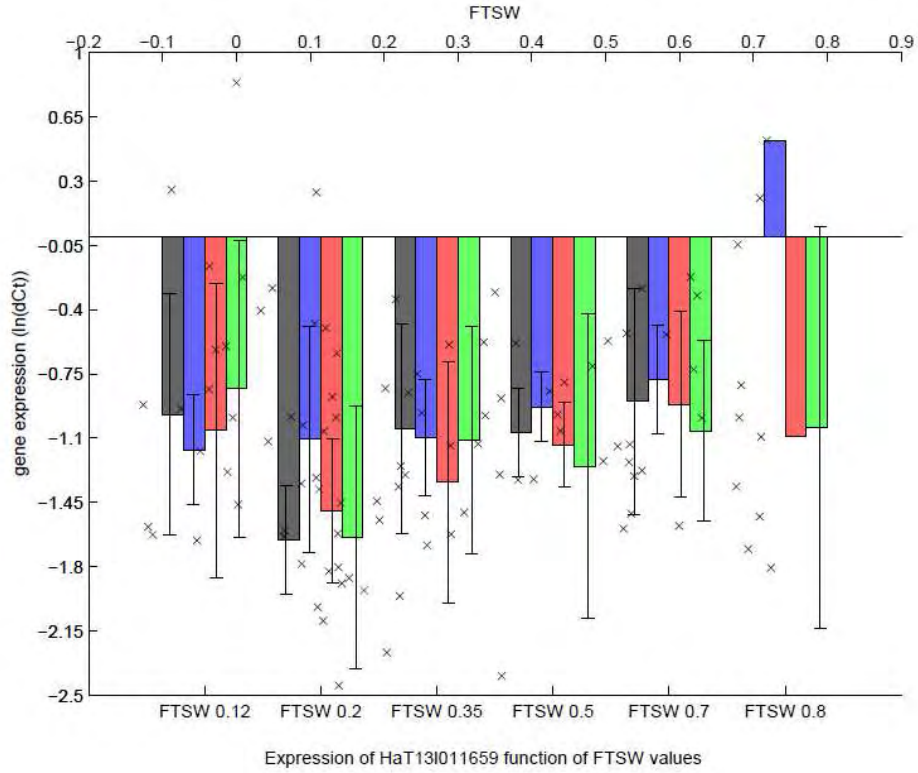
Expression of HaT13I009649 function of FTSW values



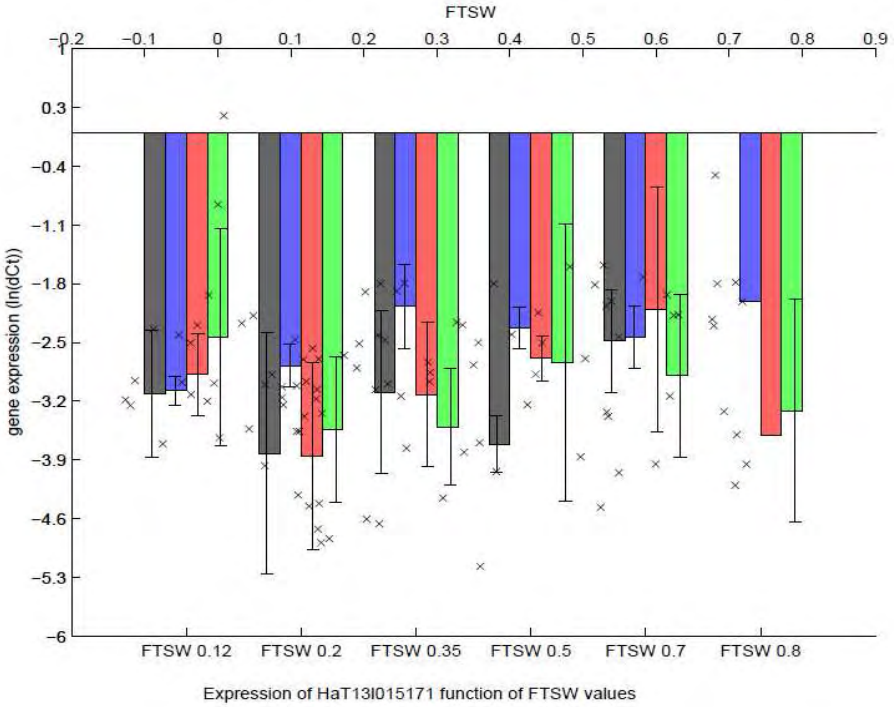
Expression of HaT13I010059 function of FTSW values



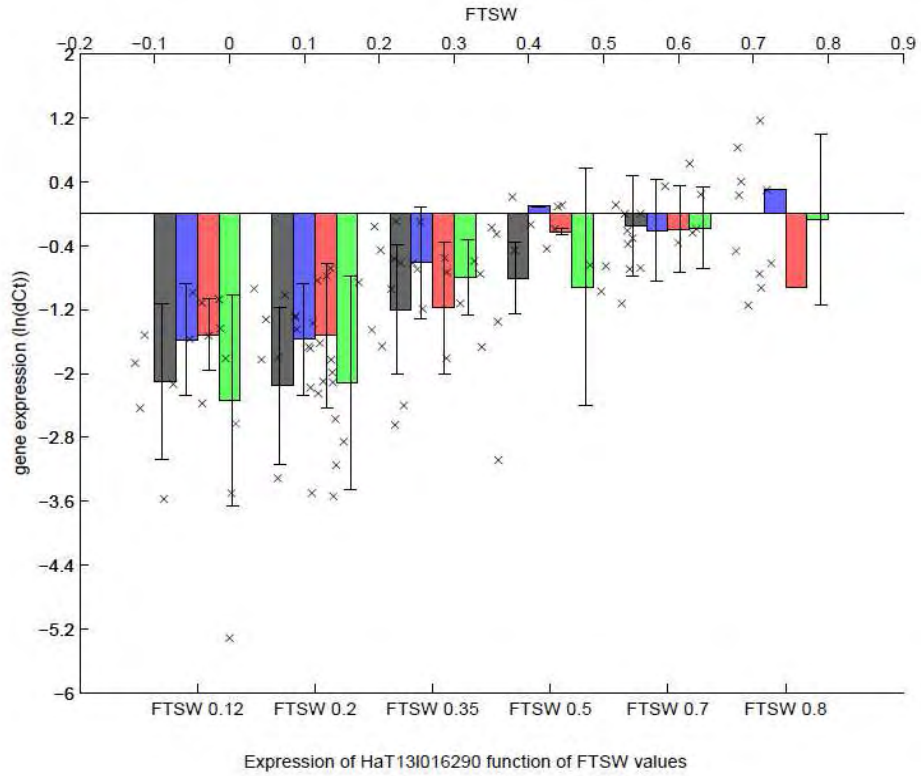
Expression of HaT13I011592 function of FTSW values



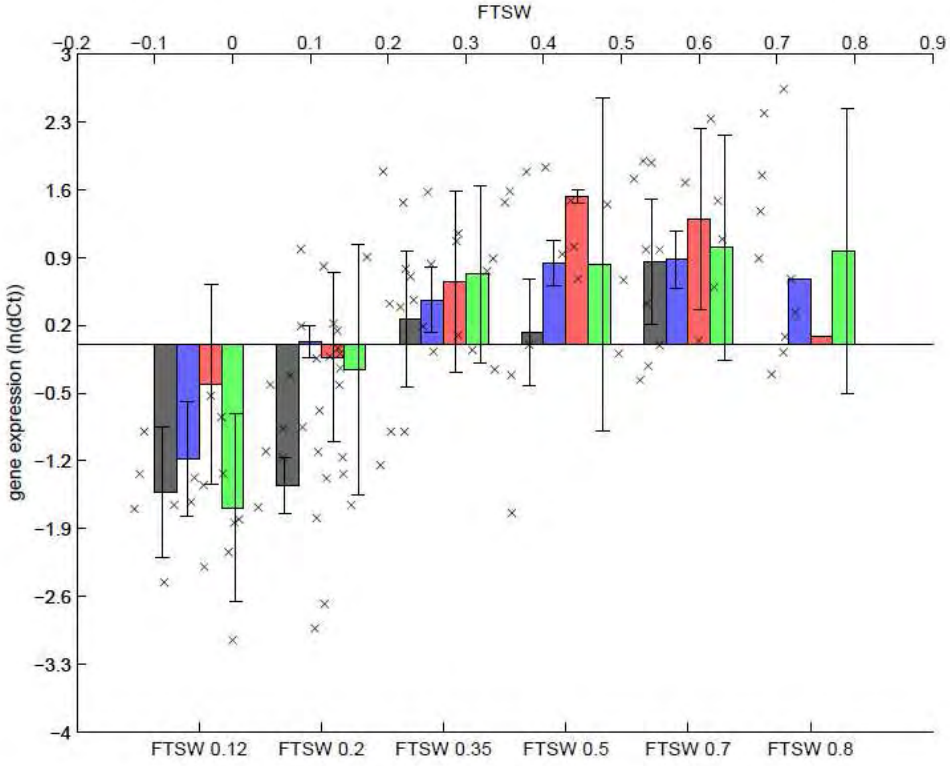
Expression of HaT13l011659 function of FTSW values



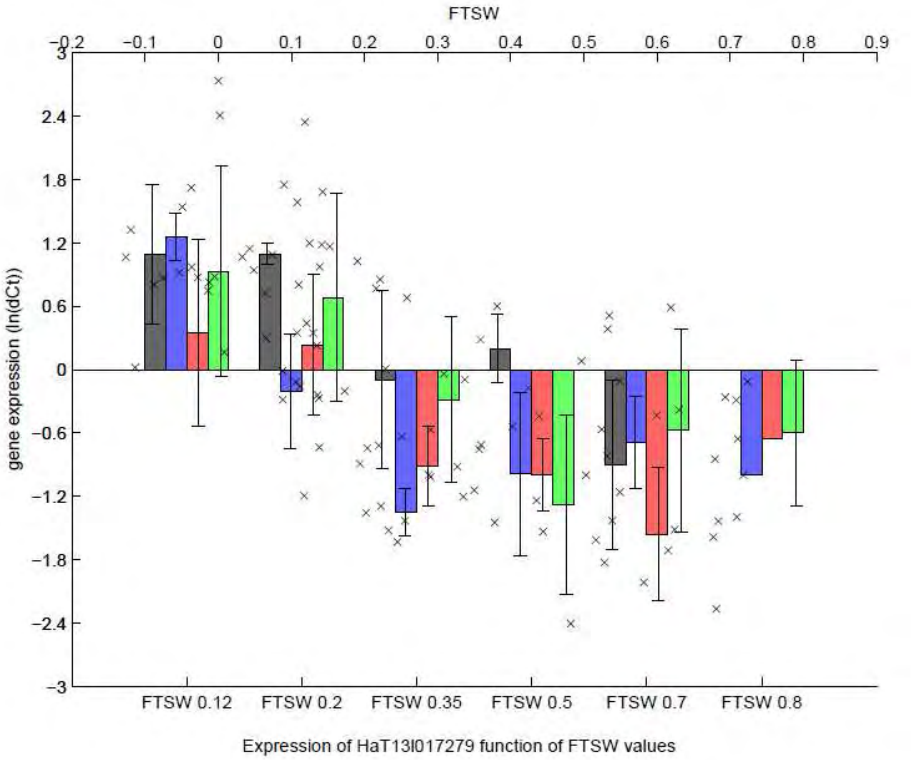
Expression of HaT13l015171 function of FTSW values



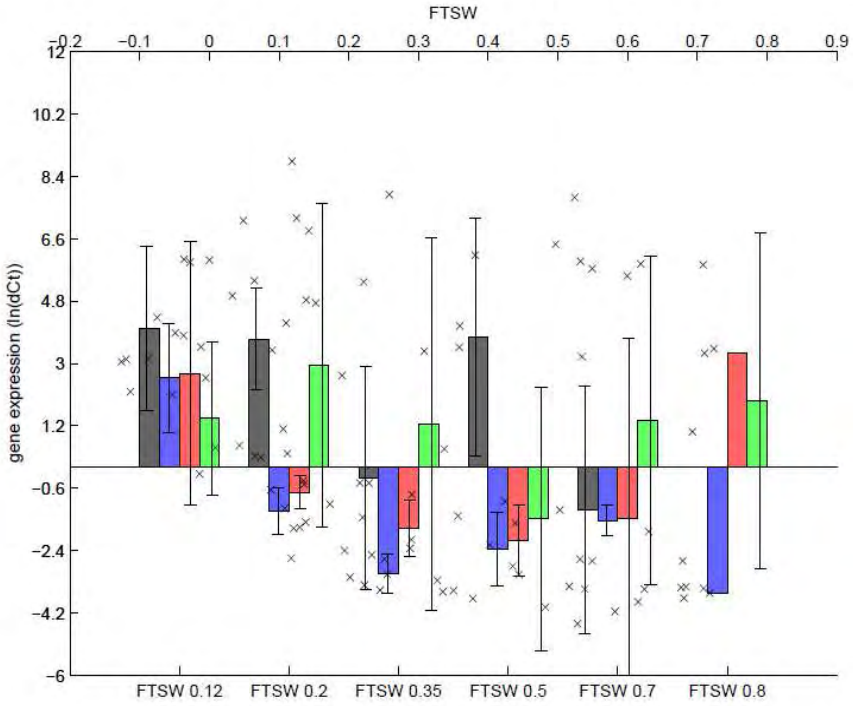
Expression of HaT13I016290 function of FTSW values



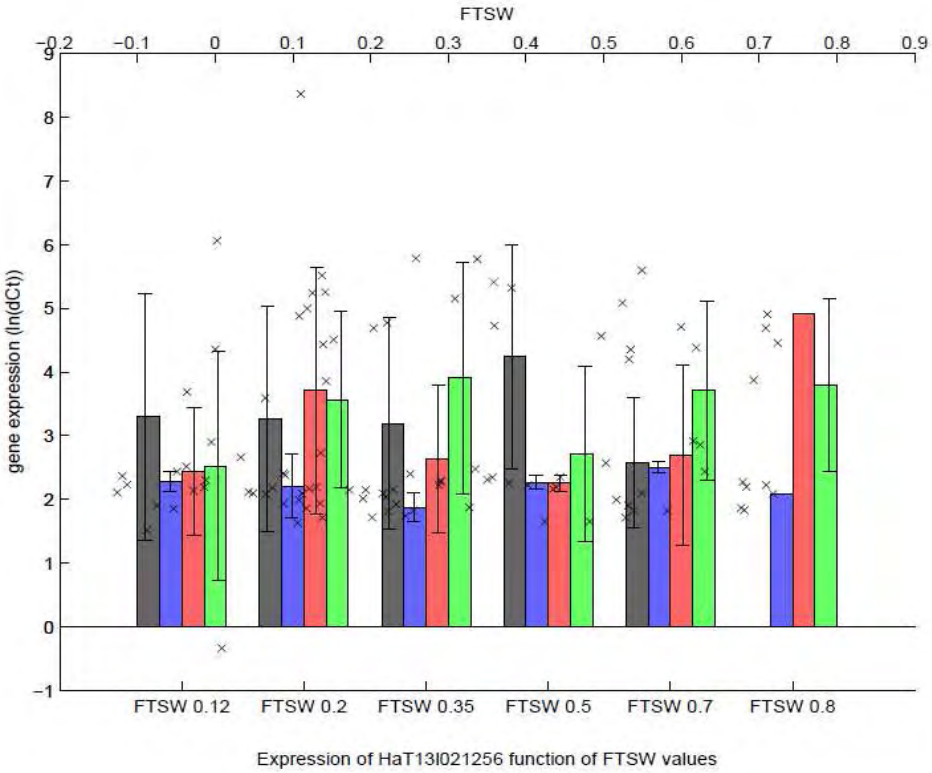
Expression of HaT13I 016327 function of FTSW values

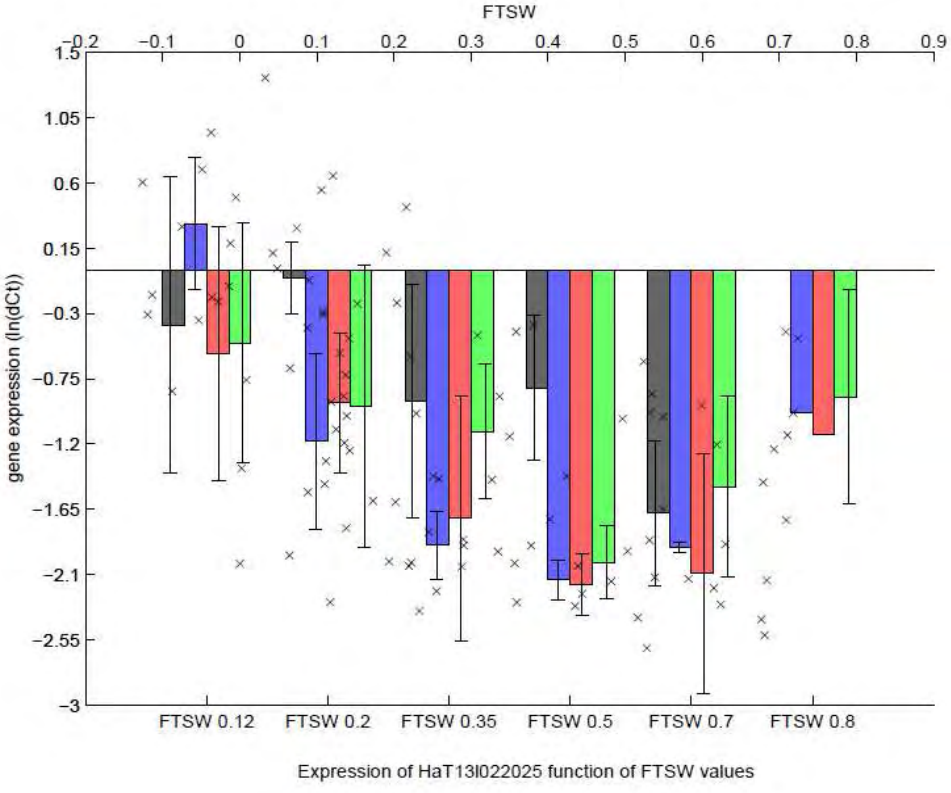


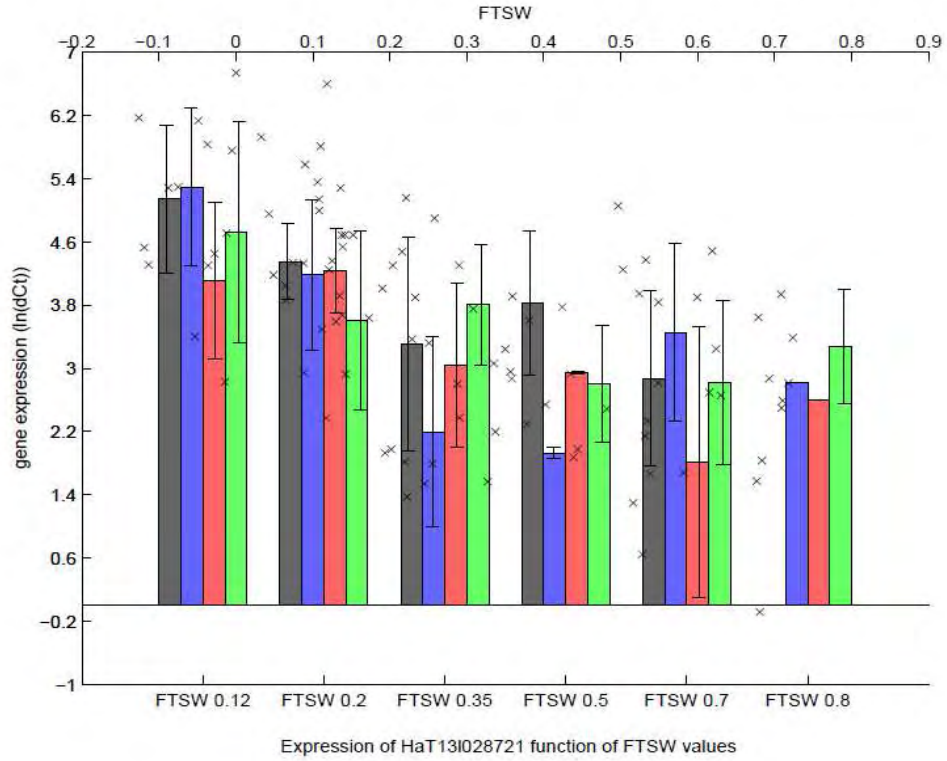
Expression of HaT13l017279 function of FTSW values



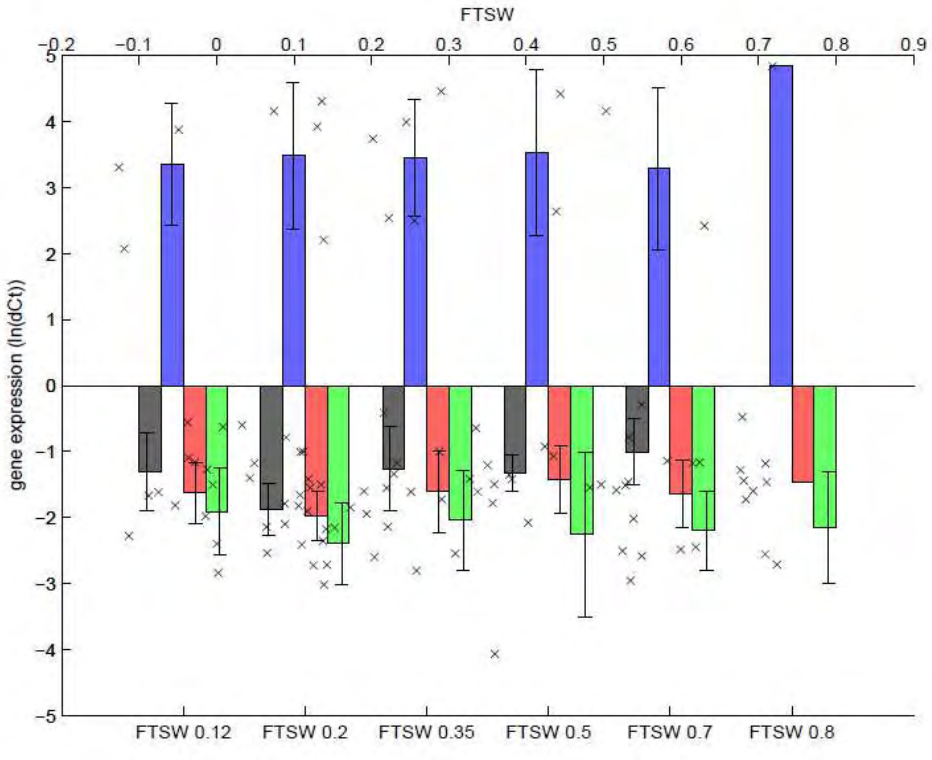
Expression of HaT13l020371 function of FTSW values


FTSW
−0.2 −0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
gene expression (ln(dCt))
9 8 7 6 5 4 3 2 1 0 −1
FTSW 0.12
FTSW 0.2
FTSW 0.35
FTSW 0.5
FTSW 0.7
FTSW 0.8
Expression of HaT13I021256 function of FTSW values

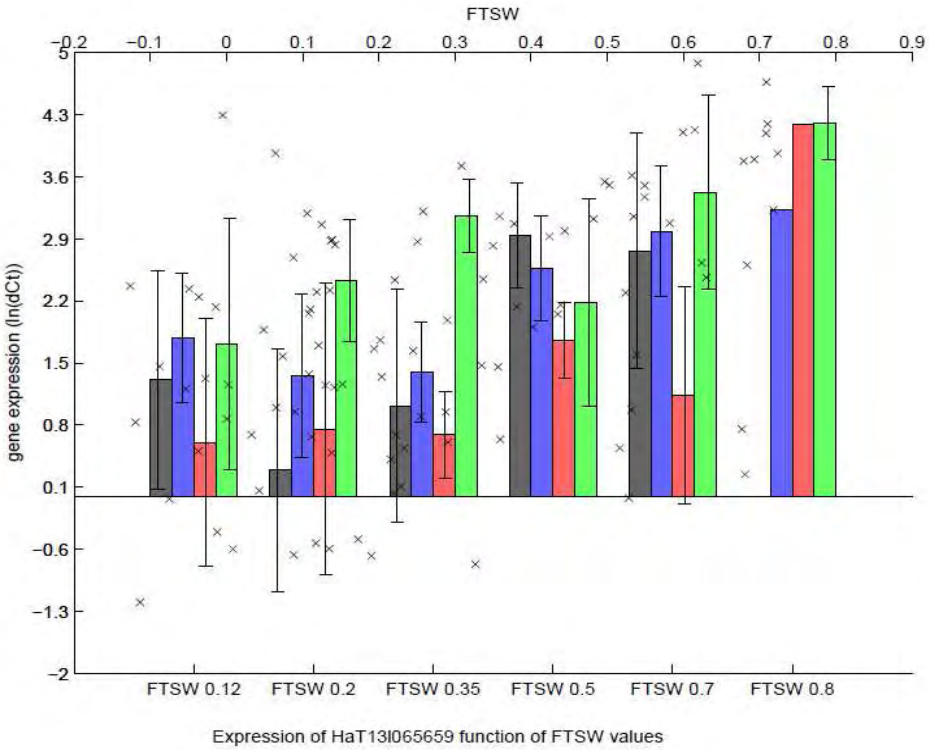

FTSW
−0.2 −0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
gene expression (ln(dCt))
1.5 1.05 0.6 0.15 −0.3 −0.75 −1.2 −1.65 −2.1 −2.55 −3
FTSW 0.12
FTSW 0.2
FTSW 0.35
FTSW 0.5
FTSW 0.7
FTSW 0.8
Expression of HaT13I022025 function of FTSW values



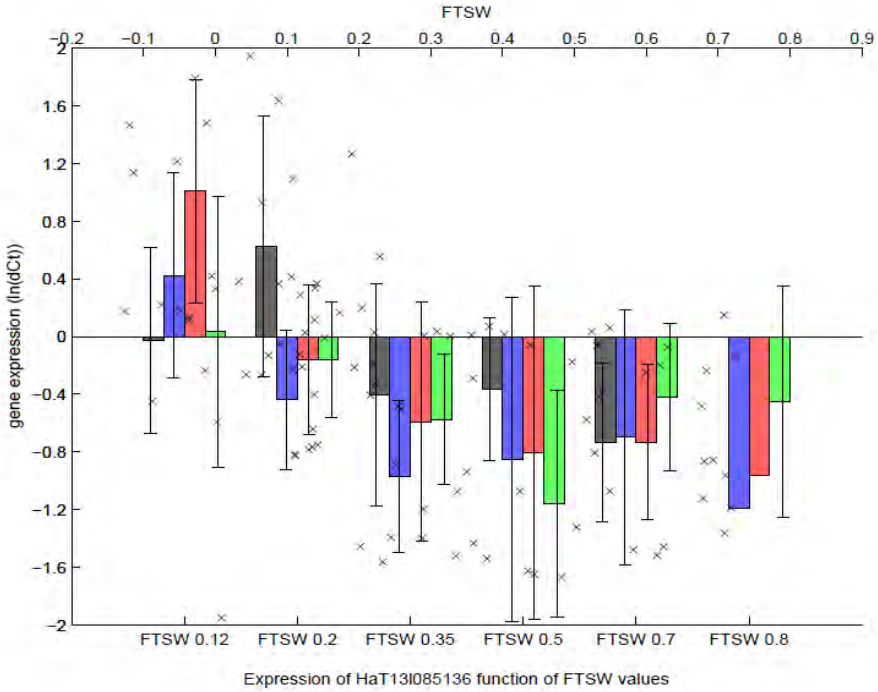
Expression of HaT13I028721 function of FTSW values



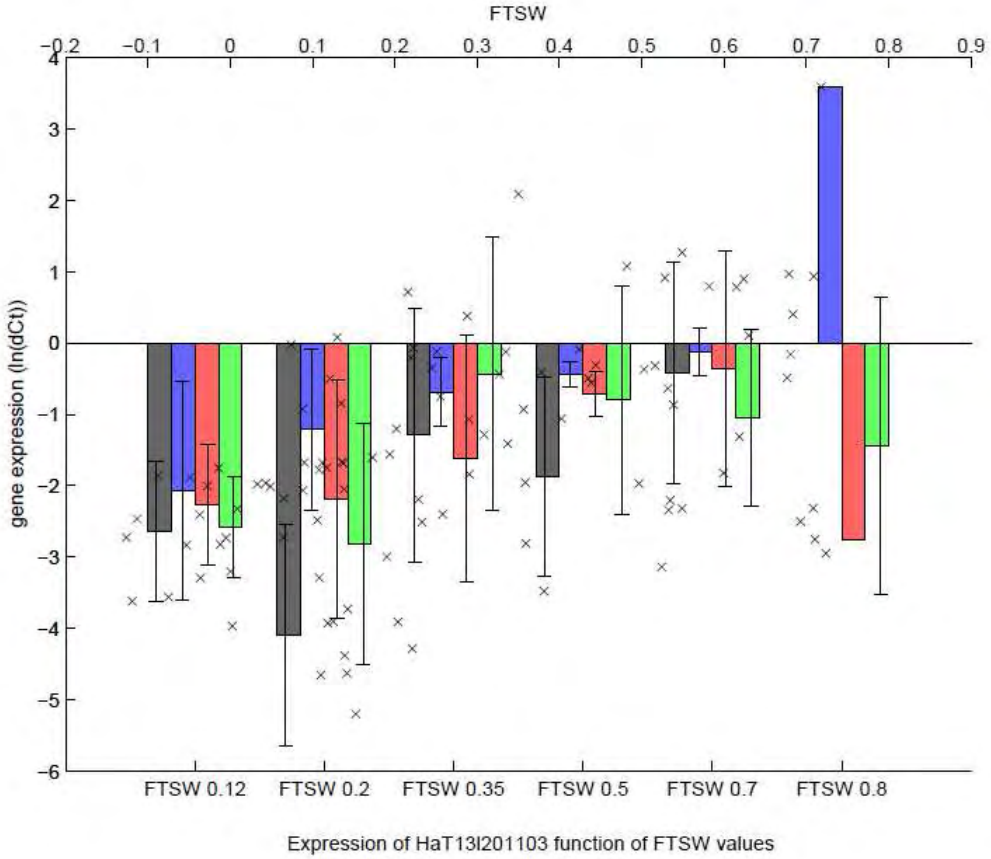
Expression of HaT13I041788 function of FTSW values



Expression of HaT13I065659 function of FTSW values



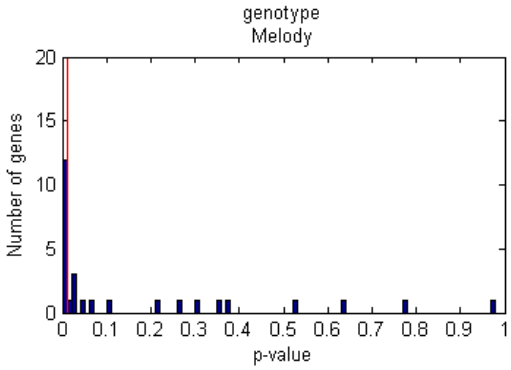
Expression of HaT13I085136 function of FTSW values

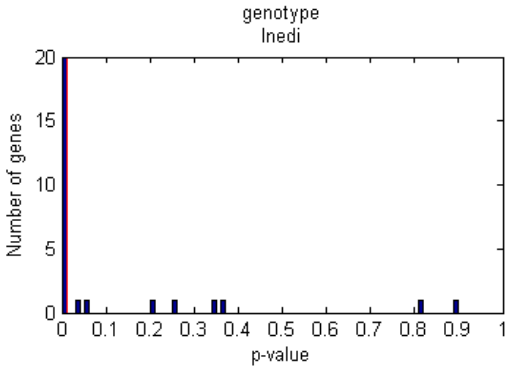

FTSW
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
gene expression (ln(dCt))
4
3
2
1
0
−1
−2
−3
−4
−5
−6
FTSW 0.12
FTSW 0.2
FTSW 0.35
FTSW 0.5
FTSW 0.7
FTSW 0.8
Expression of HaT13I201103 function of FTSW values

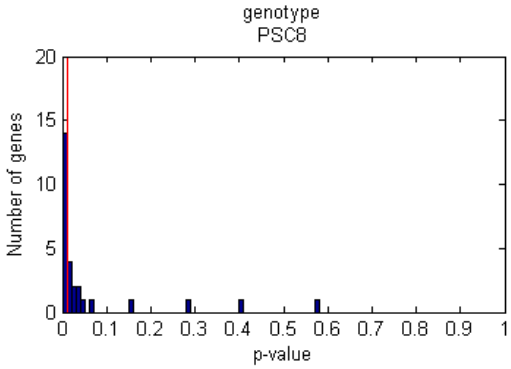
Appendix II.6: P-value distribution for the correlation between expression level of the 28 candidate genes and the four WSI.

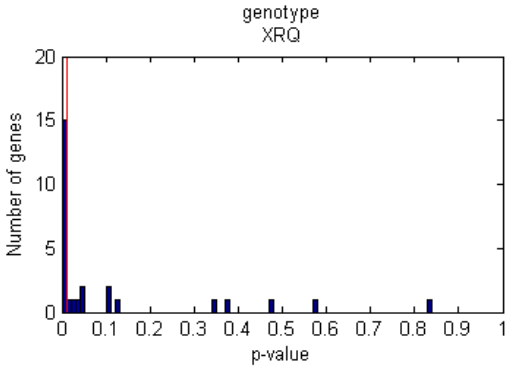
The red line show the threshold selection p-value < 0.001

a) Correlation with FTSW

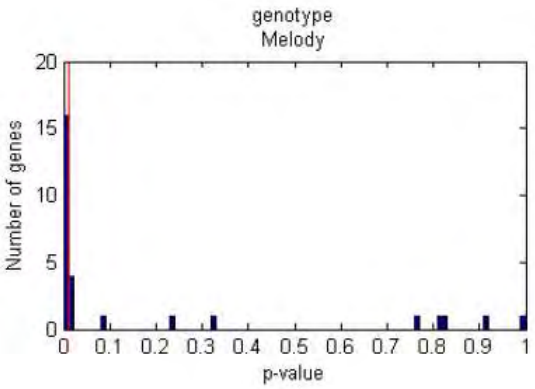


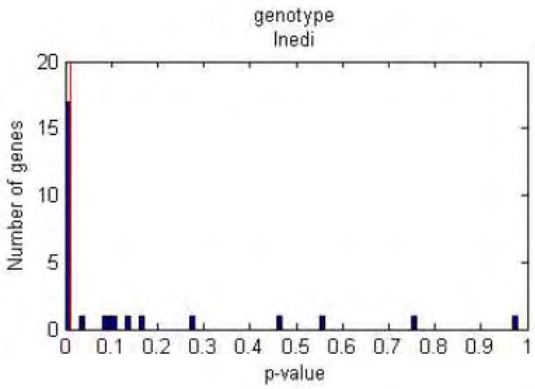


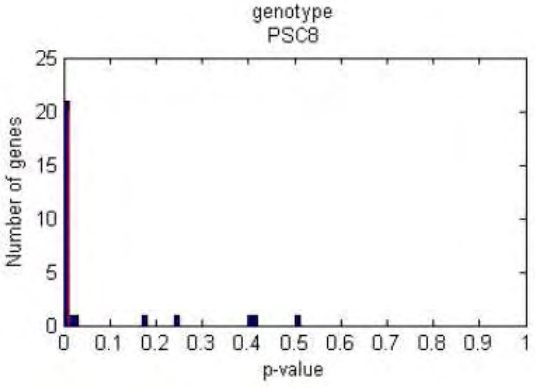


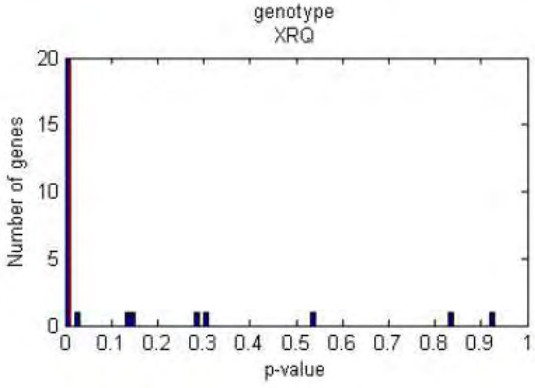


a) Correlation with Ψ'_{PD}

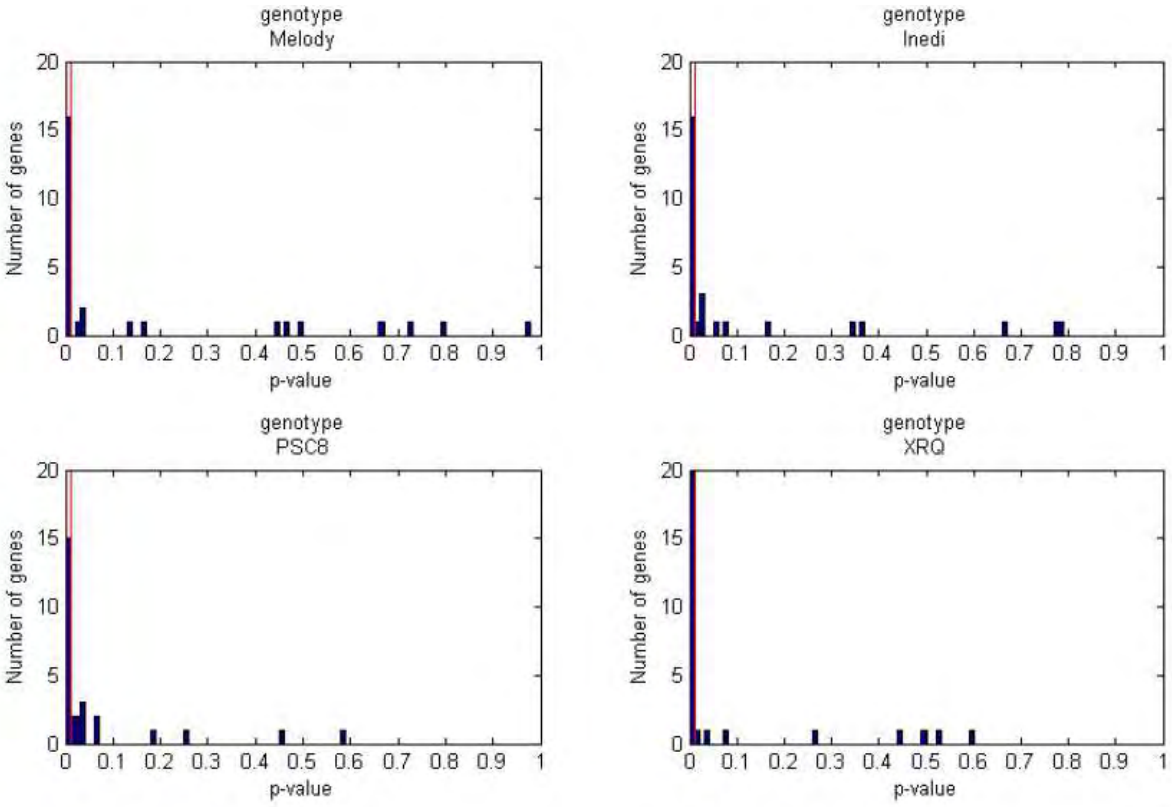




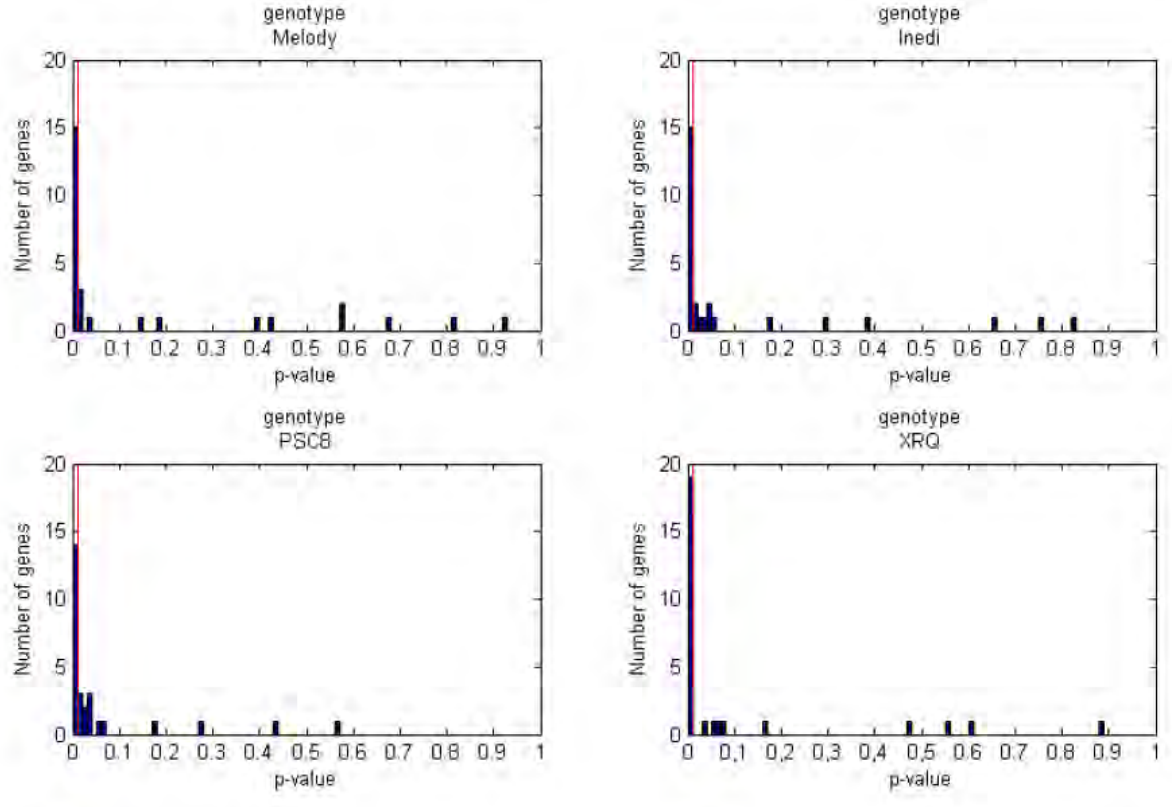




b) Correlation with SWC



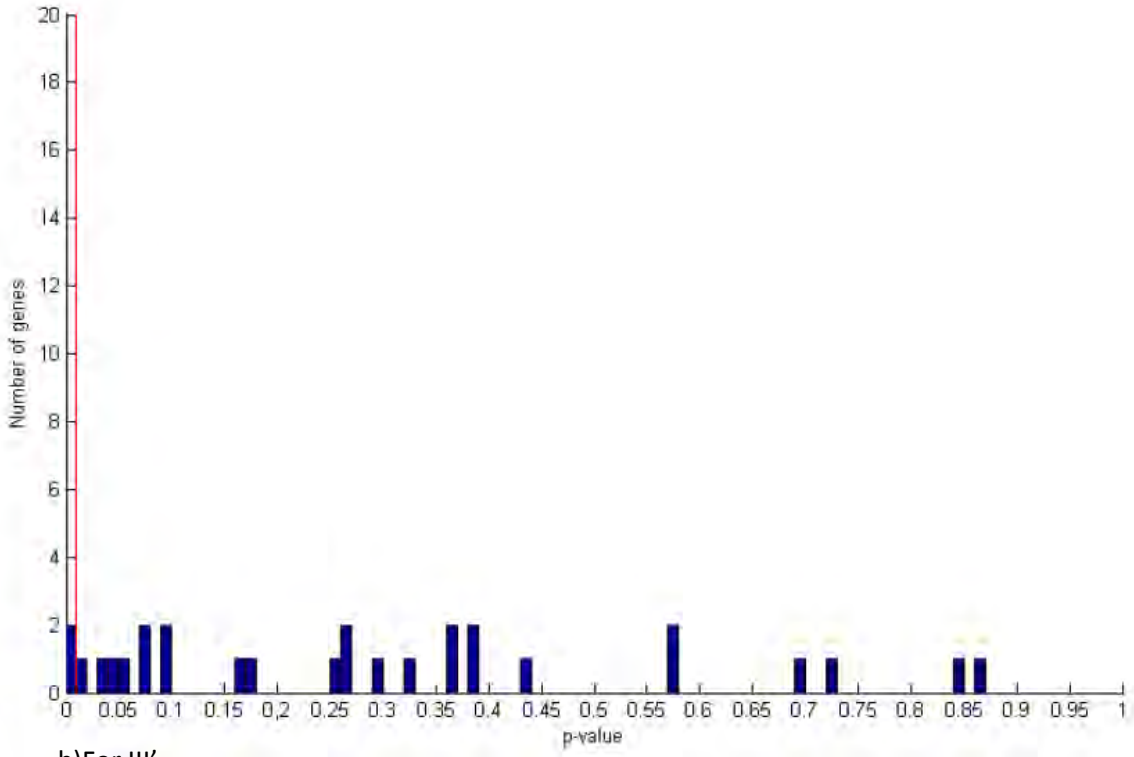
c) Correlation with FtotSW



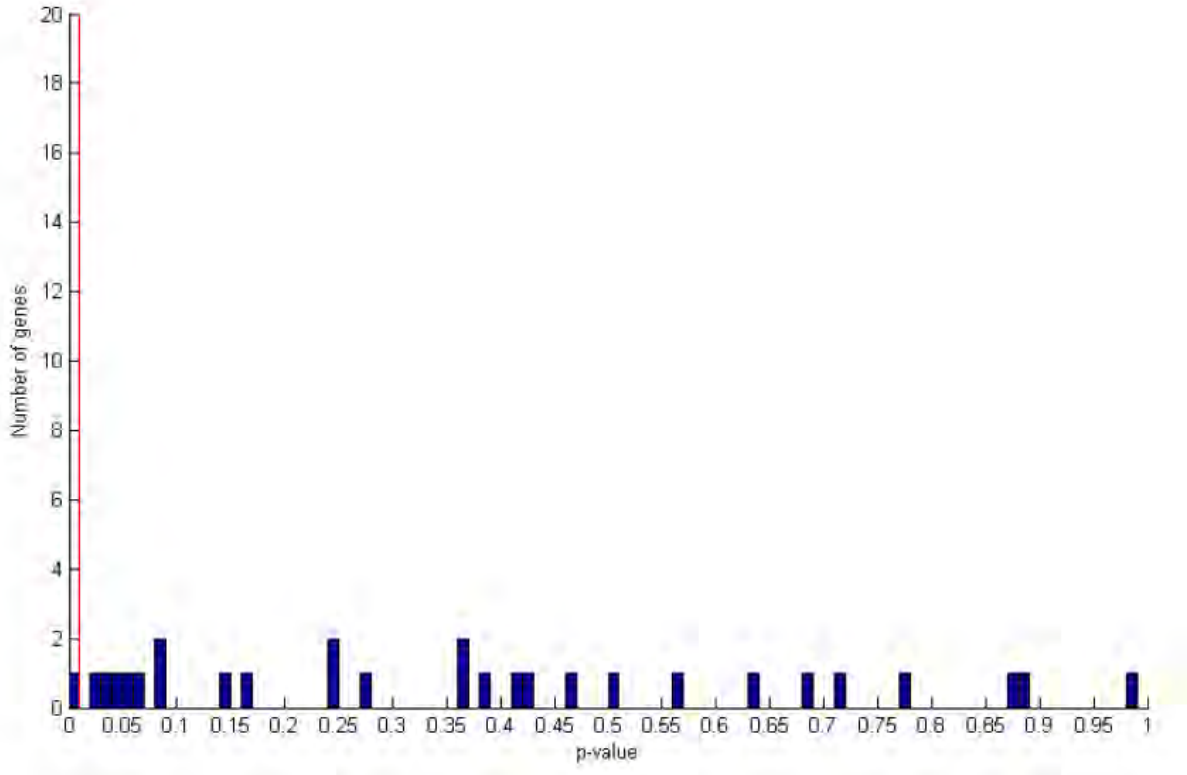
Appendix II.7: P-value distribution of F-test for comparison between correlation model independent of genotypes and correlation model dependent of genotypes between candidate gene expression and the four WSI.

The red line show threshold selection : p-value > 0.001.

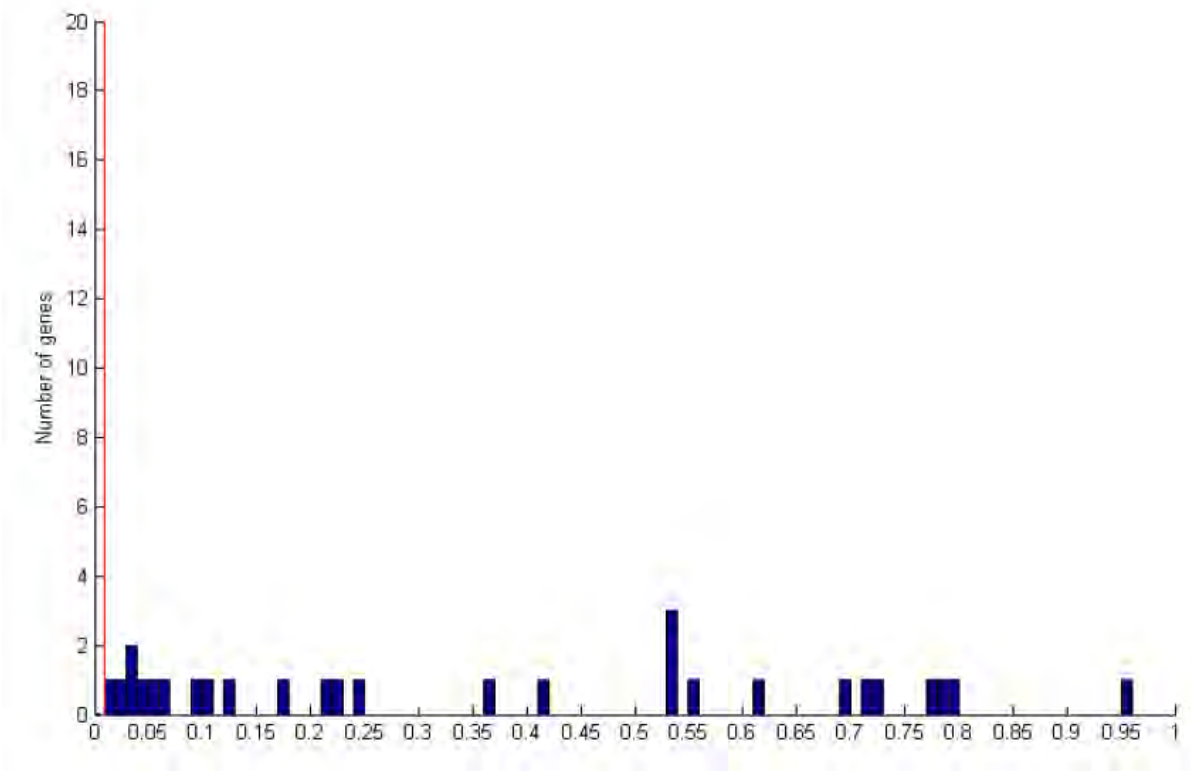
a)For FTSW



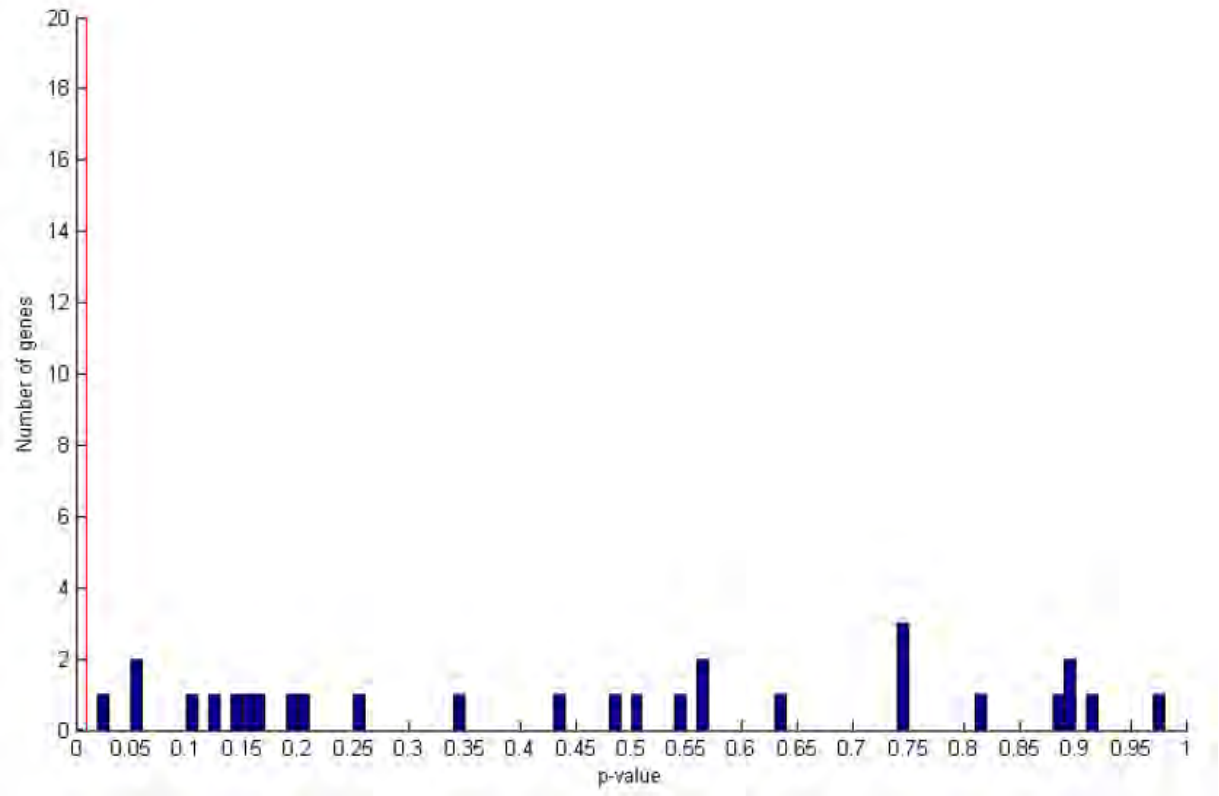
b)For Ψ'_{PD}



c)For SWC



d)For FtotSW



Appendices Chapter III

Appendix III.1: Raw and normalized values of WSB for each genotypes of tha association panel

This appendix is available on line at the address: https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents: (1) raw data for WSB and corresponding ΨPD; (2) corresponding block effect; (3) normalized data for the block effect (Excel File)

Block effect were calculated using the following ANOVA model

$Yij=\mu +Gi +Bj+\varepsilon ij$,

Where Yij is the WSB observation for the ith genotype in the jth block, μ is the intercept term, Gi is the genetic effect of the ith genotype, Bj is the effect of the jth block and εij is the residual error

Appendix III.2: List of selected genes for GWA study and their functional annotations.

This appendix is available on line at the address: https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents the 86 genes selected from Rengel *et al.* 2012 and used in the GWA study, the reference genes and the 3 genes used to calculate the WSB. (Excel File)

Appendix III.3: BLUPs for the 86 gene expressions and WSB using G or GE models

This appendix is available on line at the address: https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents (1) BLUPs for the 86 studied genes for each panel lines caculated with the G model; (2) BLUPs for the 86 studied genes for each panel lines caculated with the GE model; (3) BLUP for the WSB for each panel lines caculated with the GE model. (Excel File)

Appendix III.4: Paired t-test results to compare G and GE models

This appendix is available on line at the address:

https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents for the 68 genes with results in both G and GE models, results of the t-test, its p-value and the confidence intervalle bounds (Excel File)

Appendix III.5: Complete results of association and QTL detection

This appendix is available on line at the address:

https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents a description of the association study results for the gene expression with BLUPs different from zero (Excel File).

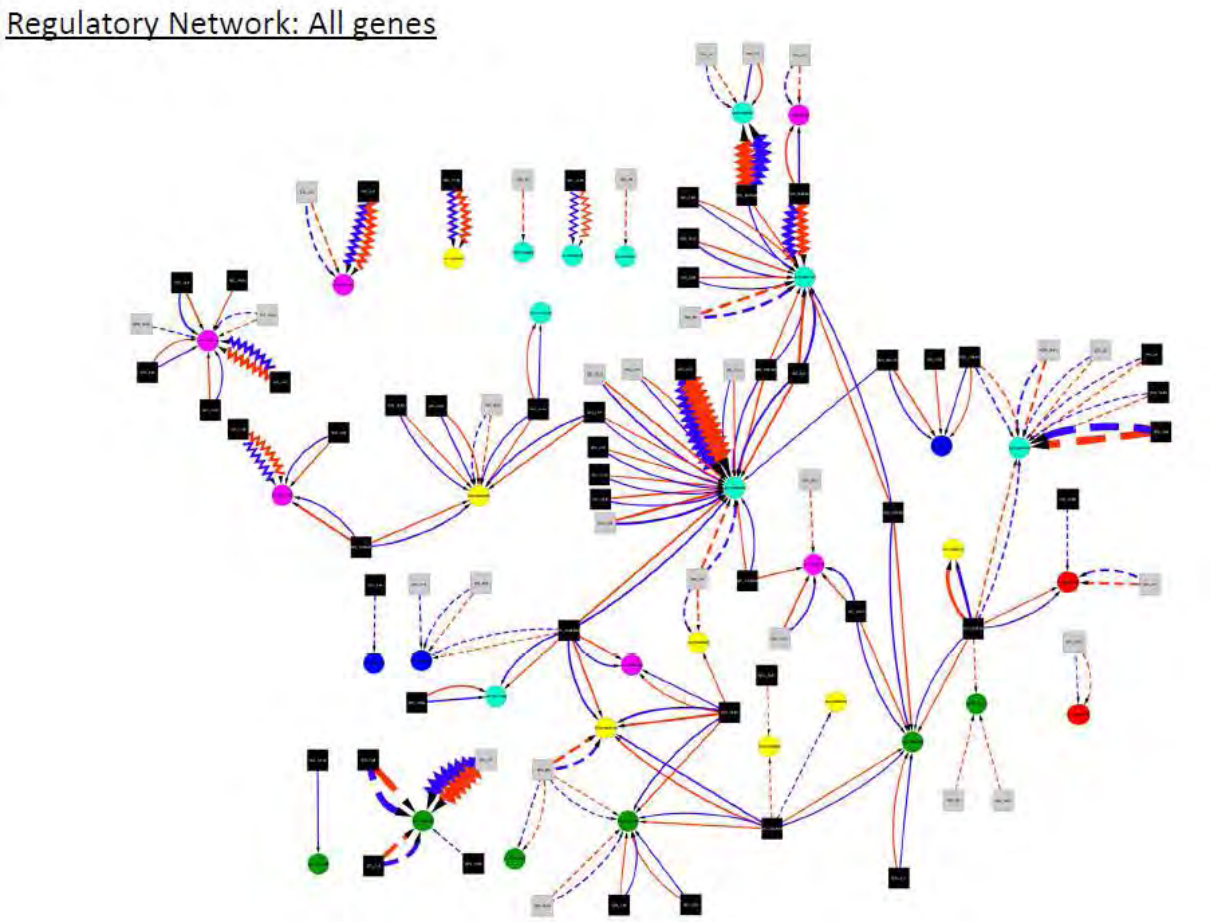
Appendix III.6: Effect of the associated SNP and comparison between G and GE model

This appendix is available on line at the address:

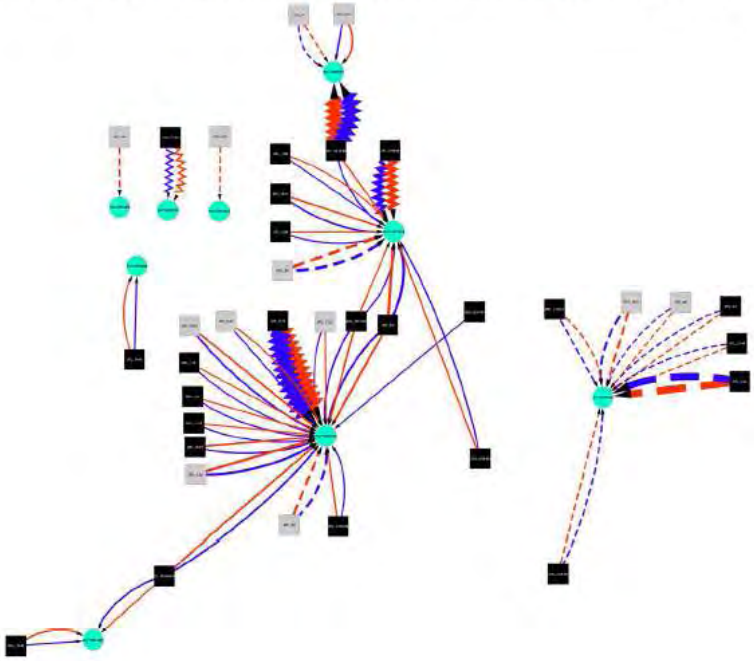
https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents for each gene the effect of the associated SNP its confidence intervalle and the comparison between the two models (Excel Files)

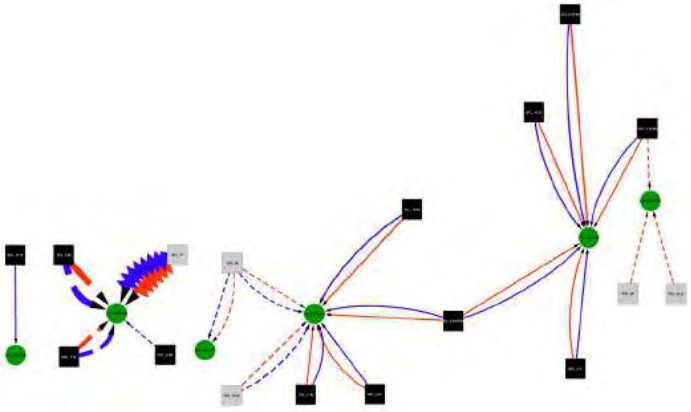
Appendix III.7: Gene regulatory Network and physiological variable correlated to genes expressions



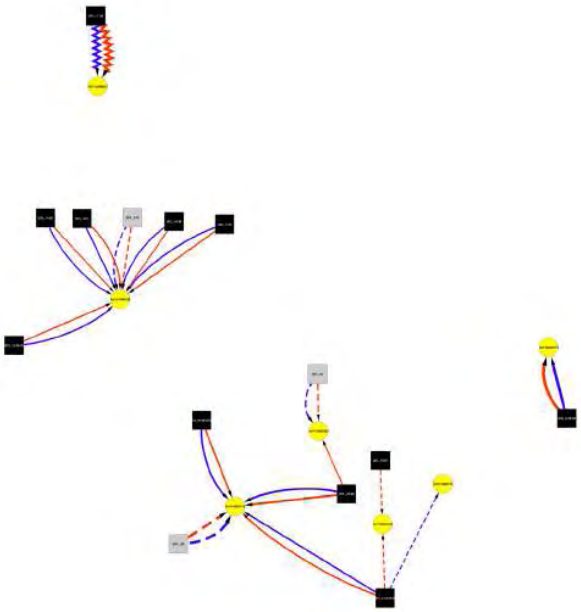
Regulatory Network: Genes correlated to Carbon Isotopic Discrimination (CID)



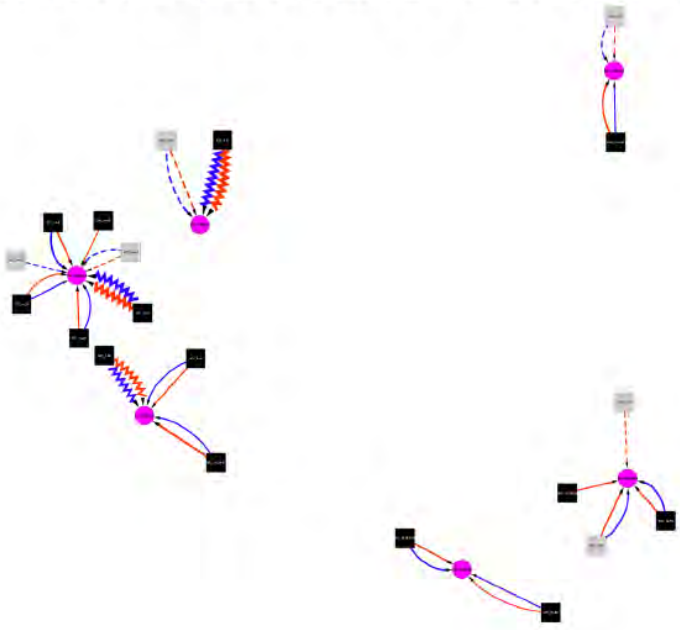
Regulatory Network: Genes correlated to Evapotranspiration (ET)



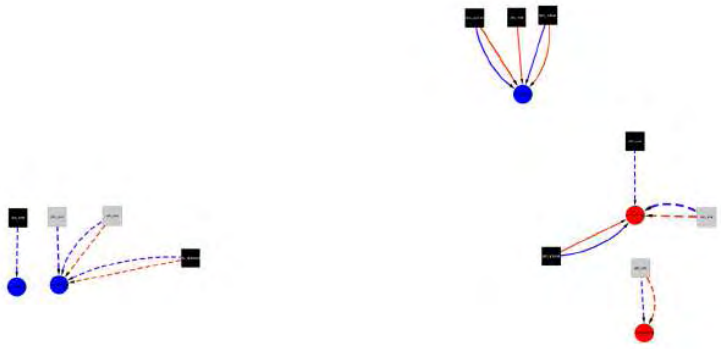
Regulatory Network: Genes correlated to FISGT



Regulatory Network: Genes correlated to Osmotic Potential (OP)



Regulatory Network: Genes correlated to Relative Water Content (RWC)



Appendices Chapter IV

Appendix IV.1: Supporting Materials and Methods for the article about GRN inference and diversity study

Plant hormonal treatment

After 14 days, the sunflower XRQ plantlets (grown in hydroponic conditions in growth chamber) were treated by adding either mock solution (DMSO only in controls) or one of the following hormonal solutions to the hydroponic solution : (i) 0.1 µM indole acetic acid (IAA); (ii) 0.25 µM 1-aminocyclopropane-1-carboxylic acid (ACC); (iii) 10 µM gibberellic acid 3 (GA3); (iv) 0.05 \micro M salicylic acid (SA); (v) 1 µM methyl-jasmonate (MeJA); (vi) 0.5 µM kinetin; (vii) 10 µM ABA (viii) 0.1 µM rac-GR24, a strigolactone (Stri) analog; or (ix) 1 µM 24-epibrassinolide (Bras).

Molecular analysis

The extraction of total RNA and cDNA synthesis were performed as described in (Rengel, D. *et al.*, 2012b). The expression levels of the 181 selected genes were analyzed in all samples by q-RT-PCR using the BioMark system (Fluidigm Corporation, San Francisco, CA, USA) as previously described (Spurgeon *et al.*, 2008) The q-RT-PCR results were analyzed following the 2ddCt method (Livak & Schmittgen, 2001). The threshold cycle Ct, which indicates the cycle number when the signal reaches the detection threshold, was calculated using Fluidigm Biomark software. The amplification efficiency for gene target X, denoted Eff_X was estimated using the robustfit function in the Matlab (version 7.11) Statistics Toolbox (version 7.4):

$$X_T = X_0 * (1 + Eff_X)^{Ct,x} \qquad (1)$$

where X_T is the threshold number of target molecules, X_0 is the initial number of target molecules, and Ct,x is the threshold cycle for target amplification.

The amount of target was then normalized to the mean of previously validated reference genes (Rengel, D. *et al.*, 2012b) and to the corresponding control sample with the mock treatment at the studied time, as shown in the following equation:

$$\text{Normalized amount of target} = \Delta\Delta Ct = ((1+Eff_X)^{(Ct,x,qCt,x,c)})/((1+Eff_R)^{(Ct,R,qCt,R,c)}) \qquad (2)$$

where Ct,x,q is the threshold cycle for gene target X for hormone q; Ct,x,c is the threshold cycle for gene target X for the corresponding control C; and Eff_R, Ct,R,q, and Ct,R,c are the corresponding

values for each reference gene. We refer to the quantity in ΔΔCt defined for each gene and each condition (time x hormonal treatment).

Five reference genes were chosen (HuCL00387C002, HuCL02526C001, HuCL05491C001, HuCL03058C001, and HuCL06237C001 available on www.heliagene.org) as they showed no response to drought stress in our previous study (Rengel, D. *et al.*, 2012b).

Validation of ABA responding genes identified from a global transcriptomic approach

In order to select robust genes regulated by drought in sunflower, we performed a global transcriptomic analysis where we compared the gene expression in control and ABA (10 µM) treated plants. We performed a t-test with a Bonferoni correction and selected genes with p-values<0.05 as sunflower genes differentially expressed after ABA application.

To validate the ABA-regulated sunflower gene set identified thanks to results of the global transcriptomic analysis, we studied and compare the expression of 226 Arabidopsis homologs. *Arabidopsis* homologs to all the sunflower genes in this study are BLAST reciprocal best hits of *Helianthus* ESTs and *Arabidopsis*.

The *Arabidopsis* data for all genes on the Affymetrix ATH1 microarray were collected from the AtGenExpress Consortium at http://www.weigelworld.org/resources/microarray/AtGenExpress/AtGe_Abiostress_gcRMA.zip/view . The authors followed through a kinetic of three points the transcriptomic response to abiotic stresses as cold, osmotic, salt, drought or heat stress in leaves. This study was of particular interest because its kinetic aspect brings more statistical power and avoids the issue of differences in kinetic parameters between sunflower and *Arabidopsis*. We performed an analysis of covariance (ANCOVA) to explain each *Arabidopsis* transcript expression by time (as a covariate) and treatment (as a factor) using Mathworks Matlab (version 7.13) Statistical toolbox (version 7.6). A Bonferroni method (Shaffer, 1995) was applied to identify significant effect of treatment, time and treatment x time interaction by setting the p-value threshold to 0.05/number of comparisons. Globally we found in *Arabidopsis* 3852 probesets (out of 22591 with an AGI annotation) significantly affected by these factors under these treatments in shoots.

Then, we compared the ABA regulated genes in sunflower having a AGI homologue (226) to this list and found 60 in common. We performed a hypergeometric test to determine if the *Arabidopsis* homologues of the sunflower ABA regulated genes are more likely to be differentially regulated during abiotic stress in *Arabidopsis*. For this, we used the Matlab function *hygecdf*.

Comparison of the drought GRN to Arabidopsis data

To compare the sunflower drought GRN to the model plant *Arabidospsis thaliana*, we collected expression data from the AtGenExpress Consortium (Goda *et al.*, 2008) (GEO accession: GSE39384 from AtGenExpress Consortium). This Arabidospsis data set includes seven hormonal treatments (Indole Acetic Acid, Cytokinin, Gibberellic Acid, Methyl Jasmonate, ABA, ACC, Brassinolide) and three time points (30 mn, 1h, 3h). We studied the expression of *Arabidopsis* homologs to the 145 sunflower genes used for network inference. Arabisopsis homologs to all the sunflower genes in this study are BLAST recipsocal best hits of *Helianthus* ESTs and *Arabidopsis*. We calculated the correlation between all selected *Arabidopsis* genes using Mathworks Matlab (version 7.13) Statistical toolbox (version 7.6). We applied an exact hypergeometric test (p=0.005) with the function *hygepdf* (Mathworks Matlab) to find if significant correlations between gene pairs were more frequent for pairs corresponding to the *Arabidopsis* homologs of the sunflower network edges.

GRN reconstruction

Problem definition

We address the problem of recovering gene regulatory networks (GRNs) from time series expression data. The targeted GRNs are directed graphs with *p* nodes, where each node represents a gene and an edge directed from one gene *i* to another gene *j* indicates that the expression of gene *j* is directly explained by that of gene *i*. However, an edge between *i* and *j* does not necessarily stand for a direct regulation in a biological sense, e.g. when a transcription factor regulates its target genes. We only consider unsigned edges; when gene *i* is connected to gene *j*, the former can be either an activator or a repressor of the latter.

In this paper, we assume that we have at our disposal an ensemble of *n* datasets D_k (k=1,…,n). We assume that these datasets are respectively obtained from *n* different perturbations of a system governed by a general regulatory network that specifies the plant response to an abiotic stress. Each perturbation corresponds to the induction of a specific hormone as described in the Methods Section. Each dataset D_k contains gene expression levels measured at T=7 different time points following a perturbation *k*:

$$D_k = \{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, \ldots, \mathbf{x}_{k,T}\}, \qquad (3)$$

where $\mathbf{x}_{k,t} \in \mathbb{R}^p, t = 1, \ldots, T$ is a vector containing the expression values of all *p* genes at the time point *t:*

$$\mathbf{x}_{k,t} = (x_{k,t}^1, x_{k,t}^2, \ldots, x_{k,t}^p)^\top, \qquad (4)$$

where $x^\top$ is the transpose of x

From these *n* datasets, our goal is to learn *n+1* GRNs: one GRN resulting from each perturbation and a global consensus GRN taking into account all the perturbations. Two complementary inference approaches were considered in this paper, respectively based on random forests and Gaussian graphical models. The results of these methods were then combined to achieve robust GRN predictions.

GRN inference with random forests

We extended a method called GENIE3 (Huynh-Thu *et al.*, 2010), which is based on Random Forests (RF, (Breiman, 2001)) and that was originally proposed for the inference of GRNs from steady-state expression data.

As in the original GENIE3 procedure, the problem of recovering a network of *p* genes is decomposed into *p* feature selection subproblems, where each of these subproblems consists in identifying the regulators of one gene of the network.

In the presence of time series data, we make the assumption that the expression of each gene of the network at time point *t+1* is a function of the expression of the other genes of the network at the preceding time point *t*. Denoting by the vector $\mathbf{x}_{k,t}^{-j}$ containing the expression values at time point *t* of all the genes except gene *j*, we thus write:

$$x_{k,t+1}^{j} = f_j(\mathbf{x}_{k,t}^{-j}) + \epsilon_{k,t}, \forall k, t, \qquad (5)$$

where $\epsilon_{k,t}$ is a random noise and functions f_j only exploit the expression in $\mathbf{X}^{-j}$ of the genes that directly regulate gene *j* in the underlying network. Recovering the regulatory links pointing to target gene *j* thus amounts to finding those genes whose expression at time *t* is predictive of the expression of the target gene at time *t+1*.

As in GENIE3, our procedure exploits feature importance scores derived from RF models to rank a candidate regulator *i* of gene *j* for perturbation *k* by its importance $w_{i,j}^{k}$.

First, a RF model is trained to predict the expression of the target gene at time *t+1* (i.e x_{t+1}^{j}) from the expression levels of all other genes at time *t* (i.e $\mathbf{x}_t^{-j}$).

Then, candidate regulators are ranked according to variable importance scores derived from the RF model. Importance scores are computed as the total variance reduction due to splits based on the corresponding regulator expression, averaged over all nodes and trees in the forest (Breiman, 2001).

Then a global ranking of all regulator-target gene edges is obtained by merging all individual target gene rankings with their associated importance scores and a network prediction is obtained by thresholding these scores.

RF importance scores are not statistically interpretable, which makes difficult the determination of an importance threshold to obtain a single and interpretable network prediction. We therefore propose to replace these scores by a new score that can be interpreted statistically.

To compute this score, we add to the dataset an artificial *p+1*th random gene, whose expression values are obtained by randomly permuting the *n* x *T* expression values of a gene randomly selected among the *p* original genes (making the new gene uncorrelated to all other genes). We then run the RF learning procedure described above to obtain a ranking of the GRN edges, including edges involving the random gene.

We repeat the experiment 1000 times and take as score for a GRN edge, the proportion of the 1000 rankings where this edge was ranked above all the edges involving the random gene. The resulting edge score is then interpreted as the probability that this edge is ranked by the RF model at a higher level than a spurious edge (the higher, the better).

The previous procedure can be applied separately on each time series D_k, k=1,...,n or on the union of all time series to obtain respectively the $ n $ perturbation-specific GRNs and the global consensus GRN. However, each individual time series being rather small and expecting only limited differences between these networks, we preferred the following procedure to obtain the *n* perturbation-specific GRNs: first, RF models for all genes are trained on the union of all time series. Then, pertubation-specific importance scores are obtained by re-propagating the instances from each dataset D_k separately into these RF models and re-computing variable importance scores only from these instances. The whole procedure is summarized in Algorithm 1.

Algorithm 1 GENIE3 for the inference of GRNs from multiple time series

Input: n time series expression datasets D_k, $k = 1, \ldots, n$.

Output: a network $G_k (k = 1, \ldots, n)$ for each time series dataset and a global consensus network G_{n+1}.

- For $m = 1$ to $M = 1000$:
  - Create a new random gene $\mathbf{x}^{p+1}$, by randomly choosing a gene among the p genes and randomly permuting its $n \times T$ expression values. Add this random gene to the other genes in the dataset.

- For gene $j = 1$ to $p + 1$:
  - Generate the learning sample of input-ouput pairs for gene j:

$$LS^j = LS_1^j \cup LS_2^j \cup \ldots \cup LS_n^j, \tag{6}$$

where

$$LS_k^j = \{(\mathbf{x}_{k,t}^{-j}, x_{k,t+1}^j), t = 1, \ldots, T-1\}, k = 1, \ldots, n. \tag{7}$$

  - Learn a Random Forest model from LS^j.
  - Use subsample $LS_k^j (\forall k = 1, \ldots, n)$ and the whole learning sample LS^j to respectively compute importance scores called confidence weights $w_{i,j}^{k,m}$ and $w_{i,j}^{n+1,m}$ for each input gene i (including the random gene).
- For $k = 1$ to $n + 1$, take $w_{rand}^{k,m}$ as the maximum weight among the weights of the edges directed from or towards the random gene:

$$w_{rand}^{k,m} = \max(w_{p+1,1}^{k,m}, \ldots, w_{p+1,p}^{k,m}, w_{1,p+1}^{k,m}, \ldots, w_{p+1,p+1}^{k,m}). \tag{8}$$

- For $k = 1$ to $n + 1$:
  - For each edge $i \rightarrow j$, compute the proportion of iterations $s_{i,j}^k$ where its weight is higher than $w_{rand}^{k,m}$:

$$s_{i,j}^k = \frac{1}{M} \times \#\{m : w_{i,j}^{k,m} > w_{rand}^{k,m}\}. \tag{9}$$

  - Choose a threshold on $s_{i,j}^k$ to obtain network G_k.

Comparison of two random forest algorithms: authorizing and not authorizing auto-loops prediction

The Random Forest algorithm used does not allow the gene *j* at time *t* to be a predictor of this same gene *j* at time *t+1*, therefore excluding auto-loops from the model prediction. In order to test if this exclusion is justified, we ran an additional test on a new simulated data set containing 100 genes related by 209 edges including 10 auto-loops. We tested both the algorithm used in our analyses (RF-NAL, for Random Forest No Auto-Loops) and a modified version which authorizes the discovery of auto-regulations (RF-AL for Random Forest Auto-Loops). We compared the capacity of both models to recover true edges and to avoid false positive by comparing the areas under the precision-recall curves. The areas are very similar (0.108 for RF-NAL and 0.110 for RF-AL), so the two methods perform equally well.

More specifically, we focused on a fixed number of edges by considering only edges which have a score no smaller than 1; it means that they were always ranked above the artificially introduced random edges. RF-NAL allowed the prediction of 422 edges, and RF-AL predicted 425 edges including 3 auto-loops. The two methods predicted 378 edges in common, 44 edges were found only by RF-NAL and among them only two were true positives. Among the edges predicted only by RF-AL algorithm, the 3 auto-loops were true positives, however, the 44 others were all false positive. In light of these results, we observe that (i) global performance are the same for RF-NAL and RF-AL, (ii)

the vast majority of edges predicted by either version of RF are common to both of them and (iii) few auto-loops are actually retrieved by RF-AL when producing a reasonable amount of edges.
Since we are more interested in edges linking different nodes, we believe that the choice of the RF-NAL algorithm is more suited regarding the biological question that focuses on relationships between genes (estimated by the network topological parameters) and sunflower evolution.

Inferring multiple GRN structures with Gaussian Graphical Models

We used here a Gaussian Graphical Modelling (GGM), a widely used statistical tool for the reconstruction of networks of regulatory relationships between genes. The main difficulty stands in the high-dimensionality of the data set: the number of variables (genes) exceeds the number of samples (combination of treatment x time point). If samples are considered independent, each of them is considered as the observation of a multivariate Gaussian random variable whose dimension is the number of considered genes in the network. Intrinsic dependencies between genes are encoded in the associated covariance matrix or more precisely in the inverse of this matrix, the precision matrix: non-zero entries of the precision matrix fully determine non independent couples of variables in the network; notice that they also exactly determine non-zero regression parameters of the regression of each gene against all other genes (Whittaker, 1990); (COX & WERMUTH, 1993). Because of the high-dimensionality of the data set at hand, we chose to rely on the lasso (Tibshirani, 1996), a widely used ℓ_1 -penalization technique, which basically assumes sparsity of the network topology.
More precisely, using notations previously introduced, we first assumed a first-order auto-regressive model on centred data in each condition *k*:

$$\mathbf{x}_{k,t} = \mathbf{x}_{k,t-1} A_k + \epsilon_{k,t},$$

where matrix A_k contains the effects of all genes at time *t-1* onto genes at time *t*. This modeling is close to that of RF of Equation (1).
In fact,

$$(A_k)_{i,j} = \frac{cov(x^j_{k,t}, x^i_{k,t-1} \mid \mathbf{x}^{-i}_{k,t-1})}{Var(x_{k,t-1} \mid \mathbf{x}^{-i}_{k,t-1})}$$

If we treat the case of one hormonal treatment and omit subscript *k*, maximizing the log-likelihood of the model is equivalent to the following optimization problem:

$$\max_A \left\{ \mathrm{Tr}(V^\top A) - 1/2 \mathrm{Tr}(A^\top S A) \right\},$$

and the solution (maximum likelihood estimator) is given by $\hat{A}^{MLE} = S^{-1}V$, where we denoted by *S* the empirical variance-covariance matrix and by *V* the empirical temporal covariance matrix (Charbonnier *et al.*, 2010) and we omitted subscript *k* which referred to the hormonal treatment.

An ℓ_1 -penalty on matrix *A* which encodes non-zero coefficient of the auto-regressive model can be used to circumvent the high-dimensionality of the problem (*S* being not invertible) under sparsity assumptions: matrix *A* has few non-zero elements with a minimum intensity (Verzelen, 2012). The multi-perturbation version of the likelihood of Equation (10) includes this ℓ_1-penalization.

More specially to choose the ℓ_1 penalty parameter lambda, we used a similar algorithm to the LARS (Efron *et al.*, 2004). It gives the model estimate for all values of the penalty parameter: from 0, no penalty, the solution is then the ordinary least square estimate, to infinity, which leads to a void model. In fact, the number of different model estimates is finite, so the number of computation is finite. Moreover, the LARS algorithm proposes a solution path, which make the computation of the different estimates which correspond to different penalty parameter values very efficient. From this comprehensive list of model estimates, we decided to select the penalty level which leads to a number of edges of the order of magnitude of the number of nodes in the network, a situation which is often encountered in sparse network settings. Since this number varies from one network to another, we arbitrarily fixed it to be as close as possible from 200.

In our framework, 9 different perturbations, which correspond to 9 different matrices $(A_k)_{k=1,\dots,9}$, have to be considered. If we ignored the relationships between the different hormonal treatments, we would simply optimize a problem which would be the sum of 9 problems similar to the one of the previous paragraph. We instead combined the temporal approach of (Charbonnier *et al.*, 2010) to the multiple graph structure inference scheme of (Chiquet *et al.*, 2011), which is written for independent identically distributed (iid) Gaussian graphical models. We used a so-called ''intertwined'' estimation of matrices A_k's. It renders the model parameter estimation over different hormonal conditions not separable anymore. More precisely, the objective function (the log-likelihood) is slightly modified and instead of using the 9 matrices V_k and S_k separately, we used a convex combination that accounts for a part which is specific to the hormonal treatment and the other part which is a mean of each matrix over all conditions. The objective function to be maximized can be expressed as

$$\max_{A_1 \dots A_9} \sum_{k=1}^{9} \left\{ \mathrm{Tr}(\tilde{V}_k{}^{\top} A_k) - 1/2\mathrm{Tr}(A_k{}^{\top} \tilde{S}_k A_k) - \lambda \mid A_k \mid_{\ell_1} \right\}, \qquad (10)$$

Where $\tilde{V}_k = \alpha V_k + (1-\alpha)\frac{\sum_{k'=1}^{9} V_{k'}}{n}$

and $\tilde{S}_k = \alpha S_k + (1-\alpha)\frac{\sum_{k'=1}^{9} V_{k'}}{n}$.

So the coupling between the A_k's which translates how the 9 networks are related to each other is made through the fitting term (likelihood of the data), not through the penalty term. Each empirical covariance S_k is being replaced by a mixture $\tilde{S}_k$ of a covariance specific to the perturbation and to a pooled estimate $\frac{\sum_{k'=1}^{9} V_{k'}}{n}$. Similarly, each empirical temporal covariance matrix V_k is being replaced by a mixture $\tilde{V}_k$ of a temporal covariance specific to the perturbation and a pooled estimate $\frac{\sum_{k'=1}^{9} V_{k'}}{n}$. Note that approaches exist in which the coupling is made cooperatively through the penalty term: model with too many different edges (possibly accounting for edge sign) between different treatments are heavily penalized.

The mixing parameter α of the convex combination is arbitrarily set to 1/2 in our experiments; if it were equal to 1, all data sets would be pooled as a single one and if it were set to 0, the estimate of the matrices corresponding to each hormonal treatment would be independent. We restricted the number of edges in each network to be no more than 200 for computational reasons.

We used the R package SIMONe (Chiquet *et al.*, 2009) to obtain estimation of matrices A_k's. We made the prediction over edges from matrices A_k more robust by applying the method we just described 200 times on bootstrapped version of the samples, in the spirit of the bootstrap lasso introduced in (Bach, 2008). A time series length was first uniformly chosen between 3 and 7 and then time points were picked up at random for each treatment, with the same time series length preserving the time ordering, so that different response times to different hormones could be considered. An edge was identified as being significant when it was predicted over 20% of the bootstrapped runs of the algorithm. The rationale behind this heuristics is that we preferred to focus on edges that appear in most bootstrapped repeats of the algorithms but in possibly varied contexts for each edge.

We summarized our Gaussian Graphical Model approach in Algorithm (2).

Algorithm 2 BootGGM for the inference of GRNs from multiple time series

Input: n time series expression datasets D_k, $k = 1, \ldots, n$.

Output: a network $G_k (k = 1, \ldots, n)$ for each time series dataset and a global consensus network G_{n+1}.

- For $m = 1$ to $N_{boot} = 200$
  - For each perturbation k, create bootstrapped version $\mathbf{x}_k$ of D_k by (i) uniformly drawing a time series length $ts.l$ (3, 4, 5, 6 or 7), (ii) and select $ts.l$ sorted times at random from the T available times to be used for each hormonal treatment. Each perturbated bootstrapped expression matrix consists in a $ts.l \times p$ matrix $\mathbf{x}_k$, whose columns are the expression values of genes in picked up conditions.
  - Learn adjacent matrices $(A_k^m)_{k=1\ldots n}$ using the intertwined estimation of the multiple time-series l_1-penalized GGM approach.
- Create $(A_k)_{k=1\ldots n}$ such that $\forall k \in \{1 \ldots n\}$, $A_k[i,j] = 0$ if $A_k^m[i,j] \neq 0$ for more than 20% of all m. A score for edges $i \to j$ is then the ratio (between 0.2 and 1 then) of its occurrence among $(A_k^m[i,j] \neq 0)_{m=1\ldots 200}$.
- We define the matrix A_{n+1} as the union of all hormonal networks. We name it Global Network for the GGM method.

GGM and RF model validation in a simulated framework

To test the accuracy of both our models and of the implementation we used, we built a simulated data set, which shares the specific features of the dynamical response to hormonal treatments data.

First, we fixed the topology of a gene network. The network is oriented, is signed and has a scale-free topology and was specifically built to assess GRN inference performances (Mendes *et al.*, 2003). It comprises 100 nodes (corresponding to genes) and 200 edges (corresponding to gene direct interactions): the density of the (undirected) network is thus approximately
equal to 4%. The term 'scale free' indicates that the network has been generated from a preferential attachment model, as described in (Barabasi & Albert, 1999).

Next, we generated 9 'child' networks from this reference network by randomly perturbing it: edges could be removed, added or reversed. 3 child networks had a 10% low perturbation level, 3 other child network had a 20% moderate perturbation level, while the last 3 child network had a 30% considerable perturbation level as compared to the 'parent' network. Hence these networks can differ up to 60% of their edges. They each model the hormone-specific GRN which could be relatively similar to a central 'stress' GRN or which could be quite different from each other.

Lastly, for each of these 9 child networks, we generated a random partial correlation matrix, whose non-zero entries exactly encode the 200 edges of the network. These partial correlation matrices are then used to simulate 9 time-series (auto-regressive model of order 1) gene expression 'ddCt-like' data. The variance was assumed to be the same for all genes.

The goal then was to reconstruct GRN from these 9 dynamical gene expression data sets using the two methods which were described above, GGM and RF. The advantage is that results could be quantitatively assessed in terms of precision, the ability of a method to produce correct edges among its predictions and of recall, the ability of the method to retrieve edges of the network used to generate the data.

As an example, we give figures which correspond to one of the moderately perturbed networks. Similar results are obtained for all networks with a slight degradation when the perturbation level is higher. Additionally, the numbers of total, specific and shared predicted edges can vary from one network to another, without it is related to the perturbation level. GGM lead to a 92 edge network and RF infers a 94 edges network for one of the child network (which had 200 correct edges to be predicted). GGM achieves a recall of 29% while RF achieves a recall of 18% at precision levels of respectively 65% and 67%. Edge prediction is hence quite good but the total network coverage is relatively poor. We prefer to produce reliable edges and accept to potentially miss some correct interactions weakly supported by the data.

Another interesting aspect, which confirmed previous observations in different settings, is that both approaches lead to good quality predictions with limited overlap (Marbach *et al.*, 2012); (Allouche *et al.*, 2013): 33 edges are jointly predicted by GGM and RF. The very interesting point here is that 32 out of these 33 edges in common to GGM and RF are true positives. Hence the intersection of both networks lead to a 97% precision , 16% recall network. A small loss in prediction coverage allows us to produce very reliable predictions. We concede that these figures are probably over-optimistic since they are only valid on simulated data. However, the trends presented here are very consistent with those obtained on the real data sets and make us confident about the networks which were produced when analyzing these data sets.

Appendix IV.2: Results of t-test demonstrating the differential expression of genes upon application of 10 µM ABA.

This appendix is available on line at the address:

https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents the Results of t-test demonstrating the differential expression of genes upon application of 10 µM ABA (Excel File)

Appendix IV.3: Results of ANCOVA showing the validation of the ABA genes dataset.

This appendix is available on line at the address:

https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents the results of ANCOVA (Excel File)

Appendix IV.4: Description of the genes selected for GRN inference.

This appendix is available on line at the address:

https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents selected genes used in the GRN inference and their origin (literature and sunflower transcriptomic experiments) (Excel File)

Appendix IV.5: Raw gene expressions.

This appendix is available on line at the address:

https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents raw expressions of selected genes for the GRN inference (Excel File)

Appendix IV.6: Gene expressions after log transformation and missing data imputation.

This appendix is available on line at the address:

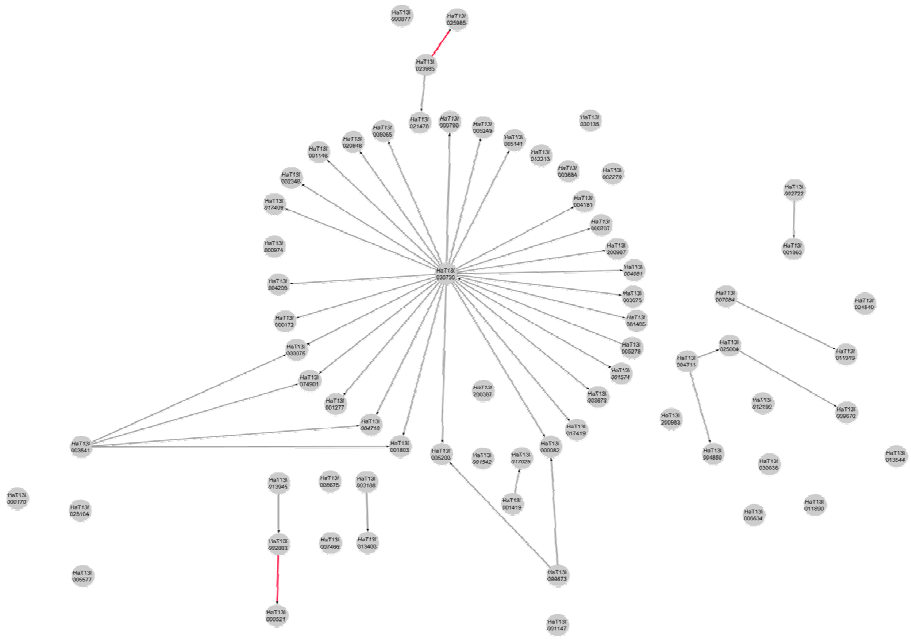
https://www.heliagene.org/cgi/heliagene.cgi?__wb_session=WBoOPfh9&__wb_main_menu=Publications&__wb_function=PhDThesis

It presents expressions of selected genes for the GRN inference after log transformation and missing data imputation. These values are directly used for the network inference (Excel File).

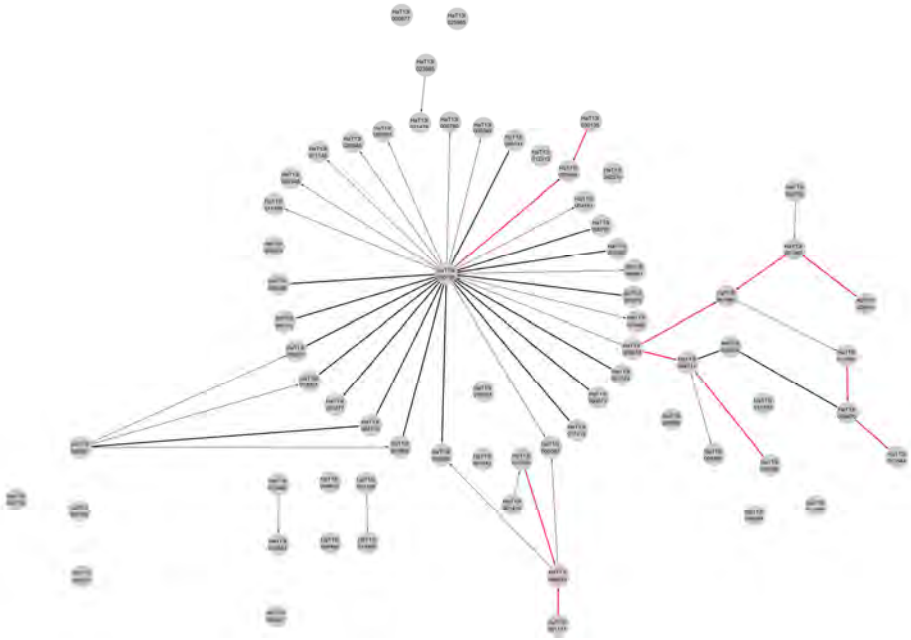
Appendix IV.7: Hormonal network representations

Grey circles represents genes. Red edges represent specific hormonal edges. Grey edges represent edges inferred with the global dataset. Black edges represent edges inferred with both the global and the specified hormonal dataset.

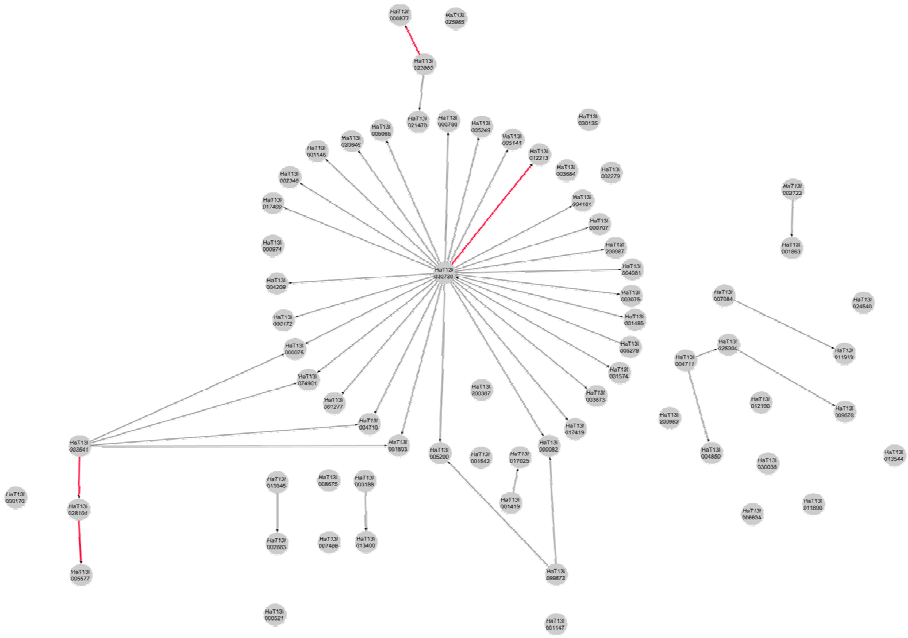
a) ABA and global networks



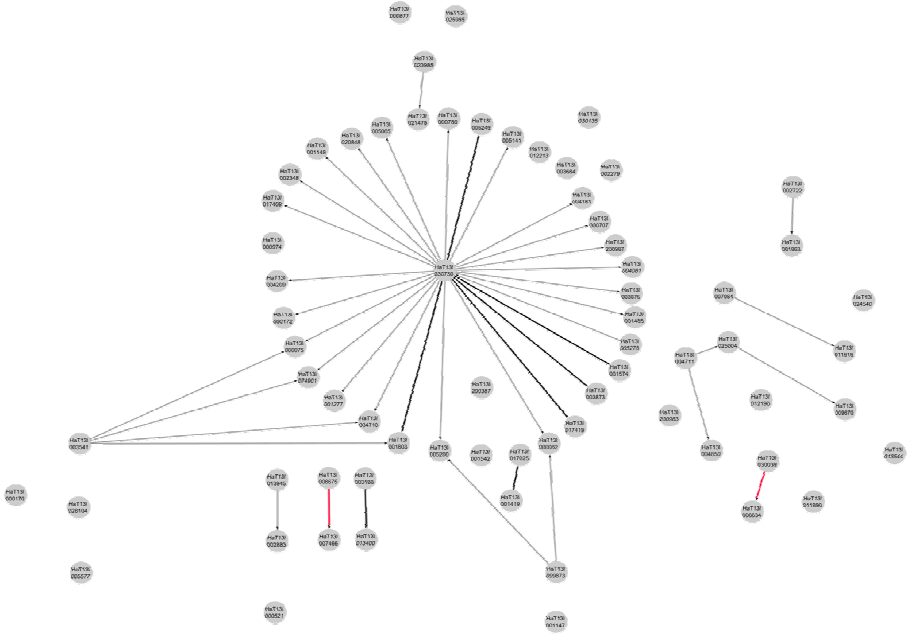
b) ACC and global networks



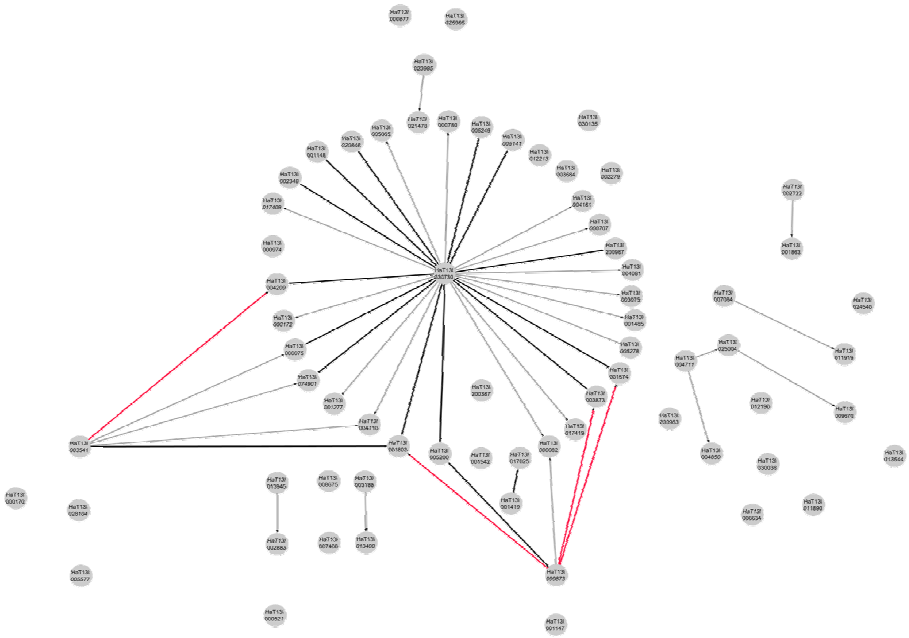
c) Bras and global networks



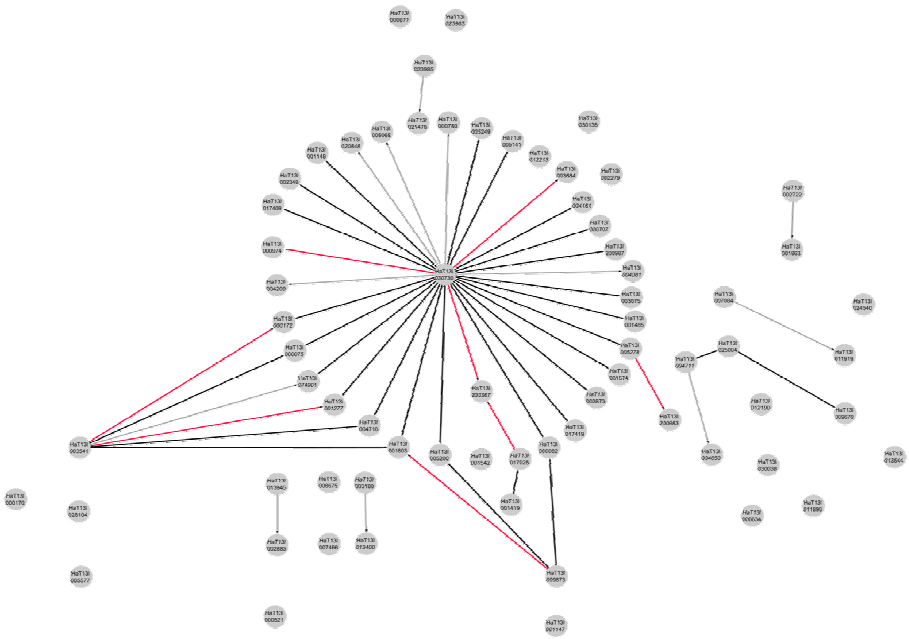
d) GA3 and global networks



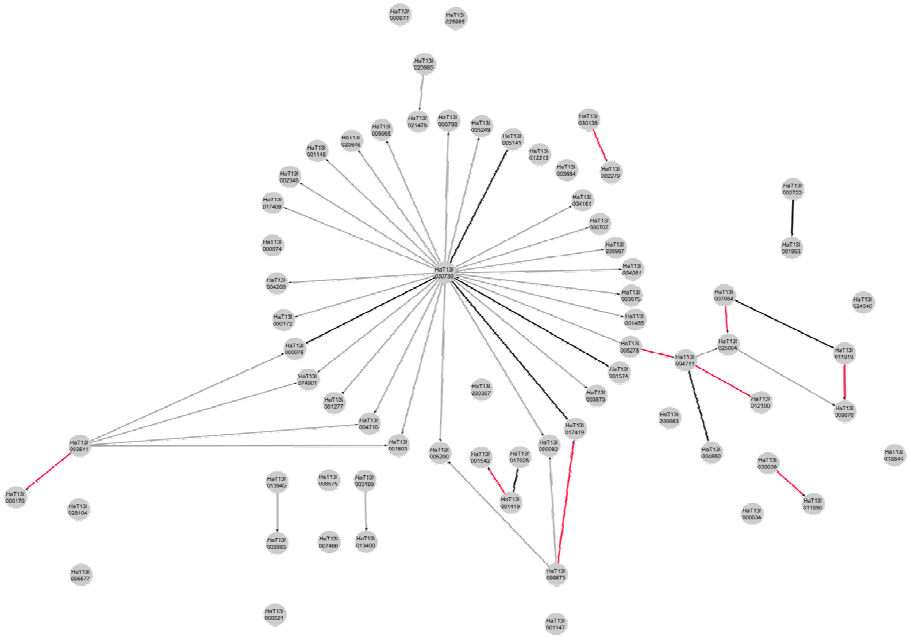
e) IAA and global networks



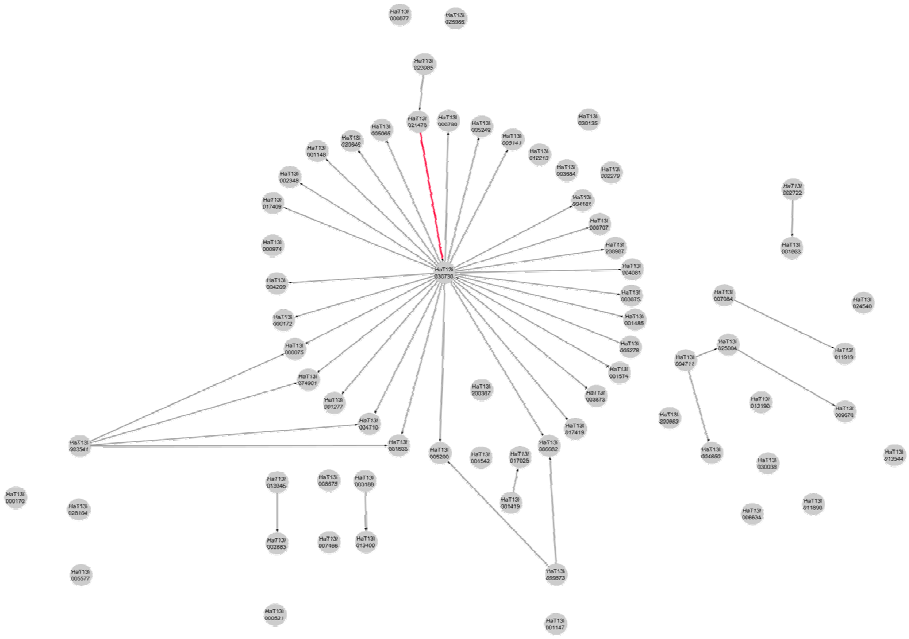
f) Kinetin and global networks



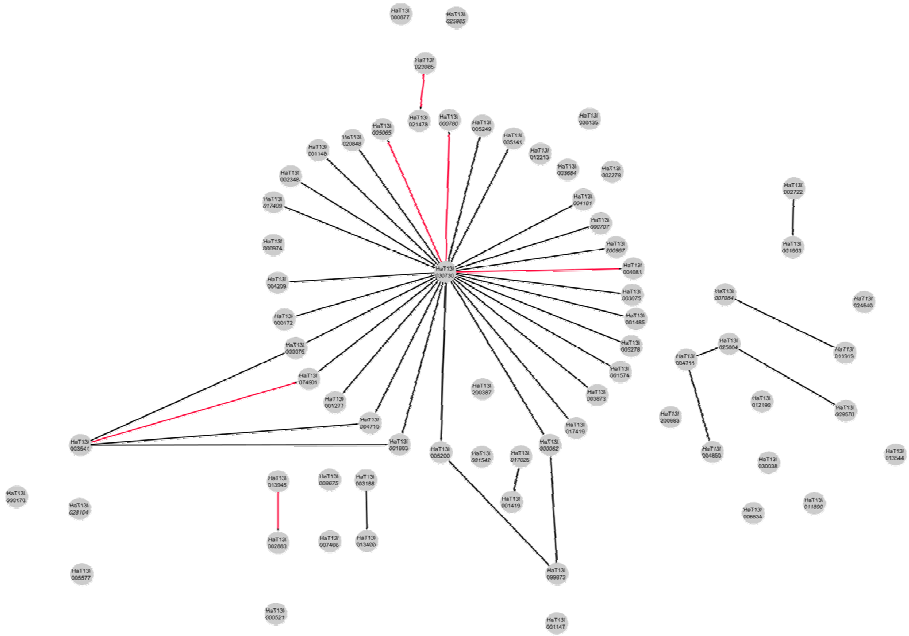
g) MeJA and global networks



h) Strigolactone and global networks



i) Global network



Appendix IV.7: Canonical analysis complete results

Argophyllus Subset

Canonical coefficients of correlation (rho)

0.6721059	0.5243957

xcoef

	Component1	Component2
ASPL	-0.4474753	-0.105491
EdgeCount	0.0124323	-0.1131035

ycoef

	Component1	Component2	Component3	Component4
fstArgvsElite	3.05982547	2.9540319	0.4602359	-1.8484929
fstArgvsLandrace	-2.77771813	-5.2569105	-1.6538927	-0.4080627
fstArgvsPet	-0.01483693	-0.7577599	1.6324079	0.3938518
fstArgvsWild	1.00345336	2.1867725	0.5463695	2.8968148

Petiolaris Subset

Canonical coefficients of correlation (rho)

0.4930524	0.3693907

xcoef

	Component1	Component2
ASPL	-0.32715772	0.32300211
EdgeCount	-0.08920463	-0.07063646

ycoef

	Component1	Component2	Component3	Component4
fstArgvsPet	-1.0040332	-1.8153277	0.5135624	0.06727046
fstElitevsPet	0.2352018	0.7120162	5.0738005	-1.19216826
fstLandracevsPet	4.8594038	-2.015399	-6.1378372	1.89442128
fstPetvsWild	-3.771505	1.7248486	-0.2025232	-1.93987094

Wild Subset

Canonical coefficients of correlation (rho)

0.7282873	0.2922409

xcoef

	Component1	Component2
ASPL	-0.38242723	-0.2551704
EdgeCount	0.05116187	-0.1016338

ycoef

	Component1	Component2	Component3	Component4
fstArgvsWild	1.3544178	-0.5737768	-0.7474485	0.2208294
fstElitevsWild	-1.7916661	2.1437681	-1.6415964	0.3512176
fstLandracevsWild	3.0080055	0.9725076	2.9611102	-0.3469812
fstPetvsWild	0.5266618	-0.5729598	-0.0935532	-1.5043243

Landrace Subset

Canonical coefficients of correlation (rho)

0.9757625	0.2986421

xcoef

	Component1	Component2
ASPL	-0.379509	-0.2594907
EdgeCount	0.0523111	-0.1010471

ycoef

	Component1	Component2	Component3	Component4
fstArgvsLandrace	0.4958675	-0.7035161	-0.8604868	-0.7221756
fstElitevsLandrace	-2.1322542	0.239832	-1.5931857	-0.9086869
fstLandracevsPet	0.4605054	0.7421679	1.1422655	-1.0628351
fstLandracevsWild	0.7258354	2.7453314	-1.9398949	0.9560778

Elite Subset

Canonical coefficients of correlation (rho)

0.946278	0.2805252

xcoef

	Component1	Component2
ASPL	-0.38304268	-0.2542456
EdgeCount	0.05091614	-0.1017572

ycoef

	Component1	Component2	Component3	Component4
fstArgvsElite	0.544304	-0.5846516	-0.7700024	-1.0952447
fstElitevsLandrace	-2.1462581	-0.4355321	-1.0276413	-1.7020776
fstElitevsPet	0.2666276	0.5114767	1.2074469	-0.7237683
fstElitevsWild	0.2503657	2.1229694	-0.9414516	1.1935652

Author: Gwenaëlle Marchand

Title: Gene regulatory networks involved in drought stress responses : identification, genetic control and variability in cultivated sunflower, *Helianthus annuus* and its relatives.

PhD supervisors: Patrick Vincourt, Nicolas Langlade

Defended: 6th June 2014

Abstract:

Drought is a major stress that affects growth, physiology and therefore yield of crops as sunflower. To become more tolerant, plants develop complex morpho-physiological responses. Various genes interacting between them and with the environment are involved in the genetic control of those responses. They form together a gene regulatory network (GRN). Here, we focused on this drought GRN, its different gene groups and their interactions in the cultivated sunflower. First, we highlighted three genes reflecting the environmental signal. From their expression we built a plant water status biomarker. Then, through an association study we built the GRN connecting drought responsive genes and we deciphered their genetic control. Finally, thanks to a systems biology approach we inferred the GRN linking regulatory and drought responsive genes. Studying this network, we examined how it could drive phenotypic changes and how it was related to *Helianthus* evolution and sunflower breeding.

Keywords: drought, sunflower, transcriptomic, gene regulatory network, *Helianthus*, association mapping, systems biology, biomarker, plant physiology.

Scientific field: genetics and interaction plant-environment

Laboratory: Laboratoire des interactions Plantes Microorganismes, UMR INRA/CNRS 441/2594, Chemin de Borde Rouge, BP52627 Castanet-Tolosan, France

Auteur : Gwenaëlle Marchand

Titre : Etude des réseaux de régulation impliqués dans la réponse au stress hydrique : caractérisation, contrôle génétique et rôle au cours de l'évolution du tournesol cultivé, *Helianthus annuus*.

Directeurs de thèse : Patrick Vincourt, Nicolas Langlade

Soutenance : Vendredi 6 juin 2014

Résumé :

La sécheresse affecte le rendement des plantes de grande culture comme le tournesol. Ces plantes développent des réponses morpho-physiologiques pour améliorer leur tolérance au manque d'eau. De nombreux gènes formant un réseau de régulation (GRN) contribuent à un contrôle génétique complexe de ces réponses. Le travail présenté étudie ce réseau, ses différents gènes et leurs interactions chez le tournesol. Tout d'abord, nous avons mis en évidence trois gènes récepteurs du signal environnemental afin de construire un biomarqueur du statut hydrique. Puis, par une étude d'association, nous avons reconstruit le GRN reliant les gènes de réponse au stress et déchiffré leur contrôle génétique. Enfin, par une approche de biologie des systèmes, nous avons inféré le GRN groupant des gènes de régulation et de réponse. Cette étude nous a permis d'identifier des mécanismes majeurs de tolérance à la sécheresse chez le tournesol, ainsi que le rôle de ce réseau dans l'évolution du genre *Helianthus.*

Mots clés : stress hydrique, tournesol, *Helianthus* annuus, transcriptomique, réseau de régulation d'expression de gènes, génétique d'association, biologie des systèmes, biomarqueur, physiologie végétale.

Discipline : Génétique et intéraction plante-environnement

Adresse : Laboratoire des interactions Plantes Microorganismes, UMR INRA/CNRS 441/2594, Chemin de Borde Rouge, BP52627 Castanet-Tolosan, France