THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En biologie Intégrative, Diversité et Amélioration des Plantes

École doctorale 584 GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Unité de recherche IPSiM – Institut des Sciences des Plantes de Montpellier

Inférence statistique des réseaux de régulation de gènes chez *Arabidopsis thaliana* en réponse à l'élévation des teneurs en CO_2 atmosphérique

Présentée par Océane CASSAN

Le 13 décembre 2022

Sous la direction de Antoine MARTIN et Sophie LÈBRE

Devant le jury composé de

Andréa RAU, Chargée de recherche, INRAE Jouy en Josas	Rapporteure
Étienne DELANNOY, Chargé de recherche, INRAE Paris Saclay	Rapporteur
Céline MASCLAUX-DAUBRESSE, Directrice de recherche, INRAE Versailles-Grignon	Examinatrice
Nathalie VIALANEIX, Directrice de recherche, INRAE Toulouse	Examinatrice
Philippe NACRY, Directeur de recherche, INRAE Montpellier	Président du jury
Antoine MARTIN, Chargé de recherche, CNRS Montpellier	Directeur de thèse

UNIVERSITÉ DE MONTPELLIER

"Even if the open windows of science at first make us shiver after the cosy indoor warmth of traditional humanizing myths, in the end the fresh air brings vigour, and the great spaces have a splendour of their own."

Bertrand Russell, "What I believe", 1925

Contents

Acknowledgements

I would like to start by acknowledging all the IPSiM members for their friendly attitude that made me feel very welcome in this lab.

I express my gratitude to Lauren Castaings, Freddy Barnèche, Laurent Laplaze, and Julien Chiquet for being part of my follow-up committee and for giving me valuable input in the course of this project. I also thank Andréa Rau, Etienne Delannoy, Céline Masclaux-Daubresse, Nathalie Vialaneix, Philippe Nacry and Vincent Ségura for accepting to review the work presented in this manuscript.

From the Plasticity team, I thank Benjamin Péret for giving access to his server, and Alexandre Soriano for his contributions to DIANE and help in various bioinformatic topics.

I am grateful to my LIRMM colleagues from the MAB team : Laurent, Charles, Mathys, Raphaël, and Mathilde, who welcomed me for lab meetings and foodtrucks in the last year. In particular, thank you Laurent for supervising my first steps into research and for keeping in touch ever since.

The SIRENE team has been a fantastic place to learn, grow, and chat on a daily basis. Laurence, Sandra, Alain, Liên, Cécile, Antoine, Jossia, I think you form a united and skillfull team that every student or researcher would like to join. Alexandre, Timothy, Paul, Salomé, Charlotte, Léna, and all other non permamnent staff who visit(ed) the team, you have been amazing co-workers and friends. I wish you all the best for the years to come.

Sophie, Antoine, it is hard to precisely say how much you contributed to my personal and scientific journey in the last years, but I suspect this is a lot. And I wouldn't wish otherwise.

Sophie, I sincerely enjoyed all our discussions and your excellent scientific guidance. Your humane, considerate approach to navigating the world of academia has been nothing short of motivational.

Antoine, I am sincerely grateful that you taught me plant biology from ground zero, and trusted me to join you in your research. I can't thank you enough for being this wise, respectful, benevolent and reliable support during those three years.

Finally, I thank you Adrien for being my day to day partner in life. In addition to all the practical help, your passion for research and knowledge fueled mine more than once. You inspire me to be the best scientist, and person, I can be.

List of Abbreviations

aCO_2	**a**mbient CO_2 level ($\sim$ 420 ppm)
ATAC-Seq	**A**ssay for **T**ransposase-**A**ccessible **C**hromatin **seq**uencing
bRF	**b**aised **R**andom **F**orests
AUPR	**A**rea **U**nder the **P**recision-**R**ecall curve
C	Carbon
ChIP-Seq	**C**hromatin *I*mmunoprecipitation *Seq*uencing
DAP-Seq	**D**NA **A**ffinity **P**urification **seq**uencing
DEA	**D**ifferential **E**xpression **A**naysis
DIANE	**D**ashboard for the **I**nference and **A**nalysis of **N**etworks from **E**xpression data
eCO_2	**e**levated CO_2 level (900 ppm)
eQTL	**e**xpression **Q**uantitative **T**rait **L**ocus
Fe	Iron
GGM	**G**aussian **G**raphical **M**odel
GHG	**G**reen **H**ouse **G**as
GRN	**G**ene **R**egulatory **N**etwork
IPCC	**I**nternational **P**anel on **C**limate **C**hange
kb	**k**ilo-**b**ase
LASSO	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
LASSO-D3S	**LASSO** with **D**ifferential **S**hrinkage and **S**tability **S**election
LMM	**L**inear **M**ixed **M**odel
MDA	**M**ean **D**ecrease in **A**ccuracy
MDI	**M**ean **D**ecrease in **I**mpurity
Mg	Magnesium
mM	**m**illi**M**olar
Mn	Manganese
mRNA	**m**essenger **r**ibo**n**ucleic **a**cid
MSE	**M**ean **S**quared **E**rror
N	Nitrogen
Na	Sodium
NH4$_4^+$	Ammonium
NO$_3^-$	Nitrate
NRT	**N**itrate **R**oot **T**ransporter
OOB	**O**ut **O**f **B**ag
PCA	**P**rincipal **C**omponent **A**nalysis
ppm	**p**arts-**p**er-**m**illion
qq-plot	**q**uantile-**q**uantile **plot**
TARGET	Transient Transformation System for Genome-Wide TF Target Discovery
TF	**T**ranscription **F**actor
TFBS	**T**ranscription **F**actor **B**inding **S**ite
RCP	**R**epresentative **C**oncentration **P**athway
SNP	**S**ingle **N**ucleotide **P**olymorphism
Zn	Zinc

Abstract

Statistical inference of the Gene Regulatory Networks in *Arabidopsis thaliana* under rising atmospheric CO_2 levels

Human activity is causing an elevation of CO_2 levels in the atmosphere, that are expected to rise from 420 ppm to approximately 1000 ppm by the end of the century. C3 plants, a major part of cultivated crops, are particularly affected by the rise of CO_2 levels. Even though a stimulation of biomass production is expected under elevated CO_2 (eCO_2), this gain is met with a marked depletion of the plant mineral composition and an especially strong decline in nitrogen (N) content. This poses a major threat to crop quality and human nutrition, that we propose to start addressing through systems biology approaches. Promising hypotheses to explain this decline invoke a disruption of signalling pathways associated to N uptake and assimilation, motivating the investigation at the genomic scale of gene expression reprogramming in the roots of the model plant *Arabidopsis thaliana* under eCO_2. To uncover the unknown regulators orchestrating such networks, we developed statistical methods for Gene Regulatory Network (GRN) inference, a challenging task hindered by high dimension and the scarcity of ground truth networks. Modelling transcriptional dependencies from gene expression data can be performed by regression-based techniques assuming that the expression variations of regulator genes hold descriptive and predictive power over the expression variations of their targets. We propose two novel approaches : (i) an extension of a Random Forest-based method, GENIE3, via permutation procedures assessing the significance of regulatory interactions that we include within a complete suite for GRN inference, and (ii) two integrative GRN inference methods based on Random Forests and sparse linear regression with stability selection, integrating Transcription Factor Binding Sites (TFBSs) with gene expression. We benchmark those methods against experimental gold standards, and show that they improve the biological relevance of inferred GRNs in *Arabidopsis thaliana*. We applied the first inference approach to a combinatorial transcriptomic dataset of root tissues under contrasted CO_2 levels and nutritional conditions, and the second to the roots of plants exposed to a gradient of CO_2 concentrations. The inferred GRNs provided candidate genes for the control of this response, and we demonstrate that some of them regulate growth stimulation under eCO_2 without penalizing shoot nutrient content. Overall, our results indicate that key nitrate and iron nutrition genes and their known regulators are misregulated by rising CO_2, and that pathways associated to high affinity nitrate transport systems are especially unfavorably altered. The last objective of this work was to leverage natural genetic variability to identify genes controlling the ionome response to eCO_2. We confirmed a mineral content decline in three populations of Arabidopsis at different geographic scales, and showed that the variability in this response can be explained by genetic determinants in the world-wide panel via linear mixed models. We put forward another set of candidate genes, highly associated to iron, N and zinc depletion in the shoots under eCO_2 that pave the way for designing plants with sustainable nutritional value for the near future.

Chapter 1

Introduction

1.1 The increase of CO_2 levels in the atmosphere reduces crops nutritional value

1.1.1 Green House Gas emissions in the context of climate change

Human activity is unprecedentedly affecting the Earth's climate. It has now been established with extremely high confidence by the International Panel on Climate Change (IPCC) that Green House Gases (GHG) emitted by the activities of the developed world are driving climate change by altering our atmosphere composition. Several GHG are currently being emitted, each of them in varying quantities, activity sectors, and with different potential impact on climate. Figure 1.1 shows the quantities of each type of GHG released in the course of the last three decades, and highlights the predominant contribution of CO_2 emissions from fossil fuel, industry, and land use to the global increase.

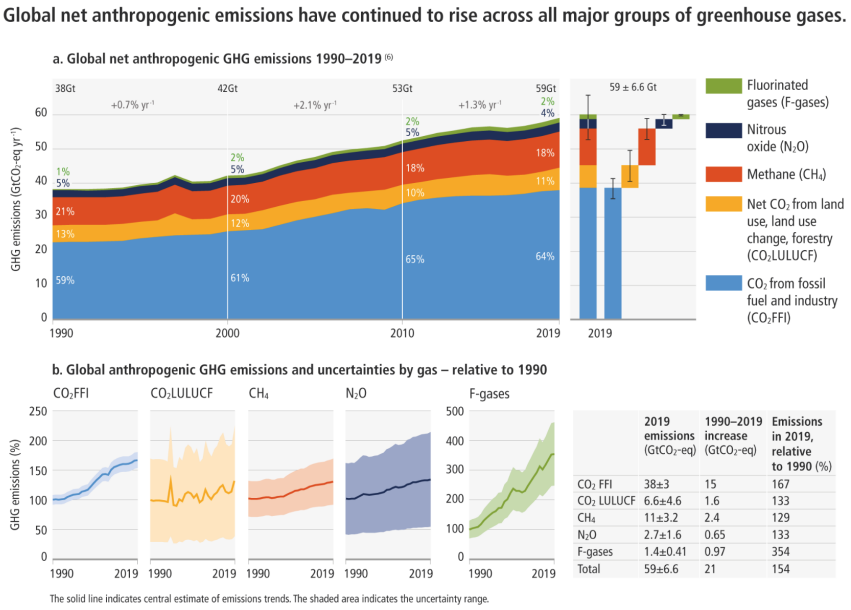


FIGURE 1.1: Global net anthropogenic GHG emissions colored by type ($GtCO_2$-eq per year) from 1990 to 2019, IPCC AR6, Working Group III, Mitigation of Climate Change, 2022

There are several ways in which GHG emissions are affecting the planet:

1. GHG in the atmosphere are blocking infra-red heat emission of the Earth back into space. Their accumulation thus leads to a global temperature elevation.

As a consequence, all ecosystems are threatened by repercussions related to climate change, like from heat stress, extreme climatic events, drought, or the rise of ocean levels.

2. GHG elevation in the atmosphere, and especially CO_2, directly and significantly interferes with planetary equilibria including oceanic pH levels or plant life.

Given the seriousness of this phenomena, the IPCC has led efforts to predict future CO_2 concentrations [IPCC, 2013]. Figure 1.2 is a forecast of the CO_2 levels in the atmosphere until the end of the century for several scenarios. Those scenarios depend on our global choices in terms of energy production, industry, transportation, agriculture, or lifestyle. Even though the political nature of such questions makes predictions challenging, the baseline foresees CO_2 levels between 900 and 1000 parts per million (ppm) in 2100, approximately twice their current amount (RCP 8.5). It is worth noting that even in the scenario in which we do not emit additional CO_2 in 2100 (RCP 2.6), the CO_2 concentration in 2100 is still above its current amount because of the strong inertia of CO_2 accumulation in the atmosphere.

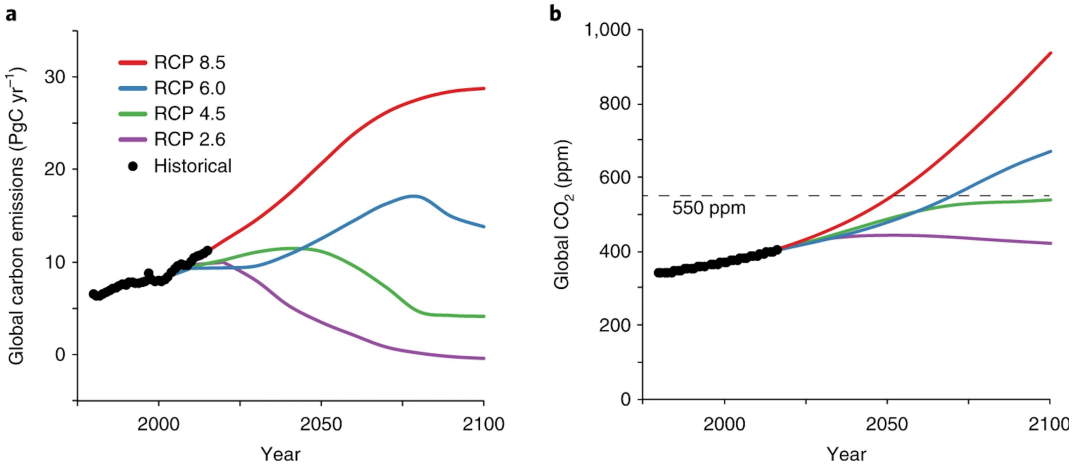


FIGURE 1.2: Carbon emissions and CO_2 concentration predictions from 2014 to 2100 for 4 IPCC scenarios called Representative Concentration Pathway (RCPs), representing scenarios of GHG emissions worldwide. **a**. Predictions for anthropogenic carbon emissions from 2014 to 2100 in Petagrams (10^{15} g) of carbon per year (equivalent to Gigatons per year). **b**. Predictions of CO_2 concentration in parts per million (ppm). Historical observations show that we are currently following the RCP 8.5 pathway, leading to 550 ppm in the middle of the century, and almost 1000 ppm in 2100. Data compiled by Smith and Myers, 2018 [Smith and Myers, 2018].

CO_2 has been fluctuating in the atmosphere since the origins of mankind: figure 1.3 shows that CO_2 levels from 800 000 years ago to today have been oscillating due to glacial cycles. They have, however, never been as high as in 2018, and the predicted levels of RCP 8.5 are dramatically exceeding all prehistorical and historical fluctuations. This highlights the unprecedented and brutal aspects of current CO_2 accumulation as compared to times when human activity did not emit carbon.

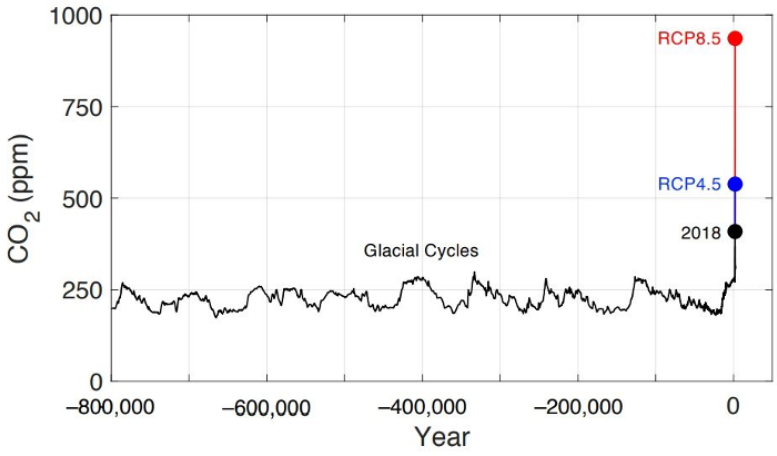


FIGURE 1.3: CO_2 concentration in the atmosphere in the 800000 leading to 2018. IPCC predictions for RCP 4.5 and 8.5 extend observed data [Lüthi et al., 2008].

1.1.2 The decline of plant mineral nutrition under rising CO_2: physiological and molecular aspects of a bad deal.

1.1.2.1 Preamble

Note: this section is intended to summarize a complete plant biology review, presented in the next section. It should give the necessary information to understand the basic motivations of this PhD project in terms of knowledge gaps about plant physiological and molecular responses to elevated CO_2.

In addition to driving climate change and threatening cultivated crops through extreme climatic events, elevated CO_2 (eCO_2) on its own can impact the physiology of plants. **CO_2 is a key substrate of photosynthesis**, a reaction in plants utilizing water, CO_2 and light to produce biomass and emit oxygen. In particular, most of the cultivated plants fall into the category of C3 plants, in which the photosynthetic reaction is still limited by current atmospheric CO_2 concentration. It is thus expected that an increase in atmospheric CO_2 will result in more primary biomass production, and even improved agricultural yield because of this **fertilization effect** [Tausz-Posch, Tausz, and Bourgault, 2019]. Still, two notable repercussions of CO_2 elevation are sources of concern:

1. As compared to ambient CO_2 conditions (aCO_2), the efficiency of photosynthesis is reduced under eCO_2. This is known as the **acclimation of photosynthesis to eCO_2** [Ainsworth and Long, 2020; Tausz-Posch, Tausz, and Bourgault, 2019]. As a result, more CO_2 is captured under eCO_2, but to a lesser extent than theoretically predicted, which lowers production gains compared to expectations in the fields.

2. **The mineral composition of C3 plants is depleted under eCO_2.** Almost all mineral nutrients in plants, composing the ionome (e.g elements such as N, P, K, S, Fe, Na, Mg, Mg, or Zn) are affected, and can be reduced from 5 to 25% depending on the species, element, and conditions. The nutrient that plants need in highest quantities, nitrogen (N), is often particularly affected [Loladze, 2014] (Figure 1.4). This mineral depletion in cultivated plants can

lead to the consumption of food with lower amounts of protein, vitamins or necessary oligo-elements, and poses the threat of malnutrition and impaired human heath at a global scale [Myers et al., 2017; Smith and Myers, 2018].

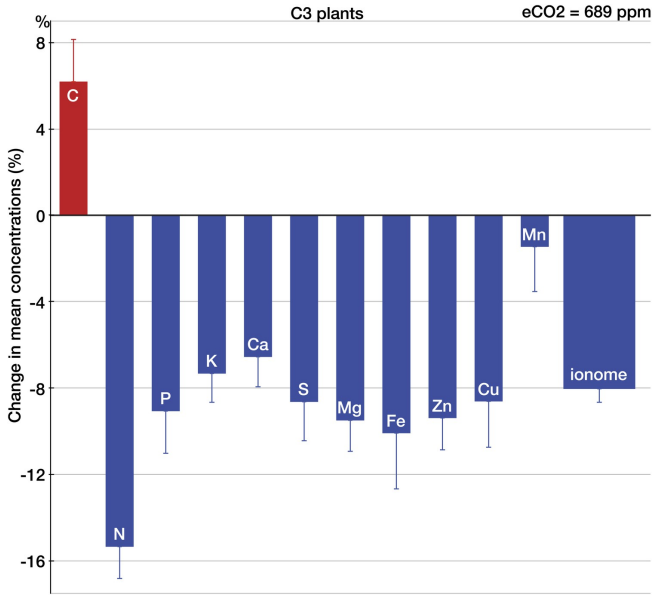


FIGURE 1.4: Relative change in different mineral elements under eCO_2, in a panel of C3 species [Loladze, 2014].

Although those negative impacts have been consensually observed in both natural and artificially controlled settings, **the mechanisms underlying the decline of mineral composition under eCO_2 remain unclear.**

Several explanatory hypotheses have been proposed, and we summarized most of them in Figure 1 of Publication #1. Some of them are likely to be involved in mineral depletion under eCO_2 but are not sufficient to explain, alone, all the physiological and molecular observations made in these conditions. Among them are the hypotheses of carbon dilution, reduced transpiration [Sun et al., 2022], or altered nutrient availability in soils.

Another category of hypotheses focuses on **the alteration of signalling, uptake and assimilation mechanisms of nutrients, and especially those of N**. Even though signalling mechanisms of N nutrition have been partly mapped, they have not been studied under eCO_2, and some transcriptomic evidence shows that signalling modules could be impacted. N assimilation also exhibits signs of disruption in the shoots, possibly caused by the alteration of photorespiration and diminished reducing power for the assimilation reactions. Finally, the activity of root uptake systems is altered in different plants where N acquisition efficiency (in the form of NO_3^-) was diminished [Rubio-Asensio and Bloom, 2017; Bloom et al., 2010], and where N transport genes were reported as transcriptionnally misregulated under eCO_2 (see section "Negative impact of eCO_2 on NO_3^- assimilation" in Publication #1 for details and references). Those transcriptional changes could not lead to the identification of clear patterns, firstly because too few data are currently available in roots tissues, but also because those regulations may be heavily influenced by developmental and environmental factors. This promotes the hypothesis of complex regulatory pathways acting upon the expression of key N nutrition genes in the roots under eCO_2, with unknown regulators orchestrating such networks.

Two main types of solutions are identified to mitigate the effect of eCO_2 on agriculture:

- **Engineering plants** in which the efficiency of nutrient uptake systems is restored. This requires gaining knowledge into the molecular mechanisms of mineral nutrition under eCO_2 to identify genes to target in the context of genetic manipulations. A handful of successful examples have already paved the way (see Box 2 of the full review).
- **Leveraging intraspecific natural variability** to cultivate more resilient crops. In fact, the genetic diversity of the ionome response to eCO_2 has been characterized in several crops of agronomic interest [Zhu et al., 2018; Myers et al., 2014a], and some promising ecotypes are not showing any mineral depletion. Although Genome-Wide Association studies (GWAs) have been performed on biomass and yield-related traits [Oguchi et al., 2022] (Section "Solutions to improve N content under eCO_2" in Publication #1), similar studies could be performed on the ionome response to eCO_2 with the objective of discovering polymorphisms involved in the control of nutrient content. Such discoveries would provide theoretical knowledge about genetic regulations under eCO_2, while putting forward crops with a preserved nutritional quality for agriculture.

1.1.2.2 Publication #1 (Published)

Note : This section has its own reference system. Citation numbers refer to bibliography items included in the present article, and not at the end of the PhD manuscript. ***This manuscript has been accepted in Trends in Plant Science and published in the February 2023 issue of the journal.***

Trends in Plant Science

CellPress

Feature Review

The decline of plant mineral nutrition under rising CO_2: physiological and molecular aspects of a bad deal

Alain Gojon,[1] Océane Cassan,[1] Liên Bach,[1] Laurence Lejay,[1] and Antoine Martin [1,*]

The elevation of atmospheric CO_2 concentration has a strong impact on the physiology of C3 plants, far beyond photosynthesis and C metabolism. In particular, it reduces the concentrations of most mineral nutrients in plant tissues, posing major threats on crop quality, nutrient cycles, and carbon sinks in terrestrial agro-ecosystems. The causes of the detrimental effect of high CO_2 levels on plant mineral status are not understood. We provide an update on the main hypotheses and review the increasing evidence that, for nitrogen, this detrimental effect is associated with direct inhibition of key mechanisms of nitrogen uptake and assimilation. We also mention promising strategies for identifying genotypes that will maintain robust nutrient status in a future high-CO_2 world.

Highlights

Elevated $[CO_2]$ (eCO_2) has a negative impact on key physiological mechanisms of nutrient acquisition and assimilation in C3 plants. The reasons are largely unknown.

eCO_2 particularly lowers nitrogen content of plants tissues, possibly through specific inhibition of nitrate uptake and assimilation.

The altered nutrient status of plants grown at eCO_2 is one likely cause of the acclimation of photosynthesis to eCO_2 that prevents full stimulation of biomass production in response to 'CO_2 fertilization'.

The high natural genetic variability of the eCO_2 impact on plant nutrient status can be exploited as a promising strategy to breed future crops better adapted to a high-CO_2 world.

Elevation of atmospheric CO_2 has positive and negative impacts on plant nutrition

The continuous elevation of atmospheric CO_2 concentration ($[CO_2]_{atm}$) since the preindustrial era (from ~280 to ~415 ppm) has been unprecedented in both rate and amplitude over the past 3 million years [1,2]. In addition to being a main driver of climate change, this elevation has a strong impact on plant nutrition owing to the pivotal substrate/signal functions of CO_2.

On the positive side, because C3 photosynthesis is limited by current $[CO_2]_{atm}$, the increase of this concentration is predicted to result in a significant enhancement of CO_2 capture by C3 plants (the so-called 'CO_2 fertilization' effect), leading to improved primary biomass production [3,4]. This is clearly welcome because enhanced photosynthesis is an absolute requirement for satisfying the increasing demand for food and for mitigating the $[CO_2]_{atm}$ rise [5–7]. The 'CO_2 fertilization' effect has been investigated using various experimental facilities to artificially increase $[CO_2]_{atm}$, including controlled closed growth chambers or greenhouses, open-top chambers (OTCs), and free-air CO_2 enrichment (FACE) facilities in the field [8,9]. Although this depends on the species and facilities, increases in yield of 20–30% are commonly reported for C3 crops grown at the $[CO_2]_{atm}$ expected during the second half of the century [9–12]. These are substantial gains that constitute an important opportunity for securing food and feed production. However, many studies reached the conclusion that the actual stimulation of photosynthesis and biomass production often remains lower than theoretically predicted [3,4,9,10]. This is generally due to a downregulation of photosynthesis efficiency in plants grown at eCO_2 compared to ambient CO_2 (aCO_2) (the so-called 'acclimation of photosynthesis to eCO_2') that is associated with an accumulation of non-structural carbohydrates, a decrease in leaf total protein content, including Rubisco, and a reduced Rubisco activation state [3,8].

On the negative side, an unexpected outcome is that growing C3 plants at eCO_2 has a detrimental impact on their mineral status because it leads to lowered concentrations of the main nutrients in most organs. Although already recorded more than 20 years ago [13–15], the

[1]Institut des Sciences des Plantes de Montpellier (IPSiM), Université de Montpellier, Centre National de la Recherche Scientifique (CNRS), Institut National de Recherche pour l'Agriculture, l'Alimentation, et l'Environnement (INRAE), Institut Agro, Montpellier, France

*Correspondence: antoine.martin@cnrs.fr (A. Martin).

actual extent and generality of this impact have been firmly established only recently [16–20]. Although differences are observed between plant functional types [10,21], almost all C3 species and nutrients (N, P, K, S, Fe, Mg, Zn) are concerned, and decreases in tissue concentrations range from 5 to 25% depending on the nutrient, the eCO_2 level, and the experimental facility. This is anticipated to have at least two strong negative consequences. First, it may deteriorate the nutritional quality of most staple crops, leading to amplified malnutrition and health problems at the global scale [17,22,23]. Second, it will significantly modify the elemental stoichiometry in the plant biomass (especially C/N and C/P ratios) which will have direct effects on the stability of the soil organic matter and on the biogeochemical processes of the nutrient cycles in the soil [24].

Several lines of evidence recently supported the validity of the above predictions from eCO_2 experiments. Concerning biomass production, it now appears that, at the global scale, the 'CO_2 fertilization' effect is real and has already resulted in enhanced photosynthetic CO_2 capture and vegetation primary production in response to the past elevation of $[CO_2]_{atm}$, explaining that the Earth has become substantially greener over the past decades [25–28]. This is considered to be the major cause of the recent increase in terrestrial carbon sink that constitutes a strong negative feedback on climate warming [25,26,29,30]. Concerning plant nutrient status, long-term studies on forests showed that the foliar mineral status of trees actually declined over the past decades [31,32]. Furthermore, analysis of archived samples confirmed that contemporary plants have lower nutrient concentrations than did plants harvested more than a century ago [33,34]. Finally, plants growing in natural eCO_2 springs (volcanic sites where $[CO_2]_{atm}$ is locally naturally elevated) have decreased leaf N content compared to their counterparts growing nearby at aCO_2 [35]. Altogether, these studies indicate that the lowered nutrient status of C3 plants is already a visible consequence of the past elevation of $[CO_2]_{atm}$.

Most importantly, the reasons and mechanisms underlying the negative impact of eCO_2 on the mineral nutrition of C3 plants are still unclear [3,19,36]. In this review we provide an overview of the latest developments on this topic, focusing on nitrogen (N), the mineral nutrient required in highest amounts in plants, and which is also the nutrient often most impacted by eCO_2 [16,20].

Hypotheses for the negative impact of eCO_2 on the mineral status of C3 plants

Multiple causes have been proposed for the lowered concentrations of nutrients observed in plants grown at eCO_2 compared to aCO_2 (Figure 1). Restricted bioavailability of nutrients in the soil may certainly contribute [37], but this cannot be the sole explanation because the negative impact of eCO_2 on plant mineral status is also seen in plants grown hydroponically [38–42]. Thus, it is now clear that eCO_2 also acts on the plant side to impair the mechanisms involved in nutrient homeostasis.

It is commonly postulated that the decrease in nutrient concentrations affecting plants grown at eCO_2 results from a simple 'dilution' effect due to the increase in biomass and C content of the tissues [14,36,43]. Although this cannot be totally ruled out, compelling evidence shows that this is not the main explanation. Indeed, there is generally no correlation between the increase in biomass or C content and the decrease in mineral status [3,17,44]. For instance, lowered nutrient concentrations are recorded even in plants showing no growth stimulation in response to eCO_2 [19,21,45]. Conversely, a strong increase in biomass production can be observed without any significant decrease in nutrient concentrations, especially in the functional groups of species that benefit the most from eCO_2, such as legumes or forest trees [13,17,46]. Furthermore, in contrast to what is expected from a general dilution effect, the changes in concentrations are often very different for the various nutrients [17,21].

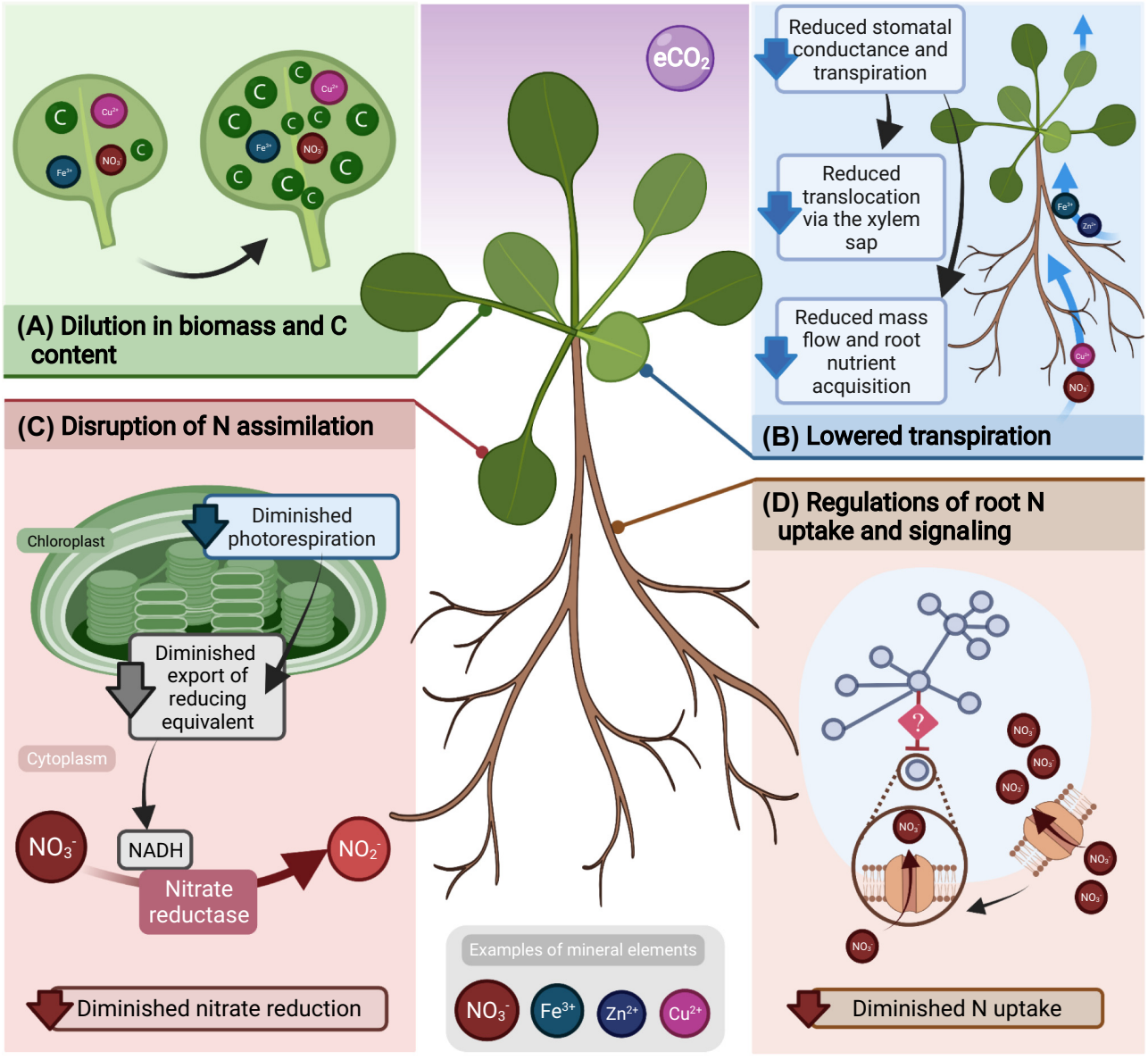


Figure 1. Main hypotheses to explain the negative impact of elevated [CO_2] (eCO_2) on the mineral composition and especially on N content of C3 plants. (A) An increase in biomass and C quantity in plant leaves can result in a lower mineral concentration via a dilution effect. (B) Under eCO_2, transpiration is reduced due to lower stomatal conductance. This could negatively impact on nutrient acquisition in roots because of the reduced mass flow in the soil, and because of diminished nutrient translocation via the xylem sap in the shoots. (C) Metabolic pathways associated with photorespiration result in the production of NADH in the cytosol, which is used as reducing power by the NO_3^- reductase enzyme. As a consequence, reduced photorespiration under eCO_2 might lead to insufficient NADH to power the NO_3^- reduction reaction. This would lead to less NO_3^- assimilation and reduced N content in shoots. (D) NO_3^- uptake systems are deregulated under eCO_2 in the roots, which might lead to a decreased rate of NO_3^- acquisition and eventually of N content in plants. The signaling pathways and regulatory mechanisms involved in this mechanism remain unknown.

Another popular hypothesis is that the negative impact of eCO_2 on plant mineral status is due to a lowering of transpiration caused by eCO_2-induced reduction of stomatal conductance (g_s) [47–50]. Transpiration can be reduced by ~30% by eCO_2, and it is postulated that this may affect not only

root nutrient acquisition through reduced mass flow in the soil but also nutrient accumulation in shoots through diminished translocation via the xylem sap. Again, this certainly explains part of the altered mineral nutrient phenotype of plants grown at eCO_2, but strong evidence indicates that this effect is not predominant. First, as already mentioned previously, the negative impact of eCO_2 on plant mineral status is seen in plants grown hydroponically, thus excluding a major role of mass flow in the soil. Second, it has been known for a long time that, unless water flux within the plant is dramatically reduced, root uptake and xylem translocation of nutrients are largely independent of transpiration [51,52]. Third, and most importantly, reduced nutrient concentrations in response to eCO_2 are observed in C3 but not in C4 plants [16], whereas the eCO_2-induced reduction of g_s similarly affects C3 and C4 species [53]. Nonetheless, recent studies confirm a possible impairment of nutrient translocation into the shoot in response to eCO_2, but relate it to a specific effect on xylem morphogenesis. Indeed, Houshmandfar *et al.* [54] and Li *et al.* [55] reported that the concentrations of cations (K^+, Ca^{2+}, Mg^{2+}) in the xylem sap of wheat plants are reduced under eCO_2 compared to aCO_2, an observation that cannot be explained by lowered transpiration. Interestingly, eCO_2 represses xylem development in tomato roots, thereby strongly decreasing root xylem area (by more than 50%), especially under non-limiting N supply, despite a strong stimulation of root system growth [56]. Gray *et al.* [57] also reported a marked reduction of root metaxylem area in tomato plants after 10 days of growth under eCO_2, following an initial increase between 7 and 10 days. These observations are consistent with the finding that eCO_2 leads to a ~10–20% increase in the ratio of cortical thickness to stele radius in a wide range of species, indicating preferential cortex over stele growth [58].

Finally, there is a growing consensus that a main cause of the negative impact of eCO_2 on plant mineral status is reduced efficiency of nutrient acquisition and/or assimilation. In many instances, the total amount of nutrients taken up by the plant is increased in response to eCO_2, but not enough to keep pace with the stimulation of biomass production [3,12,45,59–61]. Furthermore, this total amount can also be decreased, especially when the effect of eCO_2 on growth is limited [3,45,55,59]. Many studies have pointed out that the reduced efficiency of root nutrient uptake in response to eCO_2 is associated with physiological rather than developmental processes [12,21,62,63]. Indeed, this reduced efficiency cannot generally be accounted for by an impaired size or architecture of the root system because eCO_2 predominantly has a positive effect on root growth and development [64]. This suggests that a major cause of the detrimental effect of eCO_2 on the nutrient status of C3 plants is downregulation of the activity of the membrane transporters involved in root nutrient uptake, which for metabolizable ions may be associated with downregulation of the assimilatory pathways. This hypothesis is detailed in the following section specifically for N for which it is most documented.

Negative impact of eCO_2 on N acquisition

A large number of studies have investigated the effect of eCO_2 on plant nutrient uptake, and especially on N. Although it may be highly variable [3,15,65], many studies show that eCO_2 significantly impedes the efficiency of N acquisition by plants by decreasing the rate of N uptake as measured per unit of root biomass [36,66]. However, the effect of eCO_2 on plant N uptake could be different according to the form of N that is collected by plants [67]. Indeed, eCO_2 appears to target root nitrate (NO_3^-) uptake much more negatively than root ammonium (NH_4^+) or organic N uptake. In *Arabidopsis thaliana (arabidopsis)*, *Medicago truncatula*, and wheat, eCO_2 significantly reduced the rate of NO_3^- uptake by almost 50% [68,69]. The negative effect of eCO_2 on root NO_3^- uptake has been confirmed in other species, and therefore seems to be a general observation [39,40,62,70–72]. As mentioned previously, the negative impact of eCO_2 on NH_4^+ uptake is less evident than for NO_3^-. On the one hand, several studies found an effect similar to that on NO_3^- uptake [39,40,73], with a reduced rate of root NH_4^+ uptake under eCO_2 conditions. On another hand,

several experiments confirmed a negative effect on NO_3^- uptake but no effect or even a stimulation of NH_4^+ uptake by eCO_2 [62,72]. Such a difference between NO_3^- and NH_4^+ uptake may result from a more adverse effect of eCO_2 on NO_3^- reduction than on NH_4^+ assimilation (see following section).

To explain the effect of eCO_2 on N uptake systems, transcriptomic approaches have been used as a valuable source of information. The extent of changes exerted by eCO_2 on gene expression has been tested under several conditions. However, considering the major effect of eCO_2 on plant growth and physiology, relatively few transcriptomic approaches have been performed. A first striking observation emerging from these experiments is that eCO_2 leads to little changes in genome expression. In arabidopsis, soybean, rice, and poplar the number of genes differentially regulated by eCO_2 ranges from only ~10 to a few hundred [74–78]. In the tetraploid durum wheat genome, the effect of eCO_2 increases to reach ~600 differentially expressed genes [79]. In all these experiments, several observations show that eCO_2 can modify the regulation of genes involved in NO_3^- transport systems (NRT1/NPF and NRT2 families) (Figure 2A). Notably, the decrease of NO_3^- uptake observed in *Medicago* plants was associated with a downregulation of NO_3^- transporter *NRT1.1* gene expression by eCO_2 [69]. In soybean leaves, two genes corresponding to the *NRT1.3* transporter are upregulated by eCO_2, whereas another NO_3^- transporter is downregulated [74]. In soybean and arabidopsis roots, the *NRT1.5* gene is upregulated under eCO_2 [75,80]. However, *NRT1.5* is known to be involved in root-to-shoot NO_3^- transport [81], and its increased expression should not directly contribute to the negative effect of eCO_2 on shoot N content. In wheat leaves and roots, eCO_2 leads to deregulation of many *NRT* genes, but the effect of this deregulation can vary according to the tissue or stage of development [82]. The most significant effect would be the downregulation of several *NRT1* genes in leaves, with a potential effect of ambient temperature [80,82]. In roots, eCO_2 leads to the downregulation of some *NRT2* genes, especially under high-N conditions [41]. Altogether, the few transcriptomic experiments carried out under eCO_2 reveal that eCO_2 might have a significant effect on the expression of genes involved in N transport, but the variations observed often follow different and even contradictory directions (Figure 2A). Therefore, no clear trend has so far emerged concerning the role of eCO_2 on the regulation of N transport gene expression, and large-scale analysis of the molecular responses of root N uptake systems to eCO_2 is a priority to address this question.

The hypothesis that eCO_2 may repress N uptake efficiency while stimulating photosynthesis is at odds with that was previously known. Indeed, stimulation of photosynthesis has always been reported to also trigger a marked induction of root N uptake systems (Box 1), which is seen as a necessary regulatory mechanism for coordinating N and C intake into the plant. However, the stimulation of root N uptake by increased photosynthesis has generally been investigated through short-term treatments (a few hours to a few days) with light, sugar supply, or changes in $[CO_2]_{atm}$. These may not be illustrative of long-term responses to eCO_2. We are aware of only a single study that compared the short-term versus long-term responses of root N uptake to eCO_2 [40]. It showed that short-term eCO_2 treatment (24 h) stimulates root NO_3^- and NH_4^+ uptake during early vegetative growth (but not at mid-reproductive stage), whereas long-term treatment (>100 days) had a marked opposite effect. If confirmed, such a paradox in the short-term versus long-term effects of eCO_2 will certainly deserve specific attention because it would fill an important gap in our knowledge of the C/N interactions in plants.

Negative impact of eCO_2 on NO_3^- assimilation

A major component of the link between eCO_2 and N metabolism is the fact that eCO_2 might diminish the ability of plants to assimilate NO_3^-. Recent evidence has been acquired from several studies involving arabidopsis and wheat plants grown under controlled [68,83] or FACE

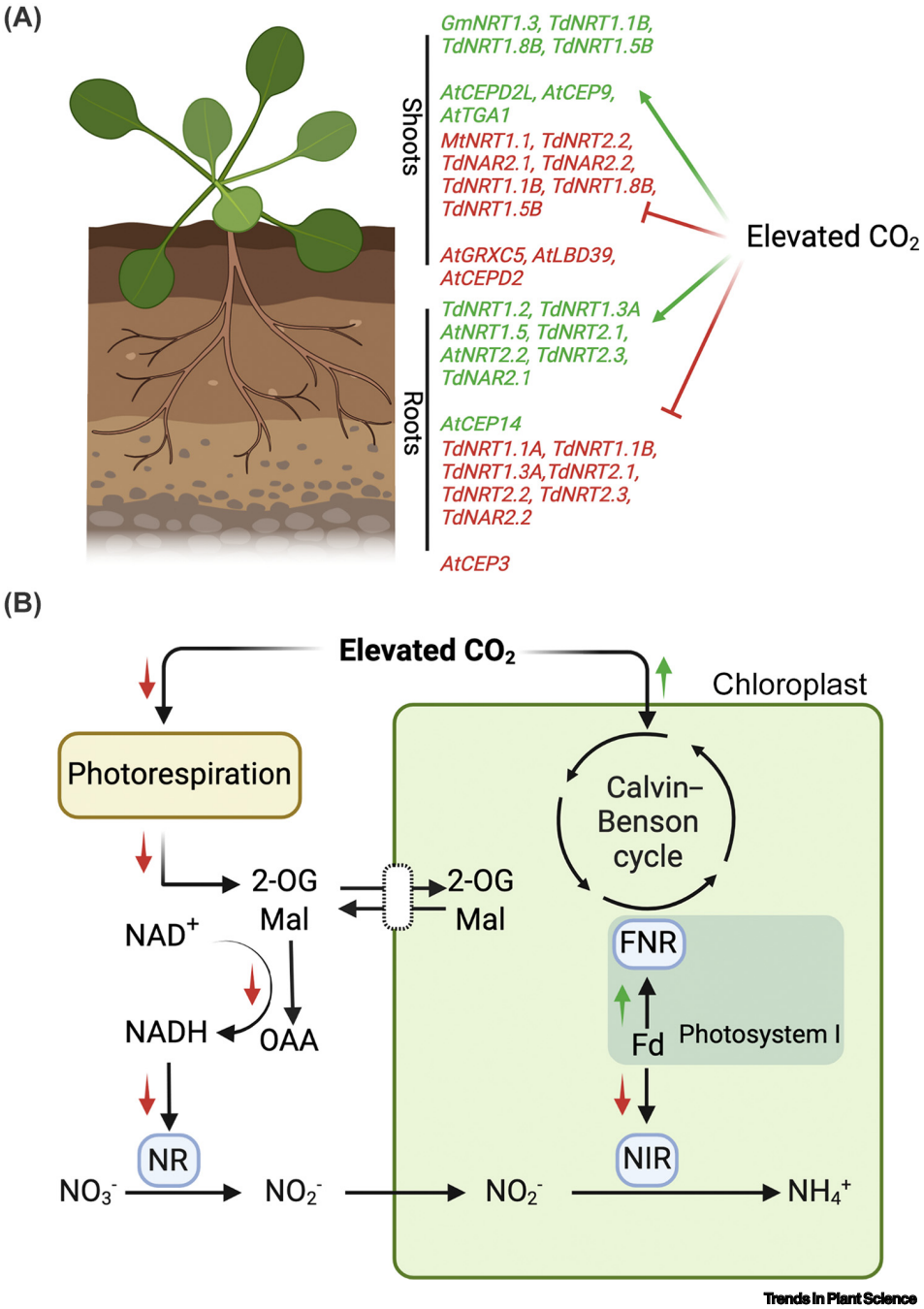


Figure 2. Proposed mechanisms underlying the effect of elevated [CO_2] (eCO_2) on nitrate transport and assimilation pathways. (A) The effect of eCO_2 on the expression of genes associated with N uptake and transport and their regulators in plants. Summary of data collected from transcriptomic experiments in various species, showing the effect of eCO_2 on genes of the *NRT1*, *NRT2*, and *NAR* families and their associated regulators, in shoots or in roots (*Gm*, *Glycine max*; *Mt*, *Medicago truncatula*; *Td*, *Triticum durum*; *At*, *Arabidopsis thaliana*). Green, genes upregulated by eCO_2. Red, genes downregulated by eCO_2. The opposite effects observed for some of these genes can be partly explained by the effect of other environmental factors (e.g., N regime or elevated temperature) or by differences in developmental stage.

(Figure legend continued at the bottom of the next page.)

Box 1. Short-term stimulation of root N acquisition by photosynthesis

Because carbon skeletons are required for amino acid synthesis, root NO_3^- transporters are regulated by the production of sugars by photosynthesis [130–132]. This regulation is characterized by fast responses (within hours) of root NO_3^- uptake, paralleling the changes in shoot CO_2 fixation, and leading for instance to a marked day/night cycle with an increase of NO_3^- uptake during the day and a decrease during the night. The daytime increase of NO_3^- uptake requires the presence of CO_2 in the atmosphere, indicating that the upregulation mechanism does not involve light *per se*, but instead depends on sugars produced by photosynthesis [130]. This conclusion is supported by the fact that addition of sugars to the nutrient solution, when plants are in the dark, can restore the level of root NO_3^- influx observed in the light [130,132]. At the molecular level, the regulation of root NO_3^- uptake by the aforementioned treatments (light/dark, sugar supply, CO_2-free atmosphere) is mirrored by similar changes in the expression of key root NO_3^- transporter genes such as *NRT2.1*, *NRT2.4*, and *NRT1.1/NPF6.3* [114,131,132]. In addition, in short-term treatments, *NRT2.1* and *NRT1.1/NPF6.3* were found to be upregulated in plants transferred to eCO_2 for 4 h in the light [114]. This suggested the action of signaling mechanisms because sugars produced by photosynthesis are known to be important signals that control different aspects of plant metabolism and development [133]. In contrast to known sugar-sensing mechanisms, the sugar signaling pathway involved in regulating root NO_3^- transporters appears to be entirely distinct [134] because it was shown to originate from the oxidative pentose phosphate pathway (OPPP) for at least *NRT2.1*, *NRT2.4*, and *NRT1.1/NPF6.3* [114]. It is interesting to note that OPPP is also involved in the regulation by sugars of genes encoding enzymes for NO_3^- assimilation [135]. Furthermore, the link between OPPP and N metabolism in roots is strong because OPPP produces the reducing power required for nitrite reductase (NiR) and glutamate synthase (GOGAT) activity [136]. If the signaling mechanism linked to the OPPP remains unknown, recent advances suggest that the reducing power produced by the OPPP could be involved in redox regulation of N uptake. Indeed, *NRT2.1* expression has been shown to be regulated by the redox status of the plant [112]. This mechanism could ensure the coordination of root NO_3^- uptake with the production of reducing equivalents required for the assimilation of N into amino acids ([113] for review).

conditions [84,85], and has been confirmed in other species [86]. The negative effect of eCO_2 on NO_3^- assimilation is in line with a range of more focused observations on genes and proteins involved in NO_3^- assimilation pathways. Indeed, a large number of studies demonstrated that eCO_2 leads to a decrease in nitrate reductase (*NR)* gene expression, and in nitrite reductase (*NiR)* and glutamine synthetase (*GS)* gene expression to a lesser extent [69,73,87–89]. Accordingly, the enzymatic activity of NR also tends to decrease under eCO_2 conditions [44,69,79,80,90,91]. The negative effect of eCO_2 on NR activity seems to be more pronounced in shoots [67,92], but has also been observed in roots [69]. Other mechanistic hypotheses have been proposed to explain the negative effect of eCO_2 on NO_3^- assimilation. One of the most advanced hypotheses lies in the possibility that available reducing power might be limiting for NO_3^- assimilation enzymes under eCO_2. Indeed, the reduction of NO_3^- to NH_4^+ by NR and NiR requires a large amount of reducing power, whose level has been proposed to be the main limiting factor for the NO_3^- reduction reaction [93]. In the chloroplast, reduced ferredoxin is required for both ferredoxin-NADP reductase, which generates NADPH at the end of the electron flux of the photosynthesis light reaction, and for NiR. Because eCO_2 boosts photosynthesis, the availability of reduced ferredoxin for NiR might be diminished under these conditions, especially because ferredoxin-NADP reductase has a higher affinity than NiR for reduced ferredoxin [94]. Another mechanism could involve photorespiration which generates cytoplasmic NADH that fuels NR reaction [95]. Among other mechanisms, 2-oxoglutarate (2-OG) produced by photorespiration is exported to the chloroplast in exchange for malate (Mal) import into the cytosol, which is recycled into oxaloacetate (OAA), generating NADH that is used by NR (Figure 2B). Because photorespiration is reduced under eCO_2, the level of cytoplasmic NADH decreases

Data were collected from [38,41,69,74–76,80,82,106,108,109]. (B) Schematic representation of metabolic pathways by which eCO_2 can modify the availability of reducing power needed for the two steps of nitrate reduction. Red and green arrows indicate the metabolic routes that can be slowed or accelerated by eCO_2, respectively. eCO_2 boosts the rate of the Calvin–Benson cycle, increasing the demand of reduced ferredoxin (Fd) by ferredoxin-NADPH reductase (FNR) to provide NADPH for the C fixation pathway. This can reduce the availability of Fd for nitrite reductase (NIR), which has a lower affinity than FNR for Fd. At the same time, eCO_2 decreases the rate of photorespiration. The reduced production of 2-oxoglutarate (2-OG) by a lower photorespiration can decrease the export of malate (Mal) to the cytosol, which is needed to provide NADH for nitrate reductase (NR). Abbreviation: OAA, oxaloacetic acid.

under these conditions and could directly affect the rate of NO_3^- assimilation [44,96,97]. This hypothesis is supported by a recent report describing the effect of the genetic manipulation of the chloroplast factor *OsCV* which is involved in rice in the repression of photorespiration-related genes. Silencing of *OsCV* under eCO_2 results in increased photorespiration, increased *NR* gene expression, restoration of NR activity and, most notably, is associated with inhibition of the negative effect of eCO_2 on protein content [98]. Interestingly, a recent approach modeling plant primary metabolism under eCO_2 also identified the limitation of reducing power by a lower rate of photorespiration as a mechanism that could significantly contribute to the inhibition of NO_3^- assimilation [99].

If confirmed, the specific inhibition of NO_3^- uptake and/or NO_3^- assimilation systems by eCO_2 has profound consequences for agriculture and ecology. In particular, it may explain why plants supplied with NO_3^- as the N source often show lower stimulation of photosynthesis and growth in response to eCO_2 compared to NH_4^+-fed or N_2-fixing plants [83,100,101]. This suggests that NO_3^- may be an inappropriate N source for taking full advantage of the 'CO_2 fertilization' effect. Nevertheless, this hypothesis is challenged by other studies which reported that NR activity was not affected by eCO_2 under the nutrient conditions investigated [102,103].

Signaling mechanisms possibly involved in the regulation of N-related processes under eCO_2

In light of the deregulation of N uptake and assimilation processes by eCO_2, identifying and characterizing the effect of eCO_2 on upstream regulatory and signaling mechanisms remains a major challenge in the field of plant N nutrition. Although N signaling pathways and regulatory networks are relatively well known [104,105], to date very few of them have been investigated under eCO_2. However, some evidence strongly suggests that N signaling modules might be misregulated by eCO_2. In particular, transcriptomic analyses show that the expression of some major transcriptional regulators of NO_3^- uptake and assimilation is deregulated by eCO_2 [76,106] (Figure 2A). In addition, several components of the C-terminally encoded peptide (CEP) pathway that is involved in systemic signaling of the N response [107], are also deregulated by eCO_2 [106,108,109] (Figure 2A). Altogether, these observations suggest that eCO_2 might disturb several steps of N signaling pathways upstream of N uptake and assimilation systems. On the other hand, eCO_2 could also profoundly modify the redox properties of plants [110]. A range of key enzymes involved in redox-based processes, such as catalases, peroxidases, and alternative oxidases, are indeed deregulated by eCO_2 [111], which might result in an accumulation of reactive oxygen species (ROS) in plant tissues. The influence of the redox status and of ROS accumulation on the expression of genes involved in NO_3^- uptake such as *NRT2.1* has been recently demonstrated [112,113], and thus the deregulation of redox properties under eCO_2 could contribute to modify N uptake systems under these conditions. In addition, several genes involved in the oxidative pentose phosphate pathway (OPPP) are downregulated by eCO_2, especially under low-N conditions [41,82,106]. Given that the OPPP is a key signaling pathway for the regulation of root NO_3^- transporter genes [114] (Box 1), dysregulation of the OPPP might also contribute to disrupting N acquisition by the roots under eCO_2.

Solutions to improve N content under eCO_2

The decline of N and protein content under eCO_2 conflicts with the imperative to increase staple crop production while maintaining its nutritional quality. For this reason there is an urgent need to identify relevant strategies for increasing plant N content under eCO_2. To achieve this objective, two research avenues stand out. First, manipulation of the genetic determinants that regulate plant physiology under eCO_2 might be a fruitful way to develop climate-resilient crops. Indeed, the few examples that are available in the current literature demonstrate that the response to

eCO_2 can be modulated, including its negative effect on plant nutrient content (Box 2). On the other hand, natural genetic variability found among plant populations can be a very promising way to understand and then overcome the deleterious effect of eCO_2 on plant N nutrition. Indeed, phenotypic variations due to intraspecific genetic diversity in the response to eCO_2 have been reported in a significant number of association studies [quantitative trait locus (QTL) analysis and genome-wide association studies (GWAS)]; however, these mainly focused on yield, biomass, and C-associated traits [115–119].

By comparison, potential genetic variations associated with phenotypic changes in N-related traits or the ionome have been suggested by a much more limited number of studies. Although not performed on large populations of plants, and therefore absent from the GWAS literature, these studies point to encouraging results. In a recent work, yield and grain composition under eCO_2, including protein and mineral content, were recorded in 10 bread wheat genotypes. Although the opposition between biomass increase and the decrease in mineral nutrients drives most of the phenotypic variability, strong differences among the genotypes were reported for these traits [120]. A meta-analysis of FACE experiments on wheat, rice, field peas, soybeans, maize, and sorghum, including intraspecific variations, pointed to a significant decrease in iron and zinc for all C3 plants, and a significant protein decrease in C3 grasses. More importantly, clear differences in the zinc and iron responses between cultivars of rice were noted, and are also foreseen for other species [17]. Similarly, in several FACE experiments, 18 rice genotypes showed diverse relative changes not only in protein but also in iron, zinc, and four types of vitamin B. In particular, some promising genotypes did not exhibit any significant loss of these nutrients [20]. Although such analyses emphasize the value of genetic diversity as a tool for breeding plants with preserved nutritional quality [17,20,120,121], this phenotypic variability remains largely unexplained. Interestingly, variation in gene expression between genotypes might explain in part the phenotypic variability of N-related traits under eCO_2. Indeed, natural variability was also found in the expression patterns of genes linked to mineral nutrition and N content under eCO_2 conditions. Among four durum wheat genotypes exposed to eCO_2 and water deficit, the decrease of N content was associated with significant variation in the expression of genes encoding Rubisco subunits [122]. In a study on arabidopsis accessions Col-0, Cvi-0, and WS exposed to eCO_2, clear ecotype-specific

Box 2. Genetic manipulations to improve the response of plants to eCO_2

Although a small number of genes involved in the response to eCO_2 have been identified, their use might represent a promising way to rapidly develop crops resilient to climate change. Several genes, even though their function is not fully characterized, have the potential to modulate CO_2 assimilation and biomass production under eCO_2. This is the case for the rice G protein qE9-1 whose overexpression leads to dysregulation of photosynthetic gene expression and of *RBCL* genes in particular, and to improvement of CO_2 assimilation and sugar production [137]. In the same line, mutant rice lines for the 14-3-3 protein OsGF14b display an increase in shoot and root biomass which is more pronounced under eCO_2 [138]. Concerning the modulation of the negative effect of eCO_2 on nutrient content, a specific effect on iron metabolism and signaling under eCO_2 has been assigned to *OsRab6a*. Growth, yield, and photosynthetic parameters were stimulated under eCO_2 in *OsRab6a*-overexpressing lines more than in the corresponding wild-type (WT) line. Most notably, *OsRab6a*-overexpressing lines display a moderate reduction in iron content in response to eCO_2 compared to the WT, and this was associated with higher expression of genes involved in iron acquisition such as the iron transporter *OsIRT1* [139]. Similar examples have been described regarding the negative effect of eCO_2 on N metabolism and content. For instance, silencing of the rice chloroplast factor *OsCV* leads to a significant reduction of the negative effect of eCO_2 on protein content, and this was associated with an increase in *NR* gene expression and NR activity [98]. On the other hand, direct manipulation of a cytosolic glutamine synthetase (*GS1*) gene successfully prevented the decline of N content observed under eCO_2. Indeed, cisgenic expression of the barley *GS1.1* isoform improves nitrogen use efficiency (NUE) under eCO_2, leading to a restoration of grain protein content under low-N conditions, and even to an increase in grain protein content under high-N conditions [140]. Finally, manipulating the level of Rubisco in rice was also described as a promising strategy to improve NUE under eCO_2. *RBCS*-RNAi rice lines display a small reduction in their Rubisco content, and this might optimize the allocation of N to Rubisco regarding other factors that limit photosynthesis. Consequently, *RBCS*-RNAi rice lines have better CO_2 assimilation and growth rates, biomass production, and most notably NUE, specifically under eCO_2 conditions [141].

divergences in gene expression were noted for several pathways including photosynthesis, amino acid metabolism, and N metabolism. As an illustration, the N metabolism gene *NIA2* was downregulated in Col-0 and Ws, whereas its expression remained stable in Cvi-0 [109]. The transcriptomic responses of Col-0 and Cvi-0 were further compared in a second FACE study including timecourse information. This work concluded that the changes observed in Col-0 under eCO_2 were similar to the changes in N-deficiency in both the short and long term, whereas Cvi-0 expression reprogramming suggested better acclimation in the long term [106]. Finally, in the leaf transcriptomes of two aspen genotypes under eCO_2, a NO_3^- transporter gene was differentially expressed [123], indicating that the efficiency of N uptake systems under eCO_2 could be genetically driven.

In conclusion, some plant varieties have evolved with a more resilient protein content, mineral status, or even nutrient-related gene expression. Despite this finding, attempts to link this phenotypic and transcriptomic diversity to causal polymorphisms are lacking: association studies between genetic determinants and the plant ionome response will be crucial to gain knowledge about genes controlling mineral nutrition under eCO_2, to elucidate the associated mechanisms, and to design more tolerant crops.

Concluding remarks and future perspectives

In addition to other likely causes (e.g., changes in nutrient bioavailability in the soil, dilution of biomass, reduced transpiration), the hypothesis that eCO_2 has a direct negative effect on key physiological processes of nutrient uptake and assimilation in C3 plants grown under eCO_2 has been increasingly documented in recent years. This negative effect remains largely unexplained because the regulatory mechanisms involved in nutrient homeostasis should theoretically act to prevent it. However, one must keep in mind that the dramatic and continuous elevation of $[CO_2]_{atm}$ is an environmental change that plants have not had to face for at least 3 million years. It may then be postulated that, unlike other abiotic constraints (hydric stress, temperature, nutrient starvation), there has been no selection pressure to drive the emergence and conservation of adaptive responses to eCO_2. Whether the negative impact of eCO_2 on the nutrient status of plants is illustrative of physiological disorders resulting from such a new challenging environment warrants consideration.

These considerations call for a much more extensive investigation of the mechanistic aspects of plant responses to eCO_2, especially at the genetic and molecular levels [124]. This is particularly

Outstanding questions

How can we resolve the paradox between the short-term induction of root N uptake by photosynthesis and the long-term inhibition by eCO_2? In particular, is there a temporal aspect beyond which the effect of eCO_2 switches from positive to negative? Is the redox status of the plant involved in this paradox?

What are the regulatory and signaling mechanisms underlying the negative effect of eCO_2 on plant N uptake and assimilation? For instance, what is the effect of eCO_2 on known N local and long-distance signaling pathways and their regulators? Do the regulators of the negative effect of eCO_2 on N nutrition correspond to new uncharacterized regulators of plant C/N interaction?

Are there common mechanisms behind the decline of plant N content under eCO_2 and that of other microelements such as iron, phosphate, and sulfate? Could these potential mechanisms lead to a global improvement of nutrient composition in crops adapted to a future climate?

Box 3. Impaired N nutrition efficiency as a main cause of the acclimation of photosynthesis to eCO_2

The 'CO_2 fertilization' effect is far from providing the expected benefits for both the mitigation of $[CO_2]_{atm}$ elevation [27,129] and the stimulation of crop yield [9,142]. Given the strategic aspects related to these two issues, it is crucial to understand the causes of this limitation. In many instances the so-called acclimation of photosynthesis to eCO_2 is involved, resulting in down-regulation of photosynthetic efficiency triggered by excessive leaf sugar accumulation [3,4,8,143]. Acclimation is in turn often postulated to be due to a sink limitation created by the inability of either the source leaves to sufficiently increase phloem sugar export [144,145] or the sink organs to use all the sugars produced [146,147]. However, there is now an impressive body of evidence indicating that nutrient limitation, and more particularly N limitation, is a major cause of the acclimation of photosynthesis to eCO_2 [91,125,143,148–153]. Accordingly, increasing N supply to the plants is often a simple way to alleviate the acclimation and to recover the full increase in photosynthesis and biomass production allowed by the 'CO_2 fertilization' effect [8,61,102,148,150,151,154]. The two aforementioned hypotheses – sink limitation or N limitation – are not exclusive because it is likely that the eCO_2-induced decrease of plant N status may well be the cause of the reduced growth of sink organs, thus creating sink limitation [150]. Therefore, understanding the causes of the negative impact of eCO_2 on plant N nutrition will not only help in securing the nutritional quality of crops but will also contribute to increasing crop productivity and mitigating climate change. Moreover, this is also relevant with regard to current biotechnological strategies aimed at improving photosynthesis efficiency. Indeed, many of these strategies rely on the development of synthetic pathways for increasing CO_2 concentration within the chloroplast [6,155]. Thus, lessons learned from eCO_2 studies, especially those related to the sink or nutrient limitations of the 'CO_2 fertilization' effect, may be highly valuable for understanding how biotechnologically improved photosynthetic efficiency can actually result in enhanced growth and yield [142].

true concerning plant nutrient status because – compared to the impact of eCO_2 on C metabolism, biomass production, and yield – its effects on the physiological and molecular mechanisms of plant nutrient acquisition have been dramatically underinvestigated ([125] for review). Stronger efforts must be devoted to addressing these questions, notably in the roots, because there is a clear lack of data concerning the underground parts of the plants because of restricted access to the root system in the vast majority of eCO_2 experiments performed on plants grown in the field or in pots (see Outstanding questions).

Finally, one main issue concerning the detrimental impact of eCO_2 on plant mineral status is to determine the extent to which it may prevent the negative feedback on climate warming associated with the 'CO_2 fertilization' effect. It appears that 'CO_2 fertilization' feedback has actually operated over the past decades and had a significant effect by recapturing as much as 15–20% anthropogenic CO_2 emissions in the atmosphere [25,26,28–30]. Several reports suggest that, in natural forests, the 'CO_2 fertilization' effect will be maintained in the long term because eCO_2 also often improves N availability in the soil and N uptake by the trees [26,61,126,127]. However, recent studies suggest that, at the global scale, the 'CO_2 fertilization' feedback will slow down this century because nutrient limitation (mostly N but also P) will prevent any further major increase in Earth vegetation photosynthesis [27,31,128,129]. This last hypothesis is fully consistent with the observation that, at the plant scale, impaired N nutrition efficiency is a major cause of the acclimation of photosynthesis to eCO_2 (Box 3). Therefore, elucidating the causes of the negative impact of eCO_2 on plant mineral status will not only help in securing the nutritional quality of crops and the stability of the soil organic matter, but will also contribute to mitigating climate change through improved photosynthesis.

Acknowledgments

This work was supported by the I-Site Montpellier Université d'Excellence (MUSE; project ECO2TREATS), the CNRS through the Mission for Transversal and Interdisciplinary Initiatives (MITI) 80 PRIME program, and the Biologie et Amélioration des Plantes (BAP) department of the INRAE. O.C. was recipient of a PhD fellowship from the CNRS.

Declaration of interests

The authors declare no conflicts of interest.

References

1. Willeit, M. *et al.* (2019) Mid-Pleistocene transition in glacial cycles explained by declining CO_2 and regolith removal. *Sci. Adv.* 5, eaav7337
2. Luthi, D. *et al.* (2008) High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature* 453, 379–382
3. Tausz-Posch, S. *et al.* (2020) Elevated [CO_2] effects on crops: advances in understanding acclimation, nitrogen dynamics and interactions with drought and other organisms. *Plant Biol. (Stuttg.)* 22, 38–51
4. Thompson, M. *et al.* (2017) Effects of elevated carbon dioxide on photosynthesis and carbon partitioning: a perspective on root sugar sensing and hormonal crosstalk. *Front. Physiol.* 8, 578
5. Harbinson, J. *et al.* (2021) Designing the crops for the future; the CropBooster Program. *Biology (Basel)* 10, 690
6. Long, S.P. *et al.* (2015) Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. *Cell* 161, 56–66
7. Ort, D.R. *et al.* (2015) Redesigning photosynthesis to sustainably meet global food and bioenergy demand. *Proc. Natl. Acad. Sci. U. S. A.* 112, 8529–8536
8. Ainsworth, E.A. and Long, S.P. (2005) What have we learned from 15 years of free-air CO_2 enrichment (FACE)? A meta-analytic review of the responses of photosynthesis, canopy properties and plant production to rising CO_2. *New Phytol.* 165, 351–371
9. Long, S.P. *et al.* (2006) Food for thought: lower-than-expected crop yield stimulation with rising CO_2 concentrations. *Science* 312, 1918–1921
10. Ainsworth, E.A. and Long, S.P. (2021) 30 years of free-air carbon dioxide enrichment (FACE): what have we learned about future crop productivity and its potential for adaptation? *Glob. Chang. Biol.* 27, 27–49
11. Kimball, B.A. (2016) Crop responses to elevated CO_2 and interactions with H_2O, N, and temperature. *Curr. Opin. Plant Biol.* 31, 36–43
12. Wang, L. *et al.* (2013) Effects of elevated atmospheric CO_2 on physiology and yield of wheat (*Triticum aestivum* L.): a meta-analytic test of current hypotheses. *Agric. Ecosyst. Environ.* 178, 57–63
13. Cotrufo, M.F. *et al.* (1998) Elevated CO_2 reduces the nitrogen concentration of plant tissues. *Glob. Chang. Biol.* 4, 43–54
14. Poorter, H. *et al.* (1997) The effect of elevated CO_2 on the chemical composition and construction costs of leaves of 27 C3 species. *Plant Cell Environ.* 20, 472–482
15. Stitt, M. and Krapp, A. (1999) The interaction between elevated carbon dioxide and nitrogen nutrition: the physiological and molecular background. *Plant Cell Environ.* 22, 583–621
16. Loladze, I. (2014) Hidden shift of the ionome of plants exposed to elevated CO_2depletes minerals at the base of human nutrition. *Elife* 3, e02245

17. Myers, S.S. *et al.* (2014) Increasing CO_2 threatens human nutrition. *Nature* 510, 139–142
18. Soares, J.C. *et al.* (2019) Preserving the nutritional quality of crop plants under a changing climate: importance and strategies. *Plant Soil* 443, 1–26
19. Uddling, J. *et al.* (2018) Crop quality under rising atmospheric CO_2. *Curr. Opin. Plant Biol.* 45, 262–267
20. Zhu, C. *et al.* (2018) Carbon dioxide (CO_2) levels this century will alter the protein, micronutrients, and vitamin content of rice grains with potential health consequences for the poorest rice-dependent countries. *Sci. Adv.* 4, eaaq1012
21. Seibert, R. *et al.* (2022) Plant functional types differ in their long-term nutrient response to eCO_2 in an extensive grassland. *Ecosystems* 25, 1084–1095
22. Högy, P. and Fangmeier, A. (2008) Effects of elevated atmospheric CO_2 on grain quality of wheat. *J. Cereal Sci.* 48, 580–591
23. Smith, M.R. and Myers, S.S. (2018) Impact of anthropogenic CO_2 emissions on global human nutrition. *Nat. Clim. Chang.* 8, 834–839
24. Bertrand, I. *et al.* (2019) Stoichiometry constraints challenge the potential of agroecological practices for the soil C storage. A review. *Agron. Sustain. Dev.* 39, 54
25. Keenan, T.F. *et al.* (2021) A constraint on historic growth in global photosynthesis due to increasing CO_2. *Nature* 600, 253–258
26. Walker, A.P. *et al.* (2021) Integrating the evidence for a terrestrial carbon sink caused by increasing atmospheric CO_2. *New Phytol.* 229, 2413–2445
27. Wang, S. *et al.* (2020) Recent global decline of CO_2 fertilization effects on vegetation photosynthesis. *Science* 370, 1295–1300
28. Zhu, Z. *et al.* (2016) Greening of the Earth and its drivers. *Nat. Clim. Chang.* 6, 791–795
29. Keenan, T.F. *et al.* (2016) Recent pause in the growth rate of atmospheric CO_2 due to enhanced terrestrial carbon uptake. *Nat. Commun.* 7, 13428
30. Schimel, D. *et al.* (2015) Effect of increasing CO_2 on the terrestrial carbon cycle. *Proc. Natl. Acad. Sci. U. S. A.* 112, 436–441
31. Jonard, M. *et al.* (2015) Tree mineral nutrition is deteriorating in Europe. *Glob. Chang. Biol.* 21, 418–430
32. Penuelas, J. *et al.* (2020) Increasing atmospheric CO_2 concentrations correlate with declining nutritional status of European forests. *Commun. Biol.* 3, 125
33. Mariem, S.B. *et al.* (2020) Assessing the evolution of wheat grain traits during the last 166 years using archived samples. *Sci. Rep.* 10, 21828
34. Penuelas, J. and Matamala, R. (1993) Variations in the mineral composition of herbarium Plant species collected during the last three centuries. *J. Exp. Bot.* 44, 1523–1525
35. Saban, J.M. *et al.* (2019) FACE facts hold for multiple generations; evidence from natural CO_2 springs. *Glob. Chang. Biol.* 25, 1–11
36. Taub, D.R. and Wang, X. (2008) Why are nitrogen concentrations in plant tissues lower under elevated CO_2? A critical examination of the hypotheses. *J. Integr. Plant Biol.* 50, 1365–1374
37. De Graaff, M.-A. *et al.* (2006) Interactions between plant growth and soil nutrient cycling under elevated CO_2: a meta-analysis. *Glob. Chang. Biol.* 12, 2077–2091
38. Jauregui, I. *et al.* (2016) Root–shoot interactions explain the reduction of leaf mineral content in *Arabidopsis* plants grown under elevated [CO_2] conditions. *Physiol. Plant.* 158, 65–79
39. Ma, Q. *et al.* (2018) Elevated CO_2 levels enhance the uptake and metabolism of organic nitrogen. *Physiol. Plant.* 162, 467–478
40. Shimono, H. and Bunce, J.A. (2009) Acclimation of nitrogen uptake capacity of rice to elevated atmospheric CO_2 concentration. *Ann. Bot.* 103, 87–94
41. Vicente, R. *et al.* (2016) Metabolic and transcriptional analysis of durum wheat responses to elevated CO_2 at low and high nitrate supply. *Plant Cell Physiol.* 57, 2133–2146
42. Vicente, R. *et al.* (2015) Nitrate supply and plant development influence nitrogen uptake and allocation under elevated CO_2 in durum wheat grown hydroponically. *Acta Physiol. Plant.* 37, 114
43. Coleman, J.S. *et al.* (1994) Interpreting phenotypic variation in plants. *Trends Ecol. Evol.* 9, 187–191
44. Wujeska-Klause, A. *et al.* (2019) Lower photorespiration in elevated CO_2 reduces leaf N concentrations in mature *Eucalyptus* trees in the field. *Glob. Chang. Biol.* 25, 1282–1295
45. Feng, Z. *et al.* (2015) Constraints to nitrogen acquisition of terrestrial plants under elevated CO_2. *Glob. Chang. Biol.* 21, 3152–3168
46. Way, D.A. *et al.* (2010) Greater seed production in elevated CO_2 is not accompanied by reduced seed quality in *Pinus taeda* L. *Glob. Chang. Biol.* 16, 1046–1056
47. Del Pozo, A. *et al.* (2007) Gas exchange acclimation to elevated CO_2 in upper-sunlit and lower-shaded canopy leaves in relation to nitrogen acquisition and partitioning in wheat grown in field chambers. *Environ. Exp. Bot.* 59, 371–380
48. Houshmandfar, A. *et al.* (2018) The relationship between transpiration and nutrient uptake in wheat changes under elevated atmospheric CO_2. *Physiol. Plant.* 163, 516–529
49. Jauregui, I. *et al.* (2015) Nitrogen assimilation and transpiration: key processes conditioning responsiveness of wheat to elevated [CO_2] and temperature. *Physiol. Plant.* 155, 338–354
50. McGrath, J.M. and Lobell, D.B. (2013) Reduction of transpiration and altered nutrient allocation contribute to nutrient decline of crops grown in elevated CO_2 concentrations. *Plant Cell Environ.* 36, 697–705
51. Schulze, E.-D. and Bloom, A.J. (1984) Relationship between mineral nitrogen influx and transpiration in radish and tomato 1. *Plant Physiol.* 76, 827–828
52. Tanner, W. and Beevers, H. (1990) Does transpiration have an essential function in long-distance ion transport in plants? *Plant Cell Environ.* 13, 745–750
53. Ainsworth, E.A. and Rogers, A. (2007) The response of photosynthesis and stomatal conductance to rising [CO_2]: mechanisms and environmental interactions. *Plant Cell Environ.* 30, 258–270
54. Houshmandfar, A. *et al.* (2015) Elevated CO_2 decreases both transpiration flow and concentrations of Ca and Mg in the xylem sap of wheat. *J. Plant Physiol.* 174, 157–160
55. Li, X. *et al.* (2016) Soil warming enhances the hidden shift of elemental stoichiometry by elevated CO_2 in wheat. *Sci. Rep.* 6, 23313
56. Cohen, I. *et al.* (2019) CO_2 and nitrogen interaction alters root anatomy, morphology, nitrogen partitioning and photosynthetic acclimation of tomato plants. *Planta* 250, 1423–1432
57. Gray, S.B. *et al.* (2020) Translational regulation contributes to the elevated CO_2 response in two Solanum species. *Plant J.* 102, 383–397
58. Wang, N. *et al.* (2020) Coordinated responses of leaf and absorptive root traits under elevated CO_2 concentration in temperate woody and herbaceous species. *Environ. Exp. Bot.* 179, 104199
59. BassiriRad, H. *et al.* (1997) Changes in root NH_4^+ and NO_3^- absorption rates of loblolly and ponderosa pine in response to CO_2 enrichment. *Plant Soil* 190, 1–9
60. Newbery, R.M. *et al.* (1995) Nitrogen, phosphorus and potassium uptake and demand in *Agrostis capillaris*: the influence of elevated CO_2 and nutrient supply. *New Phytol.* 130, 565–574
61. Wang, Z. and Wang, C. (2021) Magnitude and mechanisms of nitrogen-mediated responses of tree biomass production to elevated CO_2: a global synthesis. *J. Ecol.* 109, 4038–4055
62. Jackson, R.B. and Reynolds, H.L. (1996) Nitrate and ammonium uptake for single-and mixed-species communities grown at elevated CO_2. *Oecologia* 105, 74–80
63. Tausz, M. *et al.* (2017) Nitrogen nutrition and aspects of root growth and function of two wheat cultivars under elevated [CO_2]. *Environ. Exp. Bot.* 140, 1–7
64. Kimball, B.A. *et al.* (2002) Responses of agricultural crops to free-air CO_2 enrichment. *Adv. Agron.* 77, 293–368. https://doi.org/10.1016/s0065-2113(02)77017-x
65. BassiriRad, H. *et al.* (2001) Root system adjustments: regulation of plant nutrient uptake and growth responses to elevated CO_2. *Oecologia* 126, 305–320
66. Soussana, J.F. *et al.* (2005) A trade-off between nitrogen uptake and use increases responsiveness to elevated CO_2 in infrequently cut mixed C3 grasses. *New Phytol.* 166, 217–230
67. Rubio-Asensio, J.S. and Bloom, A.J. (2017) Inorganic nitrogen form: a major player in wheat and *Arabidopsis* responses to elevated CO_2. *J. Exp. Bot.* 68, 2611–2625
68. Bloom, A.J. *et al.* (2010) Carbon dioxide enrichment inhibits nitrate assimilation in wheat and *Arabidopsis*. *Science* 328, 899–903

69. Guo, H. *et al.* (2013) Elevated CO_2 modifies N acquisition of *Medicato truncatula* by enhancing N fixation and reducing nitrate uptake from soil. *PLoS One* 8, e81373
70. Cott, G.M. *et al.* (2018) Nitrogen uptake kinetics and saltmarsh plant responses to global change. *Sci. Rep.* 8, 5393
71. Jayawardena, D.M. *et al.* (2021) A meta-analysis of the combined effects of elevated carbon dioxide and chronic warming on plant %N, protein content and N-uptake rate. *AoB Plants* 13, plab031
72. Zerihun, A. (2000) Compensatory roles of nitrogen uptake and photosynthetic N-use efficiency in determining plant growth response to elevated CO_2: evaluation using a functional balance model. *Ann. Bot.* 86, 723–730
73. Jayawardena, D.M. *et al.* (2017) Elevated CO_2 plus chronic warming reduce nitrogen uptake and levels or activities of nitrogen-uptake and -assimilatory proteins in tomato roots. *Physiol. Plant.* 159, 354–365
74. Ainsworth, E.A. *et al.* (2006) The effects of elevated CO_2 concentration on soybean gene expression. An analysis of growing and mature leaves. *Plant Physiol.* 142, 135–147
75. Bencke-Malato, M. *et al.* (2019) Short-term responses of soybean roots to individual and combinatorial effects of elevated [CO_2] and water deficit. *Plant Sci.* 280, 283–296
76. Fukayama, H. *et al.* (2009) Rice plant response to long term CO_2 enrichment: gene expression profiling. *Plant Sci.* 177, 203–210
77. Tallis, M.J. *et al.* (2010) The transcriptome of *Populus* in elevated CO_2 reveals increased anthocyanin biosynthesis during delayed autumnal senescence. *New Phytol.* 186, 415–428
78. Taylor, G. *et al.* (2005) The transcriptome of *Populus* in elevated CO_2. *New Phytol.* 167, 143–154
79. Vicente, R. *et al.* (2019) De novo transcriptome analysis of durum wheat flag leaves provides new insights into the regulatory response to elevated CO_2 and high temperature. *Front. Plant Sci.* 10, 1605
80. Jauregui, I. *et al.* (2015) Root and shoot performance of *Arabidopsis thaliana* exposed to elevated CO_2: a physiologic, metabolic and transcriptomic response. *J. Plant Physiol.* 189, 65–76
81. Lin, S.H. *et al.* (2008) Mutation of the *Arabidopsis* NRT1.5 nitrate transporter causes defective root-to-shoot nitrate transport. *Plant Cell* 20, 2514–2528
82. Vicente, R. *et al.* (2017) Improved responses to elevated CO_2 in durum wheat at a low nitrate supply associated with the upregulation of photosynthetic genes and the activation of nitrate assimilation. *Plant Sci.* 260, 119–128
83. Bloom, A.J. *et al.* (2002) Nitrogen assimilation and growth of wheat under elevated carbon dioxide. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1730–1735
84. Bahrami, H. *et al.* (2017) The proportion of nitrate in leaf nitrogen, but not changes in root growth, are associated with decreased grain protein in wheat under elevated [CO_2]. *J. Plant Physiol.* 216, 44–51
85. Bloom, A.J. *et al.* (2014) Nitrate assimilation is inhibited by elevated CO_2 in field-grown wheat. *Nat. Clim. Chang.* 4, 477–480
86. Bloom, A.J. *et al.* (2012) CO_2 enrichment inhibits shoot nitrate assimilation in C3 but not C4 plants and slows growth under nitrate in C3 plants. *Ecology* 93, 355–367
87. Adavi, S.B. and Sathee, L. (2021) Elevated CO_2 differentially regulates root nitrate transporter kinetics in a genotype and nitrate dose-dependent manner. *Plant Sci.* 305, 110807
88. Ferrario-Méry, S. *et al.* (1997) Modulation of carbon and nitrogen metabolism, and of nitrate reductase, in untransformed and transformed *Nicotiana plumbaginifolia* during CO_2 enrichment of plants grown in pots and in hydroponic culture. *Planta* 202, 510–521
89. Vicente, R. *et al.* (2015) Quantitative RT-PCR platform to measure transcript levels of C and N metabolism-related genes in durum wheat: transcript profiles in elevated [CO_2] and high temperature at different levels of N supply. *Plant Cell Physiol.* 56, 1556–1573
90. De la Mata, L. *et al.* (2013) Elevated CO_2 concentrations alter nitrogen metabolism and accelerate senescence in sunflower (*Helianthus annuus* L.) plants. *Plant Soil Environ.* 59, 303–308
91. Geiger, M. *et al.* (1999) The nitrate and ammonium nitrate supply have a major influence on the response of photosynthesis, carbon metabolism, nitrogen metabolism and growth to elevated carbon dioxide in tobacco. *Plant Cell Environ.* 22, 1177–1199
92. Bloom, A.J. *et al.* (2020) Rising atmospheric CO_2 concentration inhibits nitrate assimilation in shoots but enhances it in roots of C3 plants. *Physiol. Plant.* 168, 963–972
93. Kaiser, W.M. *et al.* (2000) Discrepancy between nitrate reduction rates in intact leaves and nitrate reductase activity in leaf extracts: what limits nitrate reduction in situ? *Planta* 210, 801–807
94. Bloom, A.J. (2015) Photorespiration and nitrate assimilation: a major intersection between plant carbon and nitrogen. *Photosynth. Res.* 123, 117–128
95. Quesada, A. *et al.* (2000) Involvement of chloroplast and mitochondria redox valves in nitrate assimilation. *Trends Plant Sci.* 5, 463–464
96. Igamberdiev, A.U. *et al.* (2001) The role of photorespiration in redox and energy balance of photosynthetic plant cells: a study with a barley mutant deficient in glycine decarboxylase. *Physiol. Plant.* 111, 427–438
97. Rachmilevitch, S. *et al.* (2004) Nitrate assimilation in plant shoots depends on photorespiration. *Proc. Natl. Acad. Sci. U. S. A.* 101, 11506–11510
98. Umnajkitikorn, K. *et al.* (2020) Silencing of OsCV (chloroplast vesiculation) maintained photorespiration and N assimilation in rice plants grown under elevated CO_2. *Plant Cell Environ.* 43, 920–933
99. Zhao, H.-L. *et al.* (2021) Potential metabolic mechanisms for inhibited chloroplast nitrogen assimilation under high CO_2. *Plant Physiol.* 187, 1812–1833
100. Aranjuelo, I. *et al.* (2013) Pea plant responsiveness under elevated [CO_2] is conditioned by the N source (N_2 fixation versus NO_3^- fertilization). *Environ. Exp. Bot.* 95, 34–40
101. Domiciano, D. *et al.* (2020) Nitrogen sources and CO_2 concentration synergistically affect the growth and metabolism of tobacco plants. *Photosynth. Res.* 144, 327–339
102. Andrews, M. *et al.* (2019) Elevated CO_2 effects on nitrogen assimilation and growth of C3 vascular plants are similar regardless of N-form assimilated. *J. Exp. Bot.* 70, 683–690
103. Dier, M. *et al.* (2018) Effects of free air carbon dioxide enrichment (FACE) on nitrogen assimilation and growth of winter wheat under nitrate and ammonium fertilization. *Glob. Chang. Biol.* 24, e40–e54
104. Bellegarde, F. *et al.* (2017) Signals and players in the transcriptional regulation of root responses by local and systemic N signaling in *Arabidopsis thaliana*. *J. Exp. Bot.* 68, 2553–2565
105. O'Brien, J.A. *et al.* (2016) Nitrate transport, sensing, and responses in plants. *Mol. Plant* 9, 837–856
106. Li, P. *et al.* (2008) *Arabidopsis* transcript and metabolite profiles: ecotype-specific responses to open-air elevated [CO_2]. *Plant Cell Environ.* 31, 1673–1687
107. Ota, R. *et al.* (2020) Shoot-to-root mobile CEPD-like 2 integrates shoot nitrogen status to systemically regulate nitrate uptake in *Arabidopsis*. *Nat. Commun.* 11, 641
108. Delay, C. *et al.* (2013) CEP genes regulate root and shoot development in response to environmental cues and are specific to seed plants. *J. Exp. Bot.* 64, 5383–5394
109. Li, P. *et al.* (2006) Response diversity of *Arabidopsis thaliana* ecotypes in elevated [CO_2] in the field. *Plant Mol. Biol.* 62, 593–609
110. Munne-Bosch, S. *et al.* (2013) The impact of global change factors on redox signaling underpinning stress tolerance. *Plant Physiol.* 161, 5–19
111. Foyer, C.H. and Noctor, G. (2020) Redox homeostasis and signaling in a higher-CO_2 world. *Annu. Rev. Plant Biol.* 71, 157–182
112. Bellegarde, F. *et al.* (2019) The chromatin factor HNI9 and ELONGATED HYPOCOTYL5 maintain ROS homeostasis under high nitrogen provision. *Plant Physiol.* 180, 582–592
113. Chaput, V. *et al.* (2020) Redox metabolism: the hidden player in carbon and nitrogen signaling? *J. Exp. Bot.* 71, 3816–3826
114. Lejay, L. *et al.* (2008) Oxidative pentose phosphate pathway-dependent sugar sensing as a mechanism for regulation of root ion transporters by photosynthesis. *Plant Physiol.* 146, 2036–2053

115. Fabre, D. *et al.* (2020) Genotypic variation in source and sink traits affects the response of photosynthesis and growth to elevated atmospheric CO_2. *Plant Cell Environ.* 43, 579–593
116. Ingvordsen, C.H. *et al.* (2015) GWAS of barley phenotypes established under future climate conditions of elevated temperature, CO_2, O_3 and elevated temperature and CO_2 combined. *Procedia Environ. Sci.* 29, 164–165
117. Mitterbauer, E. *et al.* (2017) Growth response of 98 barley (*Hordeum vulgare* L.) genotypes to elevated CO_2 and identification of related quantitative trait loci using genome-wide association studies. *Plant Breed.* 136, 483–497
118. Oguchi, R. *et al.* (2022) Enhanced growth rate under elevated CO_2 conditions was observed for transgenic lines of genes identified by intraspecific variation analyses in *Arabidopsis thaliana*. *Plant Mol. Biol.* Published online April 9, 2022. https://doi.org/10.1007/s11103-022-01265-w
119. Rae, A.M. *et al.* (2007) Adaptation of tree growth to elevated CO_2: quantitative trait loci for biomass in *Populus*. *New Phytol.* 175, 59–69
120. Marcos-Barbero, E.L. *et al.* (2021) Genotypic variability on grain yield and grain nutritional quality characteristics of wheat grown under elevated CO_2 and high temperature. *Plants (Basel)* 10, 1043
121. Blandino, M. *et al.* (2020) Elevated CO_2 impact on common wheat (*Triticum aestivum* L.) yield, wholemeal quality, and sanitary risk. *J. Agric. Food Chem.* 68, 10574–10585
122. Medina, S. *et al.* (2016) Interactive effects of elevated [CO_2] and water stress on physiological traits and gene expression during vegetative growth in four durum wheat genotypes. *Front. Plant Sci.* 7, 1738
123. Cseke, L.J. *et al.* (2009) Transcriptomic comparison in the leaves of two aspen genotypes having similar carbon assimilation rates but different partitioning patterns under elevated [CO_2]. *New Phytol.* 182, 891–911
124. Becklin, K.M. *et al.* (2017) CO_2 studies remain key to understanding a future world. *New Phytol.* 214, 34–40
125. Jiang, M. *et al.* (2020) Low phosphorus supply constrains plant responses to elevated CO_2: a meta-analysis. *Glob. Chang. Biol.* 26, 5856–5873
126. Drake, J.E. *et al.* (2011) Increases in the flux of carbon belowground stimulate nitrogen uptake and sustain the long-term enhancement of forest productivity under elevated $CO(_2)$. *Ecol. Lett.* 14, 349–357
127. Finzi, A.C. *et al.* (2007) Increases in nitrogen uptake rather than nitrogen-use efficiency support higher rates of temperate forest productivity under elevated CO_2. *Proc. Natl. Acad. Sci. U. S. A.* 104, 14014–14019
128. Mason, R.E. *et al.* (2022) Evidence, causes, and consequences of declining nitrogen availability in terrestrial ecosystems. *Science* 376, eabh3767
129. Terrer, C. *et al.* (2019) Nitrogen and phosphorus constrain the CO_2 fertilization of global plant biomass. *Nat. Clim. Chang.* 9, 684–689
130. Delhon, P. *et al.* (1996) Diurnal regulation of NO^-_3 uptake in soybean plants IV. Dependence on current photosynthesis and sugar availability to the roots. *J. Exp. Bot.* 47, 893–900
131. Lejay, L. *et al.* (2003) Regulation of root ion transporters by photosynthesis: functional importance and relation with hexokinase. *Plant Cell* 15, 2218–2232
132. Lejay, L. *et al.* (1999) Molecular and functional regulation of two NO^-_3 uptake systems by N- and C-status of *Arabidopsis* plants. *Plant J.* 18, 509–519
133. Eveland, A.L. and Jackson, D.P. (2011) Sugars, signalling, and plant development. *J. Exp. Bot.* 63, 3367–3377
134. Sakuraba, Y. and Yanagisawa, S. (2018) Light signalling-induced regulation of nutrient acquisition and utilisation in plants. *Semin. Cell Dev. Biol.* 83, 123–132
135. Bussell, J.D. *et al.* (2013) Requirement for the plastidial oxidative pentose phosphate pathway for nitrate assimilation in *Arabidopsis*. *Plant J.* 75, 578–591
136. Esposito, S. *et al.* (2003) Glutamate synthesis in barley roots: the role of the plastidic glucose-6-phosphate dehydrogenase. *Planta* 216, 639–647
137. Wang, K. *et al.* (2021) G protein gamma subunit qPE9-1 is involved in rice adaptation under elevated CO_2 concentration by regulating leaf photosynthesis. *Rice (N. Y.)* 14, 67
138. Wu, J. *et al.* (2022) OsGF14b is involved in regulating coarse root and fine root biomass partitioning in response to elevated [CO_2] in rice. *J. Plant Physiol.* 268, 153586
139. Yang, A. *et al.* (2020) A rice small GTPase, Rab6a, is involved in the regulation of grain yield and iron nutrition in response to CO_2 enrichment. *J. Exp. Bot.* 71, 5680–5688
140. Gao, Y. *et al.* (2019) Cisgenic overexpression of cytosolic glutamine synthetase improves nitrogen utilization efficiency in barley and prevents grain protein decline under elevated CO_2. *Plant Biotechnol. J.* 17, 1209–1221
141. Kanno, K. *et al.* (2017) A small decrease in Rubisco content by individual suppression of RBCS genes leads to improvement of photosynthesis and greater biomass production in rice under conditions of elevated CO_2. *Plant Cell Physiol.* 58, 635–642
142. Kirschbaum, M.U. (2011) Does enhanced photosynthesis enhance growth? Lessons learned from CO_2 enrichment studies. *Plant Physiol.* 155, 117–124
143. Dusenge, M.E. *et al.* (2019) Plant carbon metabolism and climate change: elevated CO_2 and temperature impacts on photosynthesis, photorespiration and respiration. *New Phytol.* 221, 32–49
144. Ainsworth, E.A. and Lemonnier, P. (2018) Phloem function: a key to understanding and manipulating plant responses to rising atmospheric [CO_2]? *Curr. Opin. Plant Biol.* 43, 50–56
145. Zhang, J. *et al.* (2020) The potential role of sucrose transport gene expression in the photosynthetic and yield response of rice cultivars to future CO_2 concentration. *Physiol. Plant.* 168, 218–226
146. Dingkuhn, M. *et al.* (2020) The case for improving crop carbon sink strength or plasticity for a CO_2-rich future. *Curr. Opin. Plant Biol.* 56, 259–272
147. Parvin, S. *et al.* (2020) Carbon sink strength of nodules but not other organs modulates photosynthesis of faba bean (*Vicia faba*) grown under elevated [CO_2] and different water supply. *New Phytol.* 227, 132–145
148. Coskun, D. *et al.* (2016) Nutrient constraints on terrestrial carbon fixation: the role of nitrogen. *J. Plant Physiol.* 203, 95–109
149. Ellsworth, D.S. *et al.* (2004) Photosynthesis, carboxylation and leaf nitrogen responses of 16 species to elevated pCO_2 across four free-air CO_2 enrichment experiments in forest, grassland and desert. *Glob. Chang. Biol.* 10, 2121–2138
150. Halpern, M. *et al.* (2018) The role of nitrogen in photosynthetic acclimation to elevated [CO_2] in tomatoes. *Plant Soil* 434, 397–411
151. Pleijel, H. *et al.* (2019) Nitrogen application is required to realize wheat yield stimulation by elevated CO_2 but will not remove the CO_2-induced reduction in grain protein concentration. *Glob. Chang. Biol.* 25, 1868–1876
152. Reich, P.B. *et al.* (2006) Nitrogen limitation constrains sustainability of ecosystem response to CO_2. *Nature* 440, 922–925
153. Zhang, S. *et al.* (2013) Nutrient supply has greater influence than sink strength on photosynthetic adaptation to CO_2 elevation in white birch seedlings. *Plant Sci.* 203-204, 55–62
154. Reich, P.B. and Hobbie, S.E. (2012) Decade-long soil nitrogen constraint on the CO_2 fertilization of plant biomass. *Nat. Clim. Chang.* 3, 278–282
155. Weber, A.P.M. and Bar-Even, A. (2019) Update: improving the efficiency of photosynthetic carbon reactions. *Plant Physiol.* 179, 803–812

1.2 Objectives

The IPCC reports made it clear that the field of agriculture has a very important role to play in the decades to come [IPCC, 2013]. While needing to properly feed a growing population in more and more challenging climates, agricultural practices should also meet sustainability criteria such as limited land use, parsimonious fertilizing inputs, or carbon capture objectives. Our project positions itself in the understanding of the lower nutritional quality of plants under eCO_2, and its relation with the biomass-related fertilization and acclimation effects. More precisely, we wish to identify the mechanisms by which plant growth and physiology are affected by eCO_2 at the genomic scale. In the longer term, and based on the acquired knowledge, we envision the proposition of resilient plants overcoming these deleterious responses.

In light of the state of the art presented in previous sections, we defined several objectives for this project:

1. **Generate genome-wide data from biological material, especially in the roots, to thoroughly investigate nutritional pathways under eCO_2.** This encompasses transcriptomic combinatorial steady-state experiments, but also adaptive responses to eCO_2, combined to different nutritional settings. We focus such data collection on the model plant *Arabidopsis thaliana* because it is easily grown, and possesses high quantities of annotations and existing resources.

2. Based on the generated transcriptomic dataset and other readily available omics data in *Arabidopsis thaliana*, **statistically infer the Gene Regulatory Networks (GRNs) governing root responses to eCO_2, especially those of nutritional pathways.** Such networks should lead to the identification of candidate genes as key regulators of this response. To do so, we develop and adapt statistical inference methods to reconstruct GRNs.

3. Perform a Genome-Wide-Association study (GWAs) in *Arabidopsis thaliana*, with the aim of characterising the natural variability found in the mineral status alteration by eCO_2 of diverse ecotypes. Based on these phenotypes and available sequence information, our objective is then to **identify the genetic determinants associated with the ionome response to eCO_2.**

4. Initiate the experimental validations of the candidate genes put forward in points 2 and 3.

To investigate the nutritional pathways under eCO_2 and carry out the identification of candidate genes, we made use of systems biology.

1.3 Systems biology for candidate genes discovery

In contrast with historical approaches that broke down systems into isolated bits for independent investigation, systems biology has the objective to map and model cell function as a complex, heterogeneous, interacting whole. It is a fundamentally multidisciplinary field, motivated by biological interrogations and employing wet lab experiments, but also bioinformatics, biostatistics, mathematical modelling and,

more recently, machine learning. Because of its predictive nature, systems biology can be used to deepen existing knowledge about a system by proposing a detailed model of its behavior, but also be a tool for making discoveries when studying novel, uncharted responses or organisms.

One of the success stories of systems biology in the domain of plant nutrition is the identification of the BT1-BT2 homologs as negative regulators of Nitrogen Use Efficiency (NUE) under low nitrate conditions [Araus et al., 2016]. Rodrigo Gutierrez and his colleagues made use of supervised machine learning models guided by gene ontology information [Puelma, Gutiérrez, and Soto, 2012] to infer a network of NUE from expression micro-arrays, and extracted from this network central genes including BT2. Experimental validations further confirmed that BT2 and BT1 are indeed able to modulate NUE in *Arabidopsis thaliana* and rice plants [Araus et al., 2016].

Large omics datasets are immensely promising for the understanding of biological systems. They capture entire molecular layers of organisms and cells, such as genome sequence and polymorphisms, transcript abundance, physical interactions between DNA and other molecules, chromatin state, proteins or metabolite quantities. Such global molecular states can be generated at different moments in time to study the dynamics of a system, and spatially to account for tissues and cell-types specificities. The high resolution and number of features in such datasets gave rise to new questions and practices for their statistical analysis.

1.3.1 Statistical methods for omics data analysis

1.3.1.1 Challenges in the analysis of high throughput biological data

In the following lines, some examples of common strategies to statistically deal with omics data are presented, as well as some difficulties arising from the nature of these data.

Answering research questions from omics data often requires **dimension reduction**. Principal Component Analysis (PCA) and its extensions towards non Gaussian distributions [Chiquet, Mariadassou, and Robin, 2018] or towards data integration [Rohart et al., 2017], have the objective to decompose the total variation of a dataset, and to find the combination of the input variables that have the strongest influence in a measured output. They are routinely used to visualise the biological observations in a summarized, low-dimensional space, but also to provide some quantitative insight. For instance, variance decomposition can be used to determine the main experimental conditions causing changes in one or several omic layers, and even to pinpoint the variables contributing to outlier observations [Rau et al., 2019].

Another useful approach is **feature selection**. It is used in the context of exploratory and predictive modelling, and allows to select an interpretable and reasonable number of variables that drive significant changes in a response of interest. It can be used to answer questions like: what are the features of a sequence sufficient to predict its binding by a specific TF? What are the genetic variants responsible for a phenotype of interest? Which are the key genes influencing expression reprogramming? The answer is usually a very small subset of features as compared to the available inputs. Aside from the benefits of encoding sparsity and intelligibility in a model for biologically relevant reasons, feature selection is often a practical requirement. Parametric models such as the linear model and its generalizations can not be estimated when the number of observations is smaller than the number of input features, and they need built-in mechanisms for feature selection. Standard approaches

to feature selection are the LASSO [Tibshirani, 1996], Ridge and Elastic-net penalties [Hastie, Tibshirani, and Friedman, 2001]. Such penalized predictive models are however sensitive to **multicollinearity** in their input features, a phenomena expected to arise from large number of input features, especially in biology where redundancies and clusters of similar behaviour have been documented among input features such as genes, proteins, or other entities.

Hypothesis testing also had to be adapted to high volumes of biological data. It has become usual to test hundreds, thousands, or even more hypotheses simultaneously, for example when searching for differentially expressed genes (DEGs) across whole transcriptomes. The multiplication of statistical tests has made systems biology vulnerable to spurious deviations from null hypotheses and data snooping, *i.e* the increased chance of getting artifactually significant results by running an uncontrolled number of tests. Corrections for this problem include defining, prior to the analysis, the exact test procedures that will be used, applying corrections to p-values like Bonferroni or Benjamini and Hochberg's false discovery rate, and independently confirming significant results, either on test data untouched during hypothesis formulation and testing, or with new wet lab experiments.

1.3.1.2 Standard analysis pipeline for transcriptomic data

Expression data has been one of the first type of omic data available for large scale statistical analysis, first through micro-arrays and then RNA-Seq. Transcriptomes can be leveraged to answer biological questions of high interest, like: which genes have their expression significantly changed by a perturbation? What are the genes with a different expression value in a genotype of interest as compared to a wild type organism? Can we distinguish groups of genes with similar behaviours in a set of different conditions or developmental stages? How to model interactions and dependencies between genes in a given response? Consequently, transcriptomic datasets have been booming in the last decades (1.5).

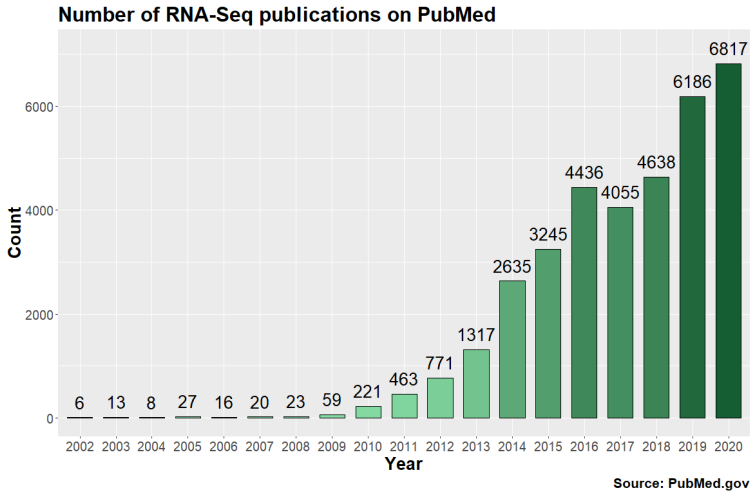


FIGURE 1.5: Increasing number of publications involving the word "RNA-Seq" in PubMed

In practice, RNA-Seq data is obtained from biological material after the extraction of RNA molecules in the cells of interest, and the preparation of the samples to the convenience of the sequencing platform. Common sequencing technologies

like Illumina provide the sequence of nucleotides corresponding to the fragments of mRNAs, generally of a size around 150 base pairs. Other technologies like PacBio or Nanopore can sequence fragments of a much larger size [Hu et al., 2021], which makes them great tools for genome assembly, but less useful for quantitative analyses: they were not used in the scope of this project. Raw sequencing files in the fasta or fastq format are the basis of a bioinformatics pre-processing pipeline:

1. The **quality control** of raw sequences is performed to remove low quality nucleotides and to trim adapters inserted at the sequencing step. Quality control also provides descriptive metrics like sequencing coverage, GC content or the distribution of insert sizes [Chen et al., 2018].

2. The **mapping** of sequenced reads to a reference genome. At this step, an alignment algorithm reports the unique or multiple locations in the reference genome where the reads could be aligned, along with alignment scores, in the bam/sam format. Many options control the stringency of the mapping, such as the maximum number of mismatches per alignment or the minimum alignment length [Dobin et al., 2012].

3. The mapping results are compared to the organism gene models in the GFF format in order to **count how many reads were aligned into the genes**. It is possible to choose how to count multiple or ambiguous alignments, and the regions of interest for the counting like entire genes, composed of the union of all their exons, or gene isoforms to study alternative splicing [Anders, Pyl, and Huber, 2014]. Counting the mRNA abundance of all genes in all samples results in an expression table or matrix, with the genes in the different rows and the N samples in columns.

4. **Normalization** ensures that the expression levels in each samples are comparable to each other, or that the expression of genes are comparable to each other, depending on the envisionned downstream statistical analyses. In particular, for the statistical analyses described below, normalization techniques such as the Median of Ratios [Love, Huber, and Anders, 2014] or TMM [Robinson and Oshlack, 2010] correcting for different sequencing depths between samples are required and routinely used.

Once the normalized expression table is available, it can be statistically analysed to understand gene expression programs. Statistical analyses to answer biological questions can be very diverse depending on the problematic, organisms, or experimental designs. Listed below are some very common types of analyses performed on RNA-Seq data to extract knowledge:

- **Differential Expression Analysis** (DEA) is carried out to identify genes that have a significant change in expression between experimental conditions. DEA in its most widespread form for RNA-Seq models gene expression as over-dispersed counts via Negative Binomial distributions. A first step is to estimate the over-dispersion parameter. Then, generalized linear models are fit to the expression values of each gene, and their coefficients are used to test the significance of the experimental variations on gene expression. This results in one p-value per gene, later corrected for multiple testing. Such models are commonly brought to researchers by the DESeq2 **[deseq2]** or edgeR [McCarthy, Chen, and Smyth, 2012] implementations. DEA results can also be controlled

by a minimum absolute log fold change between conditions, to enforce a sufficient amplitude in expression change.

- **Gene Ontology** (GO) enrichment analyses are useful to understand the functional content of a list of genes. Such analyses determine if genes belonging to certain biological functions, cellular processes or compartments are significantly over-represented in a list of interest as compared to a background (usually the entire transcriptome). Fisher's exact test and its hypergeometric null distribution are employed by common tools like clusterProfiler [Yu et al., 2012] to assess the over-representation of ontologies found in a list of genes.

- **Co-expression clustering** is a way to determine groups of genes with a common expression behavior across available experimental conditions. The partitioning of genes into co-expression clusters can be solved by the k-means algorithm, that groups genes based on their similarity across samples. Probabilistic equivalents to k-means have also been explored. Instead of characterizing clusters of genes by their centroids (mean gene expression in each of the N samples) like in k-means, clusters are characterized by a N-dimensional probability density function appropriate for count data whose parameters are estimated via Expectation-Maximization methods. Tools like coseq [Rau and Maugis-Rabusseau, 2018] bring such approaches to the community.

- **Network inference** uses transcriptomic data to reconstruct the complex structure of expression reprogramming in a given response. In order to model a specific response, network inference is usually restricted to a set of genes of interest, typically derived from DEA or co-expression clusters of interest. Statistical network inference from expression profiles is a central topic in systems biology, and various statistical frameworks have been developed to make predictions about the interactions between genes in the context of transcriptional regulation. Network inference is further introduced in section 1.3.2.

These steps of statistical analyses were performed on all the transcriptomic datasets we generated in the course of this project, but also on existing datasets that we directly analysed for methods development purposes. In particular, we make this pipeline available in the DIANE suite, with a precise choice of statistical methods for each step (Publication #2).

The investigation of transcriptomic data through statistics, whether by combining DEA, GO enrichment, clustering, network inference or other type of analyses, often results in the formulation of new research hypotheses. Such hypotheses extend existing knowledge and make new claims about the biological systems under study: they are to be tested in new experiments until they are either confirmed or not, and finally contribute to establishing knowledge.

1.3.2 Reconstructing Gene Regulatory Networks

1.3.2.1 The regulation of gene expression

Modelling gene regulation relies on some prior understanding of the biological principles at stake. The regulation of gene expression, that is to say the quantity of messenger RNA (mRNA) of genes, is at the core of cell function and acts in various molecular levels.

1. During **transcriptional** regulation, one or several Transcription Factors (TFs) bind to the gene regulatory region and contribute to the recruitment of the RNA polymerase that transcribes the gene into a mRNA. TFs bind to DNA as they recognize specific sequences of nucleotides, also called cis-regulatory motifs, or binding sites (TFBSs). TFBSs are usually localized in the gene promoter, but can also extend to other regions, like introns, or distant enhancers. Distant enhancers coming in contact with the promoter to regulate transcription have not been, however, much documented in *Arabidopsis thaliana* like in human or other mammals.

2. **Epigenetic** regulation refers to all non sequence-encoded forms of regulation. These epigenetic regulations are chemical modifications (such as methylation) of nucleic acids or histones that can impact the ability of proteins to bind to DNA and proceed to transcription, but also mRNA stability or translation.

3. **Post-transcriptional** regulations target mRNA transcripts before their translation, in the way they are capped, spliced, or exported from the nucleus. Three prime untranslated regions also contain regulatory instructions, and microRNAs or proteins can bind to those regions to inhibit translation or decrease the expression level of mRNAs. Moreover, mechanisms acting upon the stability and degradation rates of mRNA can regulate gene expression, as well as chemical modifications added to mRNAs like the methylation mark m6A.

Combined, those different layers of regulation allow cells and multi-cellular organisms to grow and adapt to their environment. Given that plants are sessile organisms, they are especially vulnerable to environmental changes, relying on fine-tuned and efficient gene expression reprogramming [López-Maury, Marguerat, and Bähler, 2008]. Regulatory programmes in higher organisms such as plants responding to the environment have several characteristics, described in the following lines.

First, gene expression reprogramming is **proportionate and dynamic**. After an external perturbation, gene expression is temporally impacted, with the expression of transcription factors and genes peaking or dropping in regulatory cascades, and then returning to a former or new steady-state [Li, Varala, and Coruzzi, 2015]. In Arabidopsis roots, the temporal response to N induction exhibits such characteristics [Varala et al., 2018; Brooks et al., 2019]. The intensity of expression changes was also shown to be positively correlated to the intensity of the environmental stimulus in several settings. [López-Maury, Marguerat, and Bähler, 2008]

Second, gene expression reprogramming displays **tissue-specificity**. Different organs or cell types can respond in a specialized and coordinated way. Different tissues may also adapt the expression levels of some gene modules responsible for short or long-distance signalling influencing the expression of target genes in other tissues. This was, for instance, observed in the signalling of nitrate starvation in Arabidopsis with the differential expression of glutaredoxins [Tabata et al., 2014; Ota et al., 2020], and with shoot-to-root signals carried by the small peptids of the CEP family.

Thirdly, gene expression reprogramming is **noisy**. As compared to gene expression in the course of development, gene expression resulting from environmental perturbations exhibits high levels of variability between cells and individuals [López-Maury, Marguerat, and Bähler, 2008; Cortijo et al., 2019]. This variability likely provides beneficial phenotypic diversity in the context of adaptation and even evolution. However, this causes technical challenges when statistically deciphering such noisy signals.

1.3.2.2 Modelling biological systems as networks

The regulation of gene expression involves thousands of genes, mRNA, proteins, connected through several layers of cellular processes. A network model suits the representation of such processes quite well, because it describes in a single object a set of entities and their relations to each other. More formally, a network or graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is composed of:

1. A set of nodes, also called **vertices** $\mathcal{V}$. They represent biological entities of interest like genes, proteins, or transcripts. Several types of vertices can coexist in the same network.

2. A set of links, also called **edges** $\mathcal{E}$, that are pairs of vertices $\in \mathcal{V}$. The set $\mathcal{E}$ is included in $\mathcal{V}$x$\mathcal{V}$. Edges represent the interactions between biological entities, and can also be of several types, for example activation, inhibition, regulation, co-expression, etc. Edges can be undirected or can be attributed a certain orientation, in this case the graph is said to be directed and $\mathcal{E}$ must be composed of ordered pairs of vertices $\in \mathcal{V}$. Some types of graphs can contain at the same time directed and undirected edges [Perković, Kalisch, and Maathuis, 2017], for example when networks are built from various sources of information. Links can also be weighted by a certain value, for example to quantify the strength of the interaction (Figure 1.6).

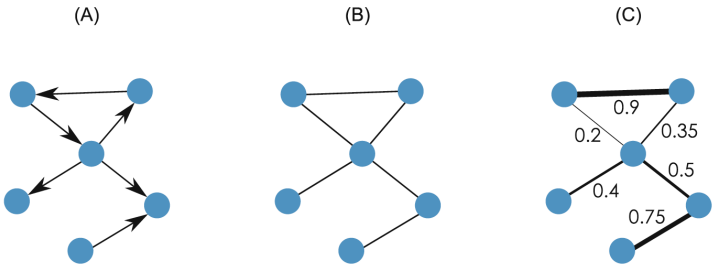


FIGURE 1.6: Examples of network types: directed (a), undirected (b), and weighted (c), where the weights are represented by edge thickness. Note that a weighted network can be directed or undirected. Source Gene Regulatory Networks, 2019 [Sanguinetti and Huynh-Thu, 2019]:

A biological network has a strong duality: it can be topologically described with the abstract tools of graph theory and discrete mathematics, while telling a semantic story about the biology of living organisms. This dual nature requires the cautious definition and disambiguation of its components in order to extract reliable interpretations. The meaning of a network stems from the type of its biological entities, the type of links between them, and the way that those links are constructed. Here are some examples of different kinds of biological networks:

- **Protein-protein interaction networks**. Nodes are proteins, and an undirected link is present between two proteins if they were reported as physically interacting.

- **Co-expression networks**. A relation of co-expression between two genes means that their expression levels vary in conjunction in all the experimental samples

available. Edges in co-expression networks are not directed, as co-expression is a symmetric relationship.

- **Gene Regulatory networks** (GRNs) contain oriented edges that represent a relationship of regulation between a regulator and its target.

To describe networks, topological properties are routinely used. The **density** of biological networks is often low, which means that biological networks are **sparse** [Leclerc, 2008; Koutrouli et al., 2020; Campos and Freyre-González, 2019; Hayes, Sun, and Pržulj, 2013]. Network density is defined as $\frac{E}{E_{max}}$, with E the number of edges of the network, and E_{max} the maximal number of edges of a network with the same nodes. For example, in a co-expression network with N nodes, $E_{max} = \frac{N(N-1)}{2}$. In an oriented regulatory network with T genes in total and R regulators, $E_{max} = R(T-1)$. Well characterized networks, when investigated, reveal densities between 0.1 and 0.001, with density decreasing as the number of nodes increases [Campos and Freyre-González, 2019; Hayes, Sun, and Pržulj, 2013]. Figure 1.7 shows this trend with several networks taken from the literature or protein-protein interactions in several organisms.

The **degree distribution** in biological networks is heavy-tailed, and has been traditionally approximated by power-law distributions [Koutrouli et al., 2020]. Even though the ubiquity of scale-freeness among real-world networks is under debate [Broido and Clauset, 2019], the degree distribution of biological networks is usually such that many nodes have a low degree, while rare genes have a high degree and behave like hubs.

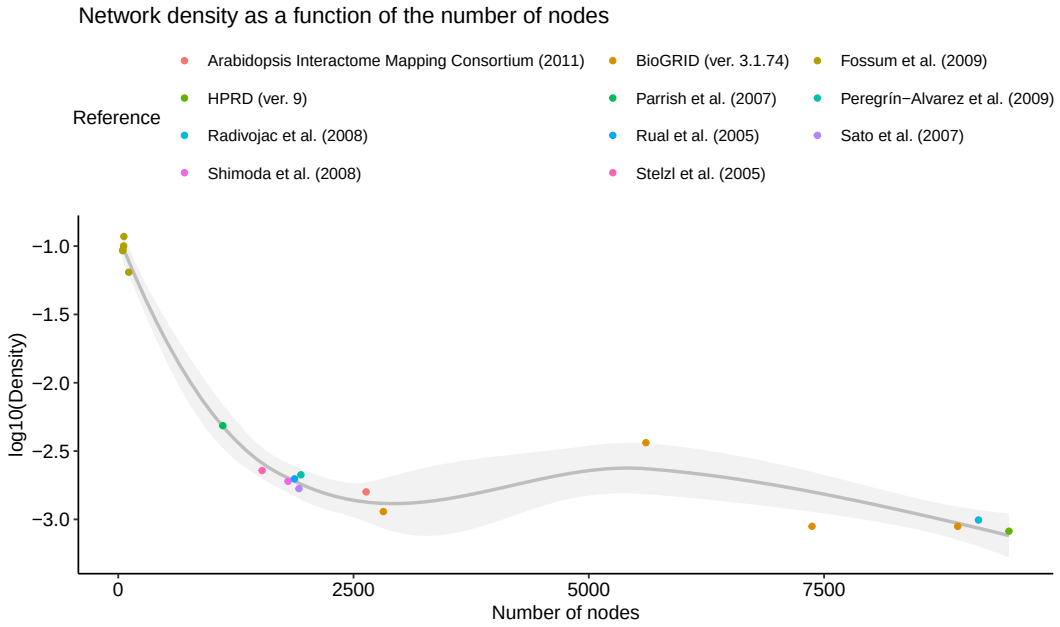


FIGURE 1.7: Log_{10} of protein-protein interaction network densities as reported in [Hayes, Sun, and Pržulj, 2013], depending on the number of entities in the network. ($Log_{10}(0.1) = -1$. $Log_{10}(0.001) = -3$)

Finally, biological networks have the tendency of being globally sparse, but to locally form tightly connected communities. This property can be measured via the **transitivity** metric, also called **clustering coefficient** [Koutrouli et al., 2020].

The network representation of a biological process offers great visualization possibilities. Software like Cytoscape or R packages like igraph or vizNetwork are great tools to explore and interact with networks in different layouts. Networks can be the foundation of biological discoveries. For example, **topological communities** can be delimited and studied, to retrieve groups of entities sharing common neighbors and biological roles. In co-expression networks and GRNs, gene communities are often mined for functional enrichment, as co-expressed genes or groups of co-regulated targets are likely involved in common pathways. Under this assumption, a GRN was used to successfully predict the function of unknown TFs as related to ROS signalling, based on their topological neighborhood [Clercq et al., 2021].

Networks are also often used to identify prevalent nodes, representing important entities in a response. The study of network **centralities** suits this purpose, and different types of centralities can be considered. **Degree** centrality of a node is the most trivial centrality measure, defined as the number of connections made by that node. Other centrality metrics can be used, for example **eccentricity**, which is the longest path from the considered node to any other node, or **betweenness**, which is the number of shortest paths in the network connecting all possible pairs of nodes passing through the considered node. In several works in *Arabidopsis thaliana*, such metrics calculated in inferred transcriptional networks were used to describe genes, rank them by order of importance and put forward candidate genes [Cheng et al., 2021; Araus et al., 2016; Clercq et al., 2021].

Main network reconstruction types The means of obtaining biological networks can be very diverse. Two broad categories appear:

- **Data extraction methods** retrieve networks from existing knowledge. Text mining algorithms and language processing can be used on literature articles or research databases to summarise results from different sources and studies. Their main disadvantage is their incapacity to discover new interactions.
- **Statistical methods** make use of experimental datasets, potentially combined with prior knowledge, to infer biological networks. Statistical methods predict networks from -omics data, a probabilistic model and an estimation method. Their predictions remain to be further investigated and validated, but they have the advantage to enable discoveries of previously unknown interactions.

What follows focuses on the category of statistical methods. Because the transcriptional reprogramming of mineral nutrition pathways under eCO_2 is unknown, it can not be directly retrieved from existing literature and databases. Instead, it needs to be statistically inferred from new experimental data, generated under environmental conditions adapted to the question of interest.

1.3.2.3 Input data for GRN statistical inference

The type of network to be inferred conditions the choice of input data. For instance, networks with the objective of modelling epigenetic relations will have to rely on methylome data or chromatin conformation. Binding site networks modelling binding potential between TFs and their targets will be based on regulatory sequence information. Co-expression networks will be formed from expression data. In the context of this project, the desired networks are GRNs, because they attempt to model

causal links between regulators and target genes, and can be used to extract important regulator genes in the eCO_2 response. Consequently, we present the different types of data that can be leveraged in the processes of GRN inference in Table 1.1. It should be noted that the process of GRN inference goes beyond the estimation of the network topology from the data, but also encompasses the steps of calibrating the model parameters, and evaluating the inference performance: biological data can be involved at any of these steps.

In addition to the data from Table 1.1, some community efforts like ATRM or REMAP [Chèneby et al., 2019] gather information about gene regulation in the literature, and may contain data from RNA-Seq, ChIP-Seq, DAP-Seq, TARGET, TFBS motifs, but also a restricted number of more functionally characterized interactions. Another example of a recent database about regulation in plants is ConnecTF [Brooks et al., 2020], that contains, for Arabidopsis, 26 ChIP-Seq experiments, 382 DAP-Seq experiments, and the interactions from the TARGET assay applied to 33 N responsive TFs identified in a dynamic study of N induction in Arabidopsis roots [Varala et al., 2018; Brooks et al., 2019]. Such heterogeneous databases are crucial tools to decipher gene regulation. Indeed, binding does not necessarily imply expression modulation, and expression modulation does not necessarily imply binding: there is a paradoxically low intersection of targets that are bound and that have their expression level altered by a TF [Alvarez et al., 2020]. Thus, the joint use of data relative to expression changes and to the physical interactions between TFs and their targets is of great interest to capture the complexity of GRNs.

In the following section, we detail some statistical GRN inference methods that are based on expression data. Indeed, expression data has been historically the most widespread type of data available genome-wide for GRN inference, from microarrays to, more recently, RNA-Seq. This explains the wide ecosystem of statistical methods to infer GRN from expression only. Further in this work, the use of other types of data in the course of inference validation, or for integrative GRN inference, will be detailed.

1.3.2.4 GRN inference methods from expression data

The challenges of GRN inference The complexity of statistical GRN inference from expression data mainly stems from:

- **The curse of dimensionality**. The complexity of a desired GRN is usually larger in respect to available data to infer it. In fact, the order of magnitude of the number of possible links in a GRN involving T genes will be T^2. However, the available expression data is an expression matrix of dimension $T * N$, with N the number of experimental samples. If $T > N$, which is almost always the case, it follows that the the number of edge coefficients to infer is larger than the total number of observations at one's disposal, and the problem is said to be **underdetermined** [De Smet and Marchal, 2010]. This means that distinct candidate GRNs can explain the observed expression data equally well, and that without additional heuristics and prior knowledge, it is impossible to discriminate between them.

- **The nature of expression data**, that is likely to be noisy or subject to technical biases. In addition, there are hidden layers of regulation that are not measured by transcriptomic datasets such as chromatin conformation, post-transcriptional or post-translational changes. This makes it harder to predict

TABLE 1.1: Brief description of -omics data commonly used for GRN inference in *Arabidopsis thaliana*, with their strengths and weaknesses.

Data	Description	Advantages	Disadvantages
RNA-Seq	Genome-wide transcript abundance	• Relatively low cost • Whole transcriptome	• Can not distinguish direct from indirect regulations • Does not report regulatory events
ChIP-Seq	*In vivo* genome-wide DNA binding sites of TFs. The bound DNA is coprecipitated, purified, and sequenced.	• Low technical biases • Preserved cellular context	• Expensive and experimentally challenging: few TFs available • No information on expression
DAP-Seq	*In vitro* genome-wide DNA binding sites of TFs. The bound DNA is coprecipitated, purified, and sequenced. [O'Malley et al., 2016]	• Less costly than ChIP-Seq: many TFs available	• Loss of chromatin context • Specificity lower than ChIP-Seq • No information on expression
TARGET	*In vivo* genome-wide direct targets of TFs. Expression changes are measured after the nuclear import of TFs. The targets are direct because of a translation inhibitor [Bargmann et al., 2013]	• Less costly than ChIP-Seq: many TFs available • The effect of a regulator on the target's expression is captured.	Notable technical biases: • Done in protoplasts • Nuclear import of TFs too sudden and abundant to be realistic
TFBS motifs	Motif search of TFs binding site in the targets' regulatory sequences. Motifs can be given in the form of PWMs available in databases like JASPAR	• Very low cost (*in silico*)	• The binding motif of many TFs is unknown • False occurrences during motif search • No information on expression
ATAC-Seq	Regions of open chromatin, accessible for regulation	• Restricts potential target genes to accessible ones • Can uncover TF footprints	• Expensive and experimentally challenging • No information on expression

accurate links between genes, especially if their relationship in the context of regulation is not purely transcriptional [Banf and Rhee, 2017].

- **The scarcity of validation data for GRN inference**. True GRNs are not known,

they are at best partially mapped in certain specific conditions or model organisms. This makes the use of supervised learning arduous, because supervised models need to learn regulatory patterns from sufficiently large amounts of known interactions. When regulatory interactions are documented, they are usually also submitted to noise, condition specificity, or technical biases. Negative examples of regulation are also needed in supervised algorithms to learn the features in the data that do not result in a regulatory link, but they are even more challenging to exhibit with certainty as existing regulatory databases exclusively contain links supporting regulation. Some examples of negative examples can be found in works like iGRN [Clercq et al., 2021]. Even when using unsupervised algorithms, the scarcity of validation data complicates the tasks of estimating the model parameters and validating the inferred GRNs. GRN validation and available metrics are further discussed in section 1.3.2.5.

Overview of statistical approaches for GRN inference from expression data Many literature reviews describe the vast landscape of network inference methods, detailing and categorising most of the available computational and statistical approaches [Banf and Rhee, 2017; Barbosa et al., 2018; Lecca, 2021; Mercatelli et al., 2020; Mochida et al., 2018; Sanguinetti and Huynh-Thu, 2019]. The line between GRN and co-expression network inference techniques can be blurry. In fact, GRN inference methods can be inspired from co-expression methods, and some co-expression techniques are extended toward GRN inference when pruned for indirect interactions. For this reason, a broad description of network inference methods from expression data is given in the following lines, with their statistical formulation, advantages and drawbacks.

1. Correlation and mutual information methods A first category of inference techniques are relevance networks. They are based on **similarity measures** between the expression profiles of genes, relying on **correlation** or **information theory**. Linear correlation between the N expression values of gene 1 (vector Y_1) and gene 2 (vector Y_2) is defined as:

$$\rho_{12} = \frac{cov(Y_1, Y_2)}{\sigma_1.\sigma_2}$$

where $cov(Y_1, Y_2)$ can be estimated by

$$\frac{\Sigma_{y_1 \in Y_1, y_2 \in Y_2}(y_1 - \mu_1)(y_2 - \mu_2)}{N - 1}$$

with μ_1, μ_2 the means of Y_1 and Y_2, and σ_1, σ_2 the standard deviations of Y_1 and Y_2.

The mutual information between genes 1 and 2 is defined, in the cases of discrete data like counts, as:

$$I_{1,2} = \Sigma_{y_1 \in Y_1} \Sigma_{y_2 \in Y_2} p(y_1, y_2) log\Big(\frac{p(y_1, y_2)}{p(y_1)p(y_2)}\Big)$$

with $p(y_1, y_2)$ the joint probability distribution of $y_1 \in Y_1$ and $y_2 \in Y_2$, and $p(y_1), p(y_2)$ are the marginal probability distributions of $y_1 \in Y_1$ and $y_2 \in Y_2$.

The core principle of correlation and mutual information methods is to quantify the degree of joint evolution of the expression between all pairs of genes, and then to select, on the basis of a threshold, the interactions of genes having the most similar or anti-correlated expression profiles. A famous correlation-based method is

WGCNA [Langfelder and Horvath, 2008], while well known information theoretic approaches are ARACNE [Margolin et al., 2006] or CLR [Faith et al., 2007]. However, the interactions predicted by such methods should be carefully interpreted because two genes can have their expression correlated because they are under the control of a common regulator, and thus end up linked in a relevance network, whereas it is an indirect link. Even if methods like WGCNA, ARACNE, and CLR attempt to prune indirect interactions by various strategies, the initial inference was still under the assumption of symmetric undirected edges, which is closer to co-expression than regulation and direct causation [Banf and Rhee, 2017; Sanguinetti and Huynh-Thu, 2019].

2. Partial Correlation methods An extension of correlation networks are networks based on **partial correlation**, in the framework of Gaussian Graphical Models (GGMs). Those techniques are able to infer the structure of gene networks by modelling their conditional dependencies. In particular, GGMs are a sound statistical and computational framework designed to estimate the inverse of the variance-covariance matrix $\Theta = \Sigma^{-1}$ of the genes, also called the precision matrix. Its dimension is T x T, with T still the total number of genes. When $\Theta_{12} = 0$, genes 1 and 2 are conditionally independent given the $T - 2$ other genes under study, and they are consequently not linked in the final network. This sparse precision matrix is usually estimated by maximizing the likelihood under regularization constraints. To model count data, that are not normally distributed, GGMs have been extended to Poisson Log Normal models and applied to RNA-Seq data for network inference [Choi et al., 2017]. They have also been extended to infer networks from multiple types of data [Chiquet, Rigaill, and Sundqvist, 2019]. A drawback of GGMs and extensions are the challenge of the precision matrix estimation from highly correlated data and under high dimensional settings, especially when scaling to large modern transcriptomic data. In addition, GGMs rely on strong assumptions about the data distribution, the relationship between genes are assumed to be linear and they only model undirected symmetric interactions [Sanguinetti and Huynh-Thu, 2019].

3. Bayesian network approaches The definition of Bayesian networks relies on a directed graph describing conditional dependencies. The probability of the entire expression data can be decomposed as the product of the probability of observing each gene conditionally on their parents in the graph:

$$P(Y_1, Y_2, ..., Y_G|\mathcal{G}) = \Pi_{g=1}^{G} P(Y_g|\text{Parents}(Y_g, \mathcal{G}))$$

This setting makes it easy to integrate prior knowledge about the links between genes, and is also appealing to model causation. The exploration and evaluation of different graph structures can be performed greedily by maximum likelihood approaches or by Markov Chains Monte Carlo sampling methods. This is however an extensive search, that grows exponentially with the number of genes, and those methods remain, even to date, difficult to scale to large datasets. In those models, the interpretation of oriented edges has to be cautious as different graph structures can represent the same set of conditional dependencies (if the graphs belong to the same equivalence class) [Maathuis et al., 2018]. Another constraint is that Bayesian methods require that the inferred graphs are directed and can not contain any loop structures. This has been overcome with the Dynamic Bayesian Networks (DBNs) introduced by [Friedman and Murphy, 1998; Murphy, Mian, et al., 1999] to model

time series gene expression data, and allowing cycles and loops. In DBNs, the parents of random variables are random variables at the previous time point. The framing of DBNs, allowing cycles, has facilitated their estimation because the conditional probabilities could now be modeled separately by the statistical laws of choice. For example, in the case of continuous normal variables, the parameters of a DBN model can be estimated via linear or non-linear regression on the past expression of potential parents. This makes DBNs closely related to time-lagged regression, also called auto-regressive models [Fujita et al., 2007; Michailidis and Buc, 2013], eventually with a time-varying graph structure [Dondelinger, Lèbre, and Husmeier, 2013; Lebre et al., 2010; Kim, 2003]. DBN modelling and inference have also been extended to model RNA-Seq data wuth a Negative Binomial distribution [Thorne, 2018] A drawback is that DBNs can only be used when time data is available.

4. Boolean methods Boolean methods start by discretizing the expression values as 0 or 1, standing for inactivated or activated states. With those binary values, a model is constructed by finding one logical function per gene, depending on the binary values of other genes, or the same gene at other time points. A directed edge is drawn from each variable used in the function toward the predicted gene. Boolean networks permit the modelling of loops and provide very interpretable and predictive networks, but discretizing expression values is very difficult because we don't known the threshold values for the activation of each genes, that can also depend on the conditions. This is thus a strong limitation. They are also limited to small datasets, as the number of nodes they model can not exceed a few tens, and can suffer from structure incorrectness in benchmarks [Liang, Fuhrman, and Somogyi, 1998; Pušnik et al., 2022].

5. Regression methods The core principle of regression-based methods is that the expression variations of regulator genes hold descriptive and predictive power over the expression variations of their target genes. Regression methods thus often decompose the problem of learning a GRN into learning one regression function per gene. Lets consider that the expression level of a target gene t form a response variable noted Y_t, and that the expression level of R potential regulator genes form the predictive variables $X = (X_1, X_2, ..., X_R)$. The regression problem for the target gene t is then framed as:

$$Y_t = f_t(X) + \epsilon_t$$

with ϵ_t the model error. Regression based methods have predictive abilities, which can be desirable: given the expression levels of regulator genes in new experimental conditions, a regression model would be able to predict the expression of the target in those new conditions. Besides, those methods are designed to model the relationship between a target gene and R regulators.

Selecting influential regulators for a target gene comes down to a feature extraction problem: fitting f_t enables the identification of the subset of regulators that are sufficient to accurately explain the expression of Y_t. These most influential regulators, once determined, will be attributed outgoing edges toward the target in the network. Methods based on regression mainly differ in their modelling choices for the function f_t, and in their way to perform feature selection based on f_t. Two broad categories of methods exist, based either on **linear or non-linear choices**.

Firstly, linear regression techniques model the expression of a target gene $y_{t,i}$ in the experimental condition i as a linear combination of the expression of the regulators $x_{r,i}$:

$$y_{t,i} = \beta_{t,0} + \Sigma_{r=1}^{R} \beta_{t,r} x_{r,i} + \epsilon_{t,i} \tag{1.1}$$

The coefficients $\beta_{t,r}$ quantify the influence of regulator r on the target gene t. It is the conditional covariance between Y_t and X_r given the other $R-1$ regulators, and divided by the variance of R_r. A direct link can be made with the framework of Gaussian Graphical Models [Lauritzen, 1996], as $\beta_{t,r}$ can be expressed as the ratio between two entries of the precision matrix Θ [Meinshausen and Bühlmann, 2006]:

$$\beta_{t,r} = -\frac{\theta_{tr}}{\theta_{tt}}. \tag{1.2}$$

However, the number of regulators typically outnumbers the number of expression measures per gene. This high dimensional setting, combined with the objective of modelling biological sparsity, mandates some form of regularization.

Regularization in high dimension are popularly performed via the LASSO, ridge, or elastic-net operators. In the context of GRN inference, the LASSO (Least Absolute Shrinkage and Selection Operator) is preferred as it attributes a coefficient of exactly 0 to uninformative variables, offering as a consequence sparser solutions. Under the LASSO penalty, the coefficients in Equation 1.1 are estimated with two objectives: minimizing the prediction error, and keeping the L_1 norm of the vector of coefficients smaller than a certain quantity. Formally, the estimated coefficients $\beta_{t,r}$ correspond to:

$$\text{argmin}_{\beta_t} \frac{1}{2N} \Sigma_{i=1}^{N} (y_{t,i} - \hat{y}_{t,i})^2 + \lambda \Sigma_{r=1}^{R} |\beta_{t,r}|, \tag{1.3}$$

with $\hat{y}_{t,i} = \beta_{t,0} + \Sigma_{r=1}^{R} (\beta_{t,r} x_{r,i})$ and N the number of experimental conditions in which gene expression was measured. Figure 1.8 graphically illustrates this constrained optimisation, and the sparsity enforced by the LASSO.

The LASSO has already been used successfully for GRN inference in methods such as LASSO-StARS [Miraldi et al., 2019], a method that was included in the Inferelator [Skok-Gibbs et al., 2022], or other works [Qin et al., 2014]. Exploiting Equation 1.2, sparse linear regression with the LASSO has even been used in the context of infering the precision matrix Θ of GGMs in the high dimensional setting, as a way of selecting relevant neighboring nodes of each genes in the graph of conditional dependencies [Meinshausen and Bühlmann, 2006]. Linear regression was also employed with another kind of feature selection: the TIGRESS method [Haury et al., 2012] uses LARS [Efron et al., 2004], a kind of forward stepwise feature selection closely related to the LASSO. Such regularization techniques are however vulnerable to multi-colinearity: if several variables are highly correlated, the LASSO or LARS approaches will become unstable and arbitrarily select one of the correlated features with no guarantee about its relevance. Because such colinearity is to be expected in large sets of potential regulator genes, this issue has been addressed via Stability Selection [Meinshausen and Bühlmann, 2010] in TIGRESS and LASSO-StARS. Stability Selection consists in running a large number of times an estimation procedure and, at each run, re-sampling the samples and the variables. The results are then aggregated into a more robust model for feature selection. In TIGRESS, the influence of a regulator over a target gene is computed as its selection frequency in a large number of LARS runs. In LASSO-StARS, Stability Selection is implemented

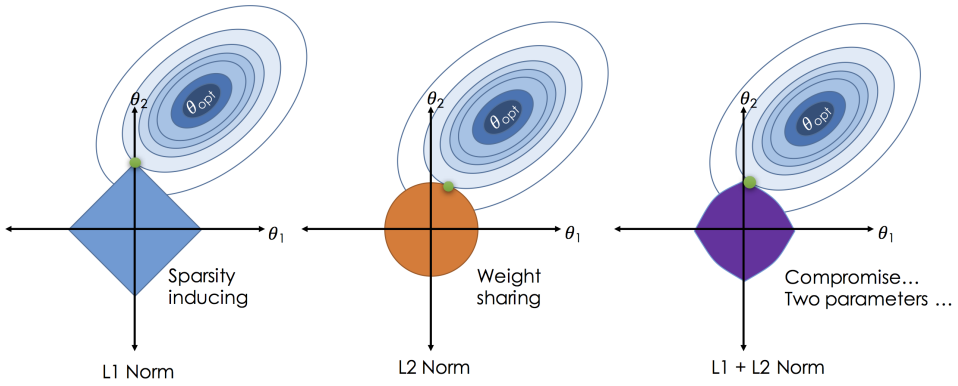


FIGURE 1.8: Different choices of norm for regularisation illustrated in a two dimensional space of coefficients. The optimal value of the coefficients θ_{opt} without regularization would be found at the center of the concentric ellipsoids. In the case of the L_1 regularization (the LASSO operator), θ_1 is put to 0 to satisfy both the norm constraint and the prediction error minimisation (left). L_2 norm corresponds to the ridge operator (center), and the combination of two norms is the elastic-net operator (right). The shapes of the unit spheres of each norm illustrates how the LASSO enforces more sparsity. Image source : https://medium.com.

according to the StARS methods [Liu, Roeder, and Wasserman, 2010], to choose a trade-off value of λ along the regularization path that provides both robustness in the variable selection, and sparsity.

Linear modelling of gene expression regulation suffers from some limitations, like:

- Linear regression makes assumptions about the normality of the data. When normality assumptions are not met, the validity of linear regression models can not be guaranteed. This usually forces practitioners to log-transform the expression data before GRN inference, or to employ Generalized linear models with a Poisson or Negative Binomial model.
- The interactions between regulators can not be modelled. As the regression problem is already in a high dimensional settings, the addition of interaction terms, increasing exponentially with the number of regulators, dramatically aggravates this issue.
- The relation between the transcript levels of a regulator and the transcript levels of its target is not necessarily linear: complex or non monotonous links, as well as step functions can reasonably be expected under some scenarios of regulation.

These arguments have motivated the choice of non linear functions for f_t. Ensembles of trees are the most popular option, and were first introduced with GENIE3 [Huynh-Thu et al., 2010]. In GENIE3, the expression of a target gene is linked to the expression of the regulators via Random Forest (RF) regression. For each gene, a regression tree is composed of test nodes. A test node is a decision rule that represents a condition on the expression level of a regulator. The succession of test nodes is found by an iterative learning algorithm, testing all possible combinations of regulators and expression thresholds, and selecting the one that best discriminates the

expression values of the response variable. The leaves of the regression trees are predictions of the expression of the target gene $\hat{y}_{t,i}$, computed as the mean of all observations ending up in this leaf. Regression trees are more flexible than linear regression because they do not make assumptions about data distribution, they can model complex non linear relations between regulators and targets, and also take into account interaction effects between regulators. Regression trees alone are however sensible to noise and over-fitting, so aggregating large numbers of trees into a Random Forest allows more robust and generalizable results [Breiman et al., 2017]. Each tree of a RF is learned from a different bootstrapped set of experimental conditions, and have also a restricted choice of regulators when learning each test nodes (i.e usually drawn from a random set of only $\sqrt{R}$ regulators), which increases diversity between trees.

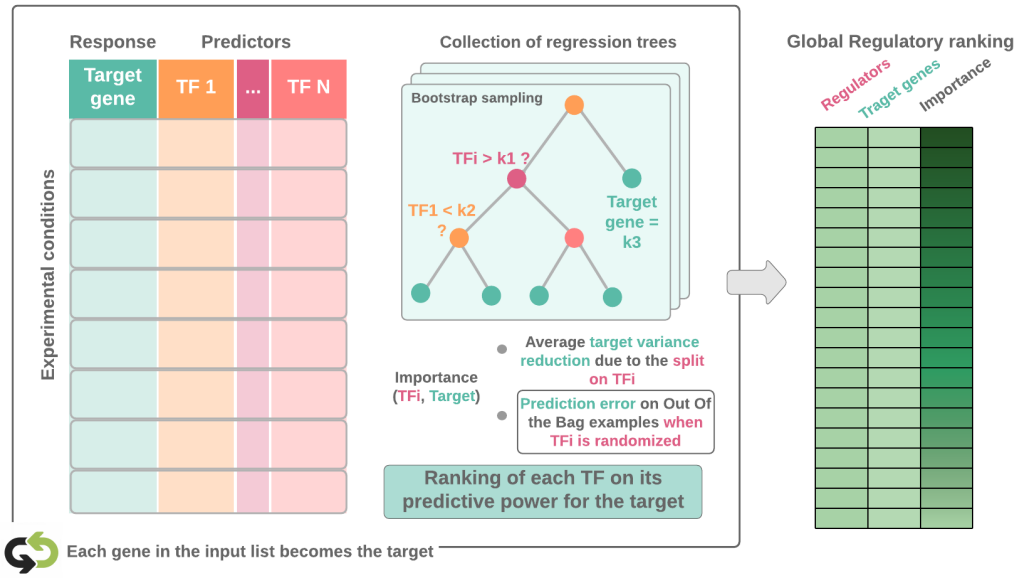


FIGURE 1.9: GRN inference process via Random Forests. The learning sample on the left is the basis for learning a RF in which test nodes represent conditions on the expression of regulator genes, and predict the expression of the target gene. This process is repeated for each target gene, for which the influence of the regulators are extracted, and lead to a global ranking of regulatory interactions.

In the context of GRN inference, RFs are used to measure the influence of the regulators for a target gene, by calculating their feature importance. Several types of feature importance can then be extracted from RFs. GENIE3 uses **node impurity** $I(N_t)$, defined for a specific test node as:

$$I(N_t) = N_{\text{t,upstream}}\text{Var}(Y_{t,\text{upstream}}) - (N_{t,\text{left}}\text{Var}(Y_{t,\text{left}}) + N_{t,\text{right}}\text{Var}(Y_{t,\text{right}}))$$

with $Y_{t,\text{upstream}}$, $Y_{t,\text{left}}$ and $Y_{t,\text{right}}$ the expression values of the target genes before the test node, on the left downstream side of the test node, or on the right downstream side of the test node. Their sizes are respectively $N_{t,\text{upstream}}$, $N_{t,\text{t,left}}$ and $N_{t,\text{right}}$. Node impurity is interpretable as the variance reduction in the response induced by the split on a specific regulator. To get the importance of a regulator, the node impurities of all test nodes concerning this regulator in a regression tree are summed,

and then averaged across all trees in the Forest. This metric however gives a higher importance to the regulators of genes with a large variance, so it requires the normalization of the expression data so that genes have a unit variance.

Another importance metric is the **relative Mean Decrease in Accuracy** (MDA). It is defined as the prediction deterioration when a regulator is randomized, as compared to prediction error the original data. The prediction error of RF regression is usually the Mean Squared Error (MSE) on Out Of Bag examples (OOB), *i.e* the experimental conditions that were not sampled in the bootstrapped sets to fit the trees.

$$MSE_t = \frac{1}{N_{OOB}} \Sigma_{i \in OOB} (y_{t,i} - \hat{y}_{t,i})^2$$

$$MDA_{t,r} = \frac{MSE_{t,\text{r shuffled}} - MSE_t}{MSE_t} * 100$$

The relative MDA metric, as defined in the above equation, has the advantage of being computed on OOB, and is thus less sensible to overfitting, while not requiring any prior data scaling.

To summarize, the inference procedure is as follows: one RF is learned for each target gene, the influence of the regulators on each target are extracted from the estimated RF, and a global ranking of all regulatory interactions is built (Figure 1.9). This ranking of all possible pairs of regulators and target genes is made on the basis of the chosen importance metric (Node impurity or relative MDA). To build a final GRN, the strongest interactions can be selected. GRN inference via RFs was further extended in the direction of temporal data with dynGENIE3 [Geurts et al., 2018], Outpredict [Cirrone et al., 2020], or with other types of trees like Boosted Trees in BTNET [Park et al., 2018]. Ensemble learning, and more broadly Machine Learning (ML) has become a promising area for GRN inference, with the development of tools unifying the use and benchmark of several ML algorithm like SVM, Boosting, RFs, like proposed in GReNaDIne.

Finally, regression-based methods are related to other types of GRN inference, namely **temporal methods**. The information of time is highly valuable in the context of deciphering gene regulation, and can be used in mainly two contexts:

- **Time-lagged models**, in which the expression of the regulators in the time point n is used to predict the expression of the target in the following time point $n+1$:

$$Y_{t,n+1} = f_t(X_n) + \epsilon_t$$

- **Differential equation models**, in which the rate of expression change of the target gene is modelled as a function of the expression levels of the target gene and of its potential regulators. If we note k the number of time units separating times n and $n+1$:

$$\frac{Y_{t,n+1} - Y_{t,n}}{k} = f_t(X_n, Y_{t,n}) + \epsilon_t$$

In both cases of temporal schemes, f_t can be determined through linear and non-linear regression. A detailed description of temporal methods is, however, not in the scope of this work.

1.3.2.5 Evaluation of GRN inference methods

The output of GRN inference is a **predictive model** of gene regulation. In order to assess to quality of predicted GRNs, several strategies have been implemented.

Experimental validation An option is to engage in wet-lab experiments to test predicted interactions between genes. For example, one could study the expression levels of the predicted targets of a regulator in a knock-out mutant of this regulator, in the environmental conditions used for network inference. Genes that have their expression altered in the mutant genotype compared to the wild type could be considered as the direct or indirect targets of the regulator. Other approaches like TARGET could be set up in a lab to determine direct targets of a TF. In addition, the generation of ChIP-Seq data for TFs of interest could be relevant to check that a regulator binds to its targets in the predicted GRN. If they are performed in the same environments and conditions than the transcriptomic data from network inference, this is a direct and high quality way of validating a restricted number of hypotheses derived from GRN models, like the transcriptional or phenotypic role of candidate genes. However, a strong limitation is that experimental validations are costly, time consuming, and could not be reasonably envisioned for the entire set of inferred interactions of a GRN (possibly up to thousands of links). Generally, experimental validations are only set up for genes of high interest, like candidate genes with a central role in a GRN, but can not be sufficient to globally estimate the performance of GRN inference. Instances of wet-lab analyses focused on candidate regulators from GRN can be seen in several works in Arabidopsis, mostly relying on candidate mutant phenotyping [Araus et al., 2016; Cheng et al., 2021; Clercq et al., 2021]. These concrete examples illustrate that GRN models are an efficient way of identifying important genes in a response.

Evaluation of inference based on external regulation databases Because of the cost of new experiments, GRN quality is commonly estimated using readily available external data relative to regulation. Depending on the organisms under study, the extent of available data varies a lot. In *Arabidopsis thaliana*, several types of omics data are generally used, such as physical contact between TFs and DNA regions provided *in vivo* by ChIP-Seq or *in vitro* by DAP-Seq. Assays like TARGET [Bargmann et al., 2013], were also more recently brought to a medium throughput and represent a valuable resource. In addition, validation information can also be brought by complementary information like scanning promoter regions for TFBSs. Complementary sources of data that can be used, especially in Arabidopsis, are summarized in Table 1.1. Such external information sources relative to regulatory interactions will serve as a reference goal for GRN evaluation, called gold standard.

The most common approach to validation is to see GRN inference as a binary classification task predicting wether there is an interaction or not between two genes. The predicted interactions are then compared to the gold standard. Metrics from statistical learning are used, namely **true positives** (the predicted interactions present in the gold standard), **false positives** (the predicted interactions that are not in the gold standard), **true negatives** (the interactions that are neither in the gold standard nor in the predicted interactions), and **false negatives** (the interactions in the gold standard but that were not predicted) [Schrynemackers, Küffner, and Geurts, 2013].

In machine learning, a common way to assesses classification performance is the area under the ROC curve (AUC). The ROC curve is formed by true positive rates and false positive rates for all possible network densities. However, GRN inference

suffers from great class imbalance as there are far more nonexistent regulatory links than existing ones. AUC was shown to be more sensible to class imbalance [Zhang and Janssen, 2006], so the area under this precision-recall curve (AUPR) is now a widely accepted metric for GRN evaluation [Banf and Rhee, 2017; Marbach et al., 2012b].

In this case, precision and recall can be calculated. Precision measures how many predicted interactions are accurate, while recall measures how many accurate interactions are predicted:

$$\text{Precision} = \frac{\text{\#True positives}}{\text{\#Predicted interactions}} = \frac{\text{\#True positives}}{\text{\#True positives + \#False positives}}$$

$$\text{Recall} = \frac{\text{\#True positives}}{\text{\#Gold standard interactions}} = \frac{\text{\#True positives}}{\text{\#True positives + \#False negatives}}$$

To globally assess model performance, precision and recall can be computed for all possible density values, ranging from no edges to a fully connected GRN. This is achieved by computing precision and recall for all possible thresholds determining the prediction of an interaction or not, from the most stringent threshold to no threshold at all. This forms the precision-recall curve (AUPR). The performance criteria is then the area under such a curve.

Several community efforts have been led to compare and evaluate the ecosystem of existing GRN inference methods, and employed AUPR. The most prominent and cited one is the Dialogue on Reverse Engineering Assessment and Methods (DREAM) initiative [Marbach et al., 2012b]. In this study, the accuracy of a broad panel of inference methods was measured, from *Escherichia coli*, *Saccharomyces cerevisiae*, and *in silico* microarray data. Gold standards were taken from RegulonDB for E. coli, and from ChIP-binding data, conserved TF motives, or systematic transcription factor deletions. The main findings of the benchmarks in DREAM were that no method had a clear advantage over all others and on all datasets, and that there was a lot of room for improving inference performance. The benchmark gave however an advantage to GENIE3, as it shows an overall score higher than other methods in Figure 1.10. Among linear regression techniques, TIGRESS was the overall best performer.

Still, several drawbacks of this kind of evaluation can be exhibited and greatly challenge GRN validation:

- Validation data is scarce. Binding experiments or regulation assays like TARGET are performed on one TF at a time, and are costly. Functionally studied and characterized interactions in the literature are even rarer. This results in a relatively small fraction of predicted interactions involving TFs that were previously studied and for which gold standard information is available. The precision and recall can thus by estimated using a restricted subset of predicted interaction, which is likely to increase chances of bias and reduce the confidence attributed to the value of AUPR [Banf and Rhee, 2017]. This makes the task of comparing models based on AUPR even more difficult.

- All omics datasets are a specific molecular snapshot of regulation, are thus limited to certain mechanisms, conditions, tissues, regulators, or may have technical biases. *In vitro* techniques like DAP-Seq are for instance, likely to miss

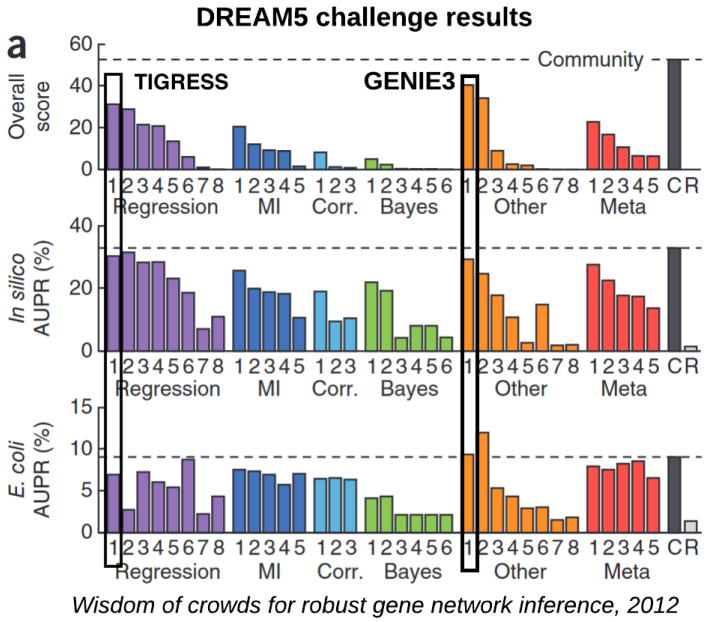


FIGURE 1.10: Benchmark results from the DREAM challenge. Each number represents a competing method, evaluated by the Area Under the Precision-Recall curve on two datasets, and their overall score. The previously described TIGRESS and GENIE3 methods are highlighted. Colors represent categories of methods based on their core statistical framework [Marbach et al., 2012b]

relevant information about the cellular context. ChIP-Seq experiments have the advantage of taking into account chromatin state, TFs combinatorics in a specific cellular environment. However, the targets of a regulator based on ChIP-Seq are dependent on the environment, and the condition-specific regulations predicted in GRNs can only be validated by ChIP-Seq experiments performed under similar conditions. These series of imperfections cause gold standard based on such data to be partly inaccurate, and contain themselves false negatives and false positives. To address this, validation efforts are based on simulated gold standards [Seçilmiş, Hillerton, and Sonnhammer, 2022; Bellot et al., 2015], but this approach is reliant on our ability to generate realistic gene expression data, which is debatable.

Evaluation of GRN properties Because of the imperfection of GRN inference validation based on an experimental gold standard, a quality criteria can also be the **biological relevance** of GRN structure [Banf and Rhee, 2017]. For example, inferred GRNs can be compared to what is currently known about biological networks, like the fact that they are **sparse** and form **highly modular** structure with **genes communities** and **hubs**. Such characteristics can be quantitatively measured and combined for a complementary assessment of GRN inference. The biological relevance of inferred interactions based on **prior knowledge** about genes has also been applied: for example we can increase our confidence that an inferred network is accurate if co-regulated genes share common functions and use this a validation criteria [Marbach et al., 2012a]. Finally, certain categories of GRN inference methods are predictive,

like regression-based methods. This makes it possible to evaluate their performance in predicting the expression levels of target genes on experimental conditions not seen in GRN model training. This is for example the case in the benchmark of Out-Predict, in which **prediction error**, the **mean square error** (MSE), serves as a comparative indicator between competing algorithms [Cirrone et al., 2020].

1.3.2.6 Summary of statistical contributions to GRN inference

In this manuscript, we use the term "GRN" in a narrower sense than the definition given in section 1.3.2.2. The broad definition of a GRN allows all genes to be considered as potential regulators and act upon the expression of other genes. In this sense, the number of candidate regulators R is equal to the total number of genes T. In this work (and as done by others), we narrow down this definition by restricting the set of candidate regulators to genes annotated as transcriptional regulators. This is done by making use of existing knowledge in *Arabidopsis thaliana*, where a significant number of regulators has been identified. This setting where $R < T$ has two advantages:

1. It alleviates the problem of high dimension
2. It provides interpretations closer to causality by reducing the chances of obtaining edges from co-expression rather than regulation

In light of the current state of the art, we made the decision to explore regression-based techniques for GRN inference. Indeed, such methods are easily scalable, multivariate, and describe regulation in a formulation oriented toward causality. They have proven in benchmarks that, even though their performance was limited, it exceeded the performance of other statistical approaches [Marbach et al., 2012b], especially in the case of ensemble of trees like GENIE3 [Huynh-Thu et al., 2010].

However, regression-based methods suffer from some limitations. For example, the extraction of regulatory influences from regression models provides a fully connected weighted GRN, but the way to threshold this fully connected GRN was not in the scope of GENIE3's original publication, and there is to date, no consensus on how to optimally obtain a sparse GRN. Given this improvement potential, we propose in Publication #2 to refine GENIE3 inference by assessing the statistical significance of predicted interactions, and benchmark the precision gain of this procedure on experimental gold standards.

In addition, regression based methods suffer from the curse of dimensionality, colinearity, and their performance could be limited by their exclusive use of transcriptomic data. For these reasons, leveraging other types of omics not only for model evaluation, but for model training as well, is considered in this work. Publication #4 provides an overview of existing solutions for data integration in regression-based inference while improving and exploring two popular types of regression models for GRN inference in a context of TFBS and expression integration. Those integrative regression approaches are compared based on precision and recall, prediction error, and biological relevance.

1.3.3 Genome Wide Association studies (GWAs)

In the previous sections, we describe the identification of candidate genes based on GRN inference in a reference genotype responding to the environment. Instead, important genes in a given response can also be discovered based on the natural variation found in genetically diverse populations of individuals.

1.3.3.1 Leveraging natural variability to identify genes of interest

The principle of association studies is to make statistical associations between one or several phenotypic traits, and the presence of genetic polymorphisms in large group of individuals. In this context, if a polymorphism is statistically associated to a phenotype, this locus and its close vicinity in the genome are potential causal elements involved in the regulation of this trait, and deserve to be prioritized in functional studies. Conducting a GWAs requires building up a large and diverse dataset on two levels:

1. **Phenotypic measures** on each individual must be obtained, and the distribution of these quantitative phenotypes should demonstrate a sufficient amount of variability to be further explored.

2. **Genotype information** must be available for each individual, and inform on wether or not they possess genetic or epigenetic variants in a high-resolution series of loci in the genome. The information of Single Nucleotide Polymorphisms (SNPs) is mostly used in GWAs.

1.3.3.2 Statistical methods for genotype-phenotype associations

GWAs modelling relies on a regression framework, in which the phenotypes of the individuals are expressed as a function of their genetic polymorphisms. Traditional approaches employ linear additive models, while more recent approaches can be based on non-linear regression and algorithms from the machine learning field [Nicholls et al., 2020]. The linear multivariate approach would express the phenotype of individual i as

$$y_i = \beta_0 + \Sigma_{k=1}^{M} \beta_k X_{ik} + \epsilon_i$$

with X_{ik} the presence or absence of the SNP k in individual i, β_k the coefficient encoding the effect of SNP k on the phenotype, M the total number of SNPs, and ϵ_i a gaussian noise centered in 0. However, the ultra-high dimension of the problem ($M >> N$, with N the number of individuals in the study) does not permit its resolution, and it has been current practice to apply regressions using each SNP separately as a descriptor of the phenotype. In this setting, another difficulty is that association studies often include individuals from heterogeneous population backgrounds. Individuals from a given sample, panel or cohort can be related to each other through hidden relatedness, or originate from different populations, causing sample stratification. Failing to account for population structure has been shown to inflate test statistics in association studies and subsequently lead to spurious associations [Kang et al., 2010]. Although the true relatedness status between individuals is not known, the high-density genomic data of the sample can be used as a basis to estimate it. Several solutions have been proposed, namely applying a genomic control inflation factor [Devlin and Roeder, 1999], or using the first principal components of a genomic PCA as covariates during regression. To date, a common and successful strategy has been the Linear Mixed Model (LMM). LMMs are fit to the vector of phenotypes Y for each marker separately, modelling **the impact of the marker as a fixed effect**, and taking into account the polygenic effects of other SNPs via a random effect, also named **additive genetic variance component**. Let's consider the LMM relative to the SNP k for the individual i :

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_i \qquad \text{Var}(\eta_i) \propto \sigma_a^2 \Phi + \sigma_e^2 I$$

where , Φ is the $N * N$ matrix of genetic relatedness between the individuals, and I is the identity matrix. Several solutions exist in order to calculate the kinship matrix Φ, like for example Identity By State, Astle [Astle and Balding, 2009], or VanRaden [VanRaden, 2008]. Once β_k has been estimated, the null hypothesis $\beta_k = 0$ is tested and provides a p-value relative to the significance of the effect of the marker k on Y. Using the estimates of σ_a and σ_e, an approximation of heritability can be given by the value of $\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$. The Efficient Mixed Model Association (EMMA) was among the first variant components models [Kang et al., 2008], and was later extended and scaled-up for large datasets by implementations like EMMA-eXpedited (EMMAX) [Kang et al., 2010].

Results from association studies are usually presented in the form of Manhattan plots, showing a transformation of the p-values of each marker along their position in the genome. Peaks in a Manhattan plot indicate candidate regions with a strong link to the phenotype. Another common representation of GWAs results is the quantile-quantile plot (qq-plot), that displays the quantiles of the p-values distribution from LMMs against the quantiles of the p-values expected by chance (i.e there is no association between SNPs and the phenotype). Under proper modelling, observed p-values should mainly match random p-values, except for a small number of true causal associations departing from the expected p-values.

1.3.3.3 Validation of candidate regions obtained by GWAs

Markers attributed low p-values by LMMs are then generally investigated to establish wether there is a true association, and uncover the cause and mechanisms underlying this association. Firstly, the vicinity of candidate markers is to be explored because of **linkage disequilibrium**. Linkage desequilibrium is the joint conservation of alleles at different loci in the course of recombination. This causes polymorphisms in proximity to be almost always observed together. It is thus required to investigate direct regions in which polymorphisms are found, but also a broader window of the size of linkage desequilibrium around those SNPs.

The interpretations and experimental validations to engage in will depend on the location of the strongly associated SNPs.

- If an associated SNP is found in a promoter region, or in an intron, this SNP could act as an **expression variant**. Such SNPs modulate the expression of surrounding genes by affecting their translation. For instance, expression variants can be found in TFBSs, and change the likelihood that a TF will bind the a regulatory region. In order to test the effect of expression variants, the expression of their putative target genes can be measured in individuals possessing the variant, and in individuals that do not.

- If an associated SNP is found in the coding region of a gene, it can result in three types of mutation. **Silent** or synonymous mutations do not alter the sequence of amino acids. **Nonsense** mutations stop the translation of the protein, which is truncated. Finally, **missense** mutations change the sequence of the amino acids of the protein. Nonsense and missense variants can be validated by the study of mutant organisms in which the candidate gene has lost its function. If the phenotype of such mutants differs from the wild type, the candidate gene is very likely to play an important role in this trait.

Before initiating functional experiments, genetic validations can be made by allele complementation showing that the trait of interest is restored by a mutation

induced on a candidate causal variant. Once cases of expression variants or coding sequence alterations have been established, the mechanisms by which the phenotype is altered by the candidate gene can be further studied and functionally characterized by new experiments.

1.3.3.4 Summary of the statistical analysis for our GWAs

In light of the current state of the art, we made the decision to rely on existing methods for our GWAs. We chose the modelling provided by LMMs to test associations between polymorphisms and phenotypes while adjusting for population structure.

Chapter 2

Statistical inference of the Gene Regulatory Networks in *Arabidopsis thaliana* under elevated CO_2 combined to nutritional limitations

2.1 Dashboard for the Inference and Analysis of Network from Expression data (DIANE)

2.1.1 Preamble

Before interrogating genome-wide expression data and inferring a GRN to understand the adaptation to eCO_2 in Arabidopsis, we first reflected on the transcriptomic analysis methods that would compose our pipeline. This reflection lead to the choice of a precise set of tools, and two conclusions. First, many graphical user interfaces conducting statistical analyses on transcriptomic data provide satisfactory methods for common steps, but more advanced statistical procedures like co-expression clustering and GRN inference were rarely included. Second, we chose the GRN inference sofwtare GENIE3. However the regulatory ranking it returns does not directly form a GRN and it has to be sparsified to build a final parsimonious GRN. As we detail in the introduction of Publication #2, we found that few existing solutions were both interpretable and practical enough in the general case. We thus developed an extension to GENIE3 that is based on permutation procedures to assess the significance of regulatory interactions.

Motivated by reproducible statistical analyses, we shared our pipeline and extension for GRN inference, via a graphical user interface deployed online, that also comes as an R package: the Dashboard for the Inference and Analysis of Network from Expression data (DIANE).

2.1.2 Publication #2 (Published)

Note : This section has its own reference system. Citation numbers refer to bibliography items included in the present article, and not at the end of the PhD manuscript. ***This manuscript was published in BMC Genomics in May, 2021.***

Cassan *et al. BMC Genomics* (2021) 22:387
https://doi.org/10.1186/s12864-021-07659-2

BMC Genomics

SOFTWARE Open Access

Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite

Océane Cassan[1*], Sophie Lèbre[2,3] and Antoine Martin[1]

Abstract

Background: High-throughput transcriptomic datasets are often examined to discover new actors and regulators of a biological response. To this end, graphical interfaces have been developed and allow a broad range of users to conduct standard analyses from RNA-seq data, even with little programming experience. Although existing solutions usually provide adequate procedures for normalization, exploration or differential expression, more advanced features, such as gene clustering or regulatory network inference, often miss or do not reflect current state of the art methodologies.

Results: We developed here a user interface called DIANE (Dashboard for the Inference and Analysis of Networks from Expression data) designed to harness the potential of multi-factorial expression datasets from any organisms through a precise set of methods. DIANE interactive workflow provides normalization, dimensionality reduction, differential expression and ontology enrichment. Gene clustering can be performed and explored via configurable Mixture Models, and Random Forests are used to infer gene regulatory networks. DIANE also includes a novel procedure to assess the statistical significance of regulator-target influence measures based on permutations for Random Forest importance metrics. All along the pipeline, session reports and results can be downloaded to ensure clear and reproducible analyses.

Conclusions: We demonstrate the value and the benefits of DIANE using a recently published data set describing the transcriptional response of Arabidopsis thaliana under the combination of temperature, drought and salinity perturbations. We show that DIANE can intuitively carry out informative exploration and statistical procedures with RNA-Seq data, perform model based gene expression profiles clustering and go further into gene network reconstruction, providing relevant candidate genes or signalling pathways to explore. DIANE is available as a web service (https://diane.bpmp.inrae.fr), or can be installed and locally launched as a complete R package.

Keywords: Gene regulatory network inference, Graphical user interface, Multifactorial transcriptomic analysis, Model-based clustering, Analysis workflow

*Correspondence: oceane.cassan@cnrs.fr
[1]BPMP, CNRS, INRAE, Institut Agro, Univ Montpellier, 34060 Montpellier, France
Full list of author information is available at the end of the article

Background

Analyzing gene expression to uncover regulatory mechanisms

A multitude of regulatory pathways have evolved in living organisms in order to properly orchestrate development, or to adapt to environmental constraints. Much of these regulatory pathways involve a reprogramming of genome expression, which is essential to acquire a cell identity corresponding to given internal and external environments. To characterize these regulatory pathways, and translate these changes in gene expression at the genome-wide level, global transcriptome study under various species, tissues, cells and biological conditions has become a fundamental and routinely performed experiment for biologists. To do so, sequencing of RNA (RNA-Seq) is now the most popular and exploited technique in next-generation sequencing (NGS) methods, and underwent a great expansion in the field functional genomics. RNA-seq will generate fragments, or short reads, that match to genes and quantitatively translate their level of expression. Standard analysis pipelines and consensus methodological frameworks have been established for RNA-Seq. Following quality control of data, reads mapping to a reference genome, and quantification on features of interest are performed, several major steps are commonly found in RNA-Seq data analysis. They usually consist in proper sample-wise normalization, identification of differential gene expression, ontology enrichment among sets of genes, clustering, co-expression studies or regulatory pathways reconstruction.

However, these analysis procedures often require important prior knowledge and skills in statistics and computer programming. In addition, tools dedicated to analysis, exploration, visualization and valorization of RNA-Seq data are very often dispersed. Most of RNA-Seq data are therefore not properly analyzed and exploited at their highest potential, due to this lack of dedicated tools that could be handled and used by (almost) anyone.

Current tools for facilitating the exploitation of RNA-seq data

Over the last few years, several tools have emerged to ease the processing of RNA-Seq data analysis, by bringing graphical interfaces to users with little programming experience. Among those tools are DEBrowser [1], DEApp [2], iGEAk [3], DEIVA [4], Shiny-Seq [5], IRIS-DEA [6], iDEP [7], or TCC-GUI [8]. All of them propose normalization and low count genes removal, exploratory transcriptome visualizations such as Principal Component Analysis (PCA), and per-sample count distributions plots. They also provide functions for interactive Differential Expression Analysis (DEA) and corresponding visualizations such as the MA-plot. Gene Ontology (GO) enrichment analysis can be performed in those applications, apart from IRIS-DEA, DEApp, and TCC-GUI.

However, when it comes to further advanced analyses such as gene expression profiles clustering or network reconstruction, solutions in those tools are either absent, or sub-optimal in terms of statistical framework or adequacy with certain biological questions. For instance, most of those applications perform clustering using similarity based methods such as k-means and hierarchical clustering, requiring both the choice of metric and criterion to be user-optimized, as well as the selection of the number of clusters. Probabilistic models such as Mixture Models are a great alternative [9–11], especially thanks to their rigorous framework to determine the number of clusters, but they are not represented in currently available tools.

Regarding Gene Regulatory Networks (GRN) inference, only three of the applications cited above propose a solution. Two of them, iDEP and Shiny-Seq rely on the popular WGCNA framework (WeiGhted Correlation Network Analysis) [12], which falls into the category of correlation networks. This inference method have the disadvantage of being very vulnerable to false positives as it easily captures indirect or spurious interactions. When the number of samples in the experiment is low or moderate, high correlations are often accidentally found [13]. Besides, linear correlations like Pearson coefficient can miss complex non-linear effects. Lastly, WGCNA addresses the question of co-expression networks, more than GRN. To infer GRN, which should link Transcription Factors (TF) to target genes, iGEAK retrieves information from external interaction databases and binding motives. This allows to exploit valuable information, but makes this step extremely dependent on already publicly available datasets. An exhaustive comparison with respect to the features and methods handled by the described interfaces for RNA-Seq analysis is given in Fig. 1.

Other frameworks focus on gene network reconstruction and visualization only. For instance, the web server GeNeCK [14] makes the combination of several probabilistic inference strategies easily available, but there is no possibility to select a subset of genes to be considered as regulators during inference. The online tool ShinyBN [15] performs Bayesian network inference and visualization. This Bayesian approach is however prohibitive when large scale datasets are involved. Lastly, neither ShinyBN nor GeNecK allow for upstream analyses and exploration of RNA-Seq expression data.

Consequently, efficient statistical and machine learning approaches for GRN inference (like for instance GENIE3 [16], TIGRESS [17], or PLNModels [18], see [19] for a review) are not available, to our knowledge, as a graphical user interfaces allowing necessary upstrem operations like normalization or DEA.

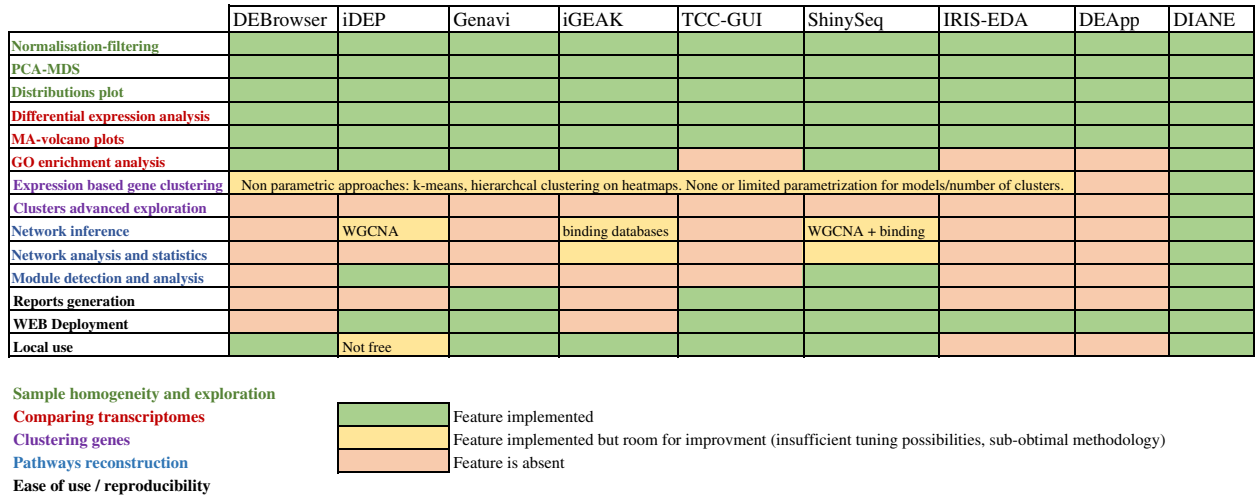


Fig. 1 Comparison of tools for facilitating the valorization of expression datasets. Eight interactive tools for analysis of count data from RNA-Seq are presented here and compared in terms of features and methodological choices. The features included are the ones we believe are the expectation from most users willing to exploit RNA-Seq experiments and understand regulatory mechanisms, and that we included to DIANE. Although not reported here for clarity reasons, many compared tools had their own features and specificities of interest. For instance, IRIS-DEA handles single cell RNA-Seq and facilitates GEO submission of the data, iDEP enables to build protein-protein interaction network and has an impressive organisms database, while Shiny-Seq can summarise results directly into power point presentations

Besides, all of the cited applications are available as online tools or as local packages with source code, although the useful possibility to provide both solutions simultaneously, in order to satisfy advanced users as much as occasional ones, is not always available. It is also worth noting that availability of organisms in current services varies a lot. Some of them like iGEAK are restricted to human or mouse only.

Proposed approach

In this article, we propose a new R-Shiny tool called DIANE (Dashboard for the Inference and Analysis of Networks from Expression data), both as an online application and as a fully encoded R package. DIANE performs gold-standard interactive operations on RNA-Seq datasets, possibly multi-factorial, for any organism (normalization, DEA, visualization, GO enrichment, data exploration, etc.), while pushing further the clustering and network inference possibilities for the community. Clustering exploits Mixture Models including RNA-seq data prior transformations [11] and GRN inference uses Random Forests [16, 20], a non-parametric machine learning method based on a collection of regression trees. In addition, a dedicated statistical approach, based on both the biological networks sparsity and the estimation of empirical p-values, is proposed for the selection of the edges. Step-by-step reporting is included all along the analyses, allowing reproducible and traceable experiments.

In order to illustrate the different features of DIANE, we have used a recently published RNA-seq data set, describing the combinatorial effects of salt (S), osmotic (M), and heat (H) stresses in the model plant *Arabidopsis thaliana* [21]. RNA-seq were performed under single (H, S, M), double (SM, SH, MH), and triple (SMH) combinations of salt, osmotic, and heat stresses. In the course of our paper, we will demonstrate that DIANE can be a simple and straightforward tool to override common tools for transcriptome analyses, and can easily and robustly lead to GRN inference and to the identification of candidate genes.

Implementation and results

DIANE is an R Shiny [22, 23] application available as an online web service, as well as a package for local use. To perform relevant bioinformatic and bio-statistical work, different existing CRAN and Bioconductor packages as well as novel functions are brought together. Its development was carried out via the golem [24] framework, allowing a modular and robust package-driven design for complex production-grade Shiny applications. Each main feature or analysis step is programmed as a shiny module, making use of the appropriate server-side functions. In the case of local use, those functions are exported by the package so they can be called from any R script to be part of an automated pipeline or more user-specific analyses. We also provide a Dockerfile [25] and instructions so that interested users can deploy DIANE to their own team servers. Figure 2 presents the application workflow and main possibilities. The analysis steps in DIANE are shown in a sequential order, from data import, pre-preprocessing and exploration, to more advanced studies such as co-expression or GRN inference.

Fig. 2 DIANE's workflow. The main steps of the pipeline available in the application -data import, normalization, exploration, differential expression analyses, clustering, network inference- alongside with some chosen visual outputs

Data upload

Expression file and design

To benefit from the vast majority of DIANE's features, the only required input is an expression matrix, giving the raw expression levels of genes for each biological replicate across experimental samples. It is assumed that this expression matrix file originates from a standard bioinformatics pipeline applied to the raw RNA-Seq fastq files. This typically consists in quality control followed by reads mapping to the reference genome, and quantification of the aligned reads on loci of interest.

Organism and gene annotation

Several model organisms are included in DIANE to allow for a fast and effortless annotation and pathway analysis. For now, automatically recognized model organisms are *Arabidopsis thaliana*, *Homo sapiens*, *Mus musculus*, *Drosophilia melanogaster*, *Ceanorhabditis elegans*, and *Escherichia coli*. DIANE takes advantage of the unified annotation data for those organisms offered by the corresponding Bioconductor organisms database packages [26–31]. Other plant species are annotated such as white lupin, and users can easily upload their custom files to describe any other organism whenever it is needed or possible along the pipeline. Organism specific information needed can be common gene names and descriptions, gene - GO terms associations, or known transcriptional regulators.

Normalization and low count genes removal

DIANE proposes several strategies of normalization to account for uneven sequencing depth between samples. One step normalization can be performed using either the Trimmed Mean of M values method (TMM) [32] or the median of ratios strategy from DESeq2 [33]. The TCC package [34] also allows to perform a prior DEA to remove

potential differentially expressed genes (DEG), and then compute less biased normalization factors using one of the previous methods. DIANE also includes a user-defined threshold for low-abundance genes, which may reduce the sensitivity of DEG detection in subsequent analyses [35]. The effect of normalization and filtering threshold on the count distributions can be interactively observed and adjusted.

Exploratory analysis of RNA-seq data

PCA - MDS

Dimensionality reduction techniques are frequently employed on normalized expression data to explore how experimental factors drive gene expression, and to estimate replicate homogeneity. In particular, the Multi-Dimensional Scaling (MDS) plot takes samples in a high dimensional space, and represents them as close in a two-dimensional projection plane [36] depending on their similarity. Principal Component Analysis (PCA) is also a powerful examination of expression data. Through linear algebra, new variables are built as a linear combination of the initial samples, that condense and summarize gene expression variation. By studying the contribution of the samples to each of these new variables, the experimenter can assess the impact of the experimental conditions on gene expression. DIANE offers those two features on expression data, where each gene is divided by its mean expression to remove the bias of baseline expression intensity.

As presented in Fig. 3a, we applied PCA to the normalized transcriptomes after low gene counts removal. No normalization was applied in DIANE as raw data

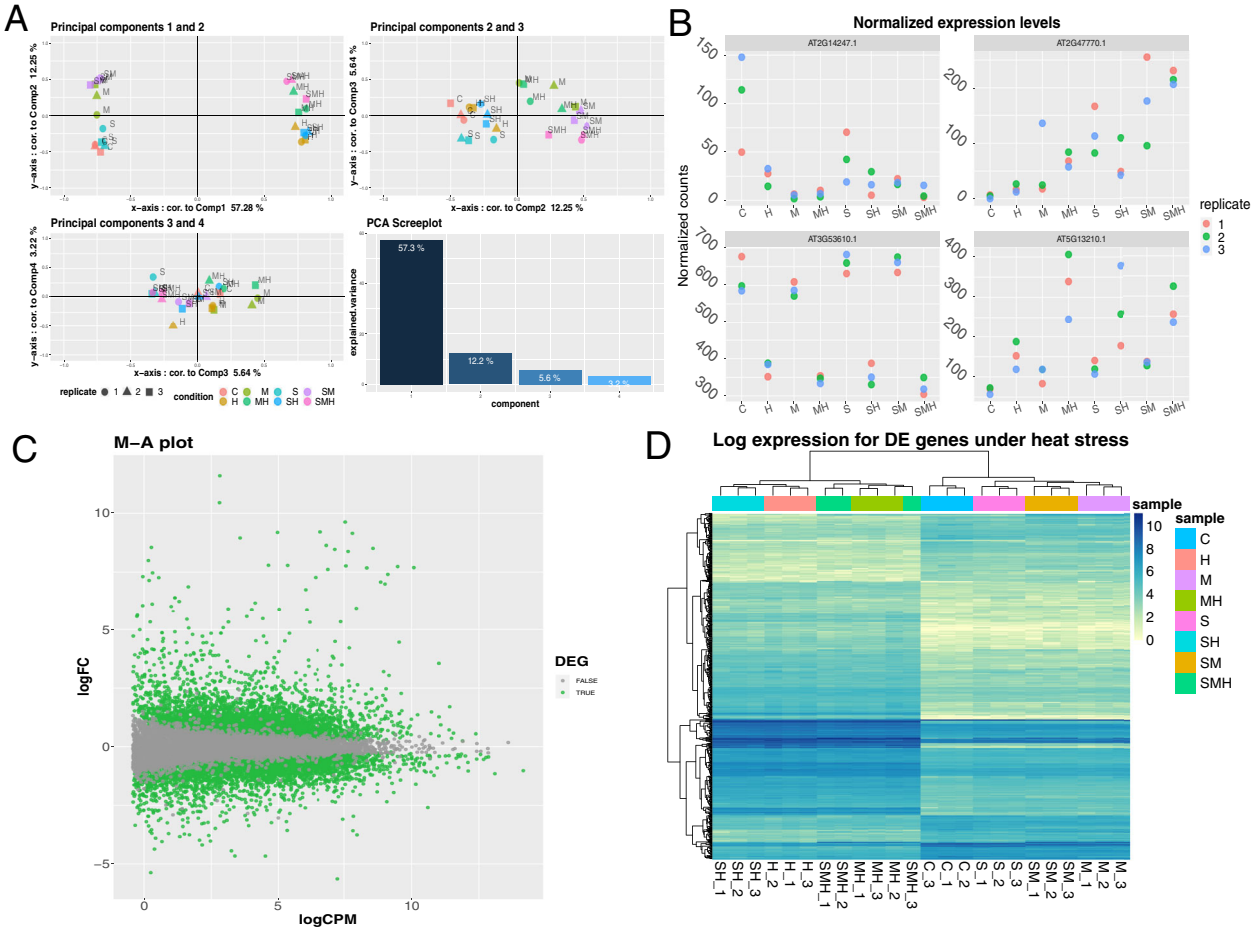


Fig. 3 Normalization and exploration of RNA-seq dataset with DIANE. **a** PCA analysis for the normalized expression table. The experimental conditions have for coordinates their contributions (correlations) to the first four principal components. The scree-plot shows, for each principal component, the part of global variability explained. **b** Example of normalized gene expression levels across all seven perturbations and control. **c** MA-plot for the DEG in response to a single heat stress. The x-axis is the average expression, and the y-axis is the LFC in expression between heat stress and control. DEG with FDR < 0.05 and an absolute LFC > 2 appear in green. **d** Log normalized expression heatmap for the DEG under heat across all perturbations and control

was presented as Tags Per Millions. We found consistent conclusions regarding how heat, salinity and osmotic stresses affect gene expression. The first principal component, clearly linked to high temperature, discriminates the experimental conditions based on heat stress while explaining 57% of the total gene expression variability. The second principal component, to which mannitol-perturbed conditions strongly contributes, accounts for 12% of gene expression variability. The effect of salinity is more subtle and can be discerned in the third principal component.

Normalized gene expression profiles

The "expression levels" tab of the application is a simple exploratory visualization, that allows the user to observe the normalized expression levels of a several genes of interest, among the experimental conditions of its choice. Each replicate is marked as different shapes. Besides rapidly showing the behavior a desired gene, it can provide valuable insights about a replicate being notably different from the others.

Using this feature of DIANE, we represented in Fig. 3b four genes showing different behaviors in response to the combination of stresses, and illustrating the variation that can be found among biological replicates.

Differential expression analysis

DEA in DIANE is carried out through the EdgeR framework [37], which relies on Negative Binomial Modelling. After gene dispersions are estimated, Generalized Linear Models are fitted to explain the log average gene expressions as a linear combination of experimental conditions. The user can then set the desired contrasts to perform statistical tests comparing experimental conditions. The adjusted p-value (FDR) threshold and the minimal absolute Log Fold Change (LFC) can both be adjusted on the fly. A data table of DEG and their description is generated, along with descriptive graphics such as MA-plot, volcano plot, and interactive heat-map. The result DEG are stored to be used as input genes for downstream studies, such as GO enrichment analysis, clustering or GRN inference.

Figure 3c and d represent DEG under heat perturbation. Selection criteria were adjusted p-values greater than 0.05, and an absolute log-fold-change over 2. The 561 up-regulated genes and 175 down-regulated genes are indicated in green in the MA-plot, and correspond to the rows of the heatmap. The high values of LFC for those genes, along with their expression pattern in the heatmap across all conditions confirm the strong impact of heat stress on the plants transcriptome.

In the case where several DEA were performed, it might be useful to compare the resulting lists of DEG. DIANE can perform gene lists intersection, and provide visualizations through Venn diagrams, as well as the possibility to download the list of the intersection. This feature is available for all genes, or specifically for up or down regulated genes.

GO enrichment analysis

Among a list of DEG, it is of great interest to look for enriched biological processes, molecular functions, of cellular components. This functionality is brought to DIANE by the clusterProfiler R package [38], that employs Fischer-exact tests on hypergeometric distribution to determine which GO terms are significantly more represented. Results can be obtained as a downloadable data table, a dotplot of enriched GO terms with associated gene counts and p-values, or as en enrichment map linking co-occurring GO terms.

Gene clustering

Method

In order to identify co-expressed genes among a list of DEGs, DIANE enables gene expression profiles clustering using the statistical framework for inferring mixture models through an Expectation-Maximisation (EM) algorithm introduced by [9, 10]. We chose to use the approach implemented in the Bioconductor Coseq package [11]. Coseq makes it possible to apply transformation to expression values prior to fitting either Gaussian or Poisson multivariate distributions to gene clusters. A penalized model selection criterion is then used to determine the best number of clusters in the data. With DIANE, users simply have to select which DEG should be clustered among previously realized DEA, the experimental conditions to use for clustering, as well as the range of number of clusters to test.

Exploring the clusters

Once clustering was performed, a new tab enables a detailed exploration of the created clusters. It includes interactive profiles visualization, downloadable gene data table, GO enrichment analysis. In addition, if the experimental design file was uploaded, Poisson generalized linear models are fitted to the chosen cluster in order to characterize the effect of each factor on gene expression.

To validate and extend the work done around our demonstration dataset, we performed clustering analysis similarly to what was done in the original paper [21]. We considered all genes from the seven DEA computed between control and perturbation treatments, with a 0.05 FDR threshold and an absolute LFC above 2.

Figure 4 presents the clusters of interest as given by the Poisson Mixtures estimation. They provide a gene partitioning representative of all behaviors in the dataset. In particular, we found that the 3 biggest clusters (2, 3, 6) were composed of heat responsive genes. Among those clusters, statistically enriched GO terms are in majority

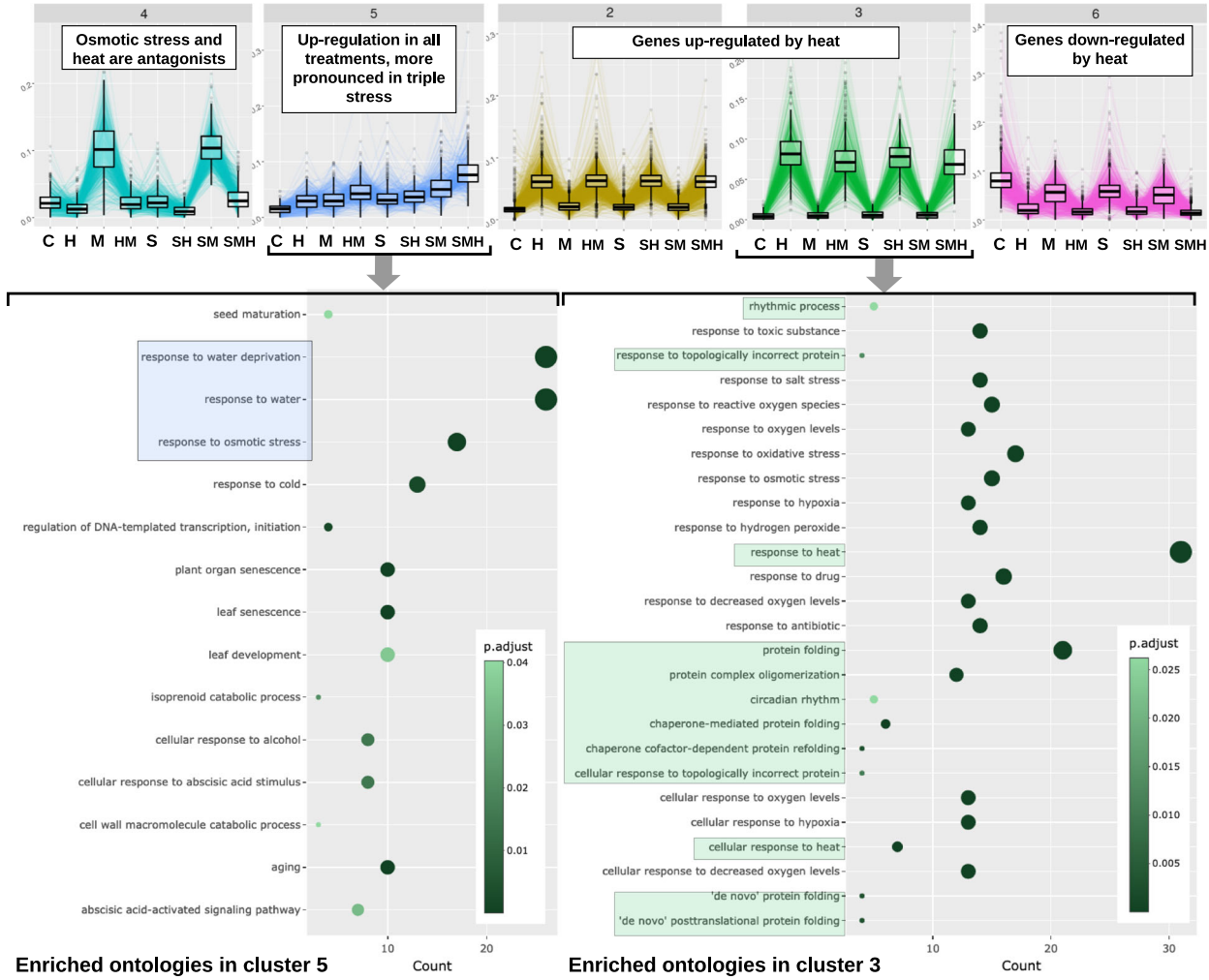


Fig. 4 Clustering of combinatorial RNA-seq data with DIANE. Clusters of interest generated by Coseq in DIANE. Gene expression profiles are defined as the normalized expression divided by the mean normalized expression across all conditions. Graphical results of ontology enrichment analysis are presented for clusters 3 and 5. Highlighted ontologies are relevant categories in line with previously published findings [21]. Ontology enrichment plots show detected GO terms (under 0.05 in Fischer's exact tests), color-coded by their adjusted *p*-value, and shifted in the x-axis depending on the number of genes matching this ontology

linked to heat and protein conformation. Indeed, proteins misfolding and degradation are direct consequences of high temperatures, thus requiring rapid expression reprogramming to ensure viable protein folding in topology control [39]. Two enriched ontologies involved in rhythmic and circadian processes also support evidence for disrupted biological clock. Second, the cluster 5 brings together genes up-regulated in all stress treatments, with the highest induction being observed in the combination of the three perturbations. Those genes, also noted in [21] to exhibit a synergistic response to mannitol and salt, contain three ontologies related to osmotic stress and water deprivation. Lastly, cluster 4 corroborates the existence of genes characterized by opposite reactions to osmotic stress and heat. They are specifically induced in all mannitol perturbations, except under high temperature, where they are strongly repressed.

Gene regulatory network inference

GRN inference is a major contribution of DIANE compared to similar existing applications, the latter offering either no possibility for such task, or either limited ones, as described in the "Background" section.

Estimating regulatory weights

GRN inference aims to abstract transcriptional dependencies between genes based on the observation of their resulting expression patterns. Each gene is represented by

a node in the network. The aim is to recover a weight associated with each edge (i.e. pair of nodes). This is a complex retro-engineering process, challenged by the Curse of dimensionality. Many methods are available, and can be divided into two main categories : statistical and data-driven approaches [13]. Statistical strategies rely on assumptions regarding the data distribution, whose parameters are estimated by maximum-likelihood techniques, often in the case of Bayesian [40] or Lasso inference [17, 41]. However, the underlying modelling assumptions may be inaccurate or difficult to verify in practice. In the second category, the objective is to quantify interaction strengths between pairs of nodes directly from the data. This is typically achieved by using similarity measures such as correlation [12], information theory metrics [42, 43], or feature importances extracted from regression contexts [16]. This second category is less restrictive in terms of hypothesis. However, once the inference is performed, the problem of defining a threshold above which an interaction will be part of the network is far from easy.

There is a large variety of tools available for the task of network inference. Many of them have been benchmarked against one another at the occasion of the DREAM challenges [44, 45]. Those challenges aim at comparing state of the art network inference methods on both simulated and validated biological data. They provide performance metrics for 27 methods based on regression techniques, mutual information metrics, correlation or Bayesian framework among other methods. The performance metrics gathered by DREAM5 [45] (i.e Area Under Precision and Recall curves or overall scores), as well as more recent efforts to compare new methods on those gold standards (i.e F-meausres, ROC curves) are useful resources to help making a choice. For example, existing methods to learn GRN structures are WGCNA [12], ARACNE, CLR, TIGRESS, GENIE3 (see [45] for an exhaustive and referenced list of methods), or also SORDER [46] or CMI2NI [47].

In DIANE, the package chosen for GRN reconstruction is GENIE3 [16], a machine learning procedure that was among the best performers of the DREAM challenges. GENIE3 uses Random Forests [20] which is a machine learning method based on the inference of a collection of regression trees. It has the advantage of being a non-parametric procedure, requiring very few modelling or biological priors, while being able to capture interactions and high order combinatorics between regulators. After having defined a set of regulators among the genes under study, the regression framework allows to infer oriented edges from regulators to targets. With GENIE3, for each target gene, a Random Forest determines the predictive power of each regulator on the target gene expression. The regulatory interactions can then be thresholded according to their importance, so that the strongest links are kept to build a sparse final network. However, choosing such a threshold is not trivial, left as an open question by GENIE3's authors and ever since.

Selecting meaningful regulatory weights

Proposed approach To avoid the unsatisfying hard-thresholding solution, some researchers make use of TF binding experiments, TF-perturbation assays, or literature data to select a threshold influence measure maximizing prediction precision [48–50]. Network backboning [51, 52] and BRANE Cut [53] are mathematical frameworks that try to extract an informative structure from weighted fully connected networks, but they rely on mathematical modelling and assumptions that we suppose might be too strong or not valid in the precise case of gene regulatory network topology. Feeling the lack of an appropriate model-agnostic strategy with no need for external data, we conceived a method that provides a statistical testing framework for weighted regulator-gene pairs. The main steps of the method, as schematized in Fig. 5, are:

Inference of the importance values for all regulator-target gene pairs using Random Forests according to GENIE3's strategy [16] on a chosen list of DEG as input. Transcriptional regulators with a very high value of non linear correlation (typically 0.9 or 0.95) can lead to spurious or missed connections in the final network, and cause robustness issues during the regression procedure. DIANE allows to group them together and to consider them as unique genes.

Selection of the strongest inferred regulatory influences. As biological networks are known for their pronounced sparsity [54–56], testing all possible regulator-target pairs would be of very little interest, as well as a waste of computation time. We therefore create a first graph, topologically consistent with biological network density standards, which will be further refined by statistical tests.

Empirical *p*-values are computed for the selected regulatory weights. To assess weather the importance value of a pair is significant or not, the rfPermute package [57] fits Random Forests and repeatedly shuffles the target gene expression profile so that the null distribution of each regulator influence is estimated. Hence, the empirical *p*-value of a regulator-gene pair is given by the extremeness of its importance as compared to the estimated null distribution. For a faster and more exploratory-oriented network inference, it is possible to skip edges testing (this step and the following).

FDR correction for multiple testing [58] is applied to the *p*-values, and only the edges above an FDR threshold are kept to form the final network. After edges statistical testing, graphics that show the *p*-values distribution and the final number of edges depending on the FDR choice

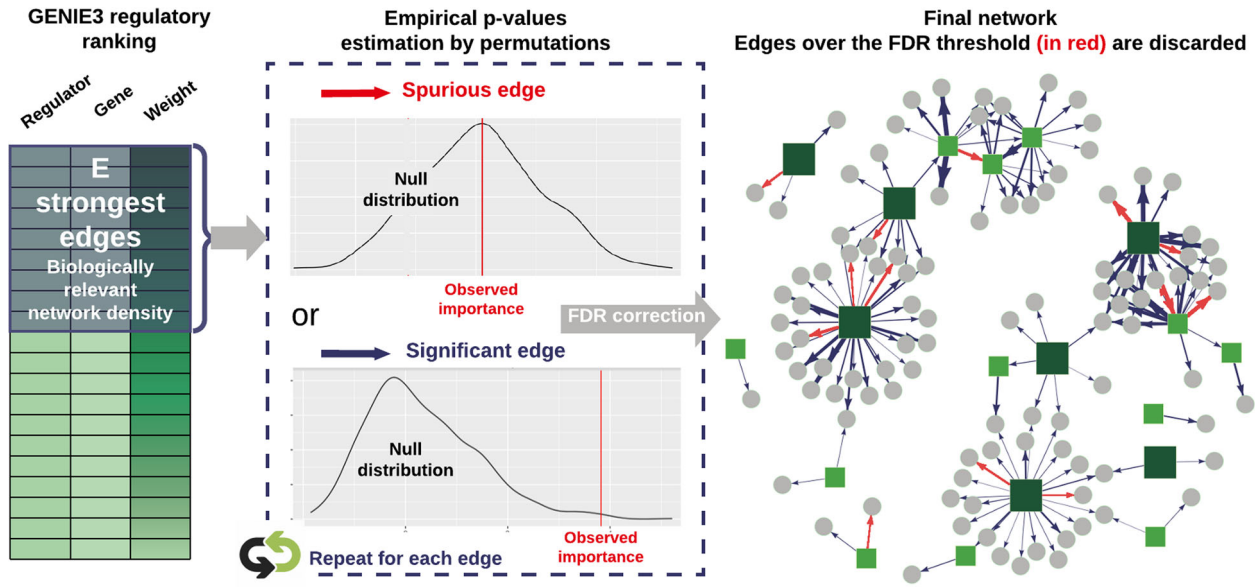


Fig. 5 Statistical testing procedure for edges selection among weighted regulatory links. A first network is built from the influence measures resulting from network inference, by choosing an appropriate network connectivity density. The statistical significance of all its edges is then assessed by empirical tests based on permutation for Random Forest importance metrics

are displayed, providing the user with additional decision guidance.

See Additional file 1 for more details on the statistical procedure and implementation. Thanks to this procedure, the main user-defined parameters are the network density prior to statistical tests, and the FDR cut-off. Together, they bring much more biological meaning and decision help than an arbitrary importance threshold.

Benchmark of the proposed approach We benchmark this novel procedure designed to keep the most significant interactions from a complete GRN. As GENIE3's performance was already assessed in several comparative studies, we focus here only on the edges testing strategy, that we compare to a more naive approach, hard thresholding. To do so, we applied our edges selection strategy to GENIE3 edges ranking on two different datasets, for which robust regulator-gene validation information is available.

The first expression dataset is the RNA-Seq experiment on *Arabidopsis thaliana* we present in this article. We inferred a GRN of heat responsive genes in all experimental conditions (1497 genes from C versus H DEA, LFC $\geq$ 1.5, FDR $\leq$ 0.05, containing 118 regulators). To validate the inferred connections, we made use of connecTF [59], a recent database containing regulatory interactions in Arabidopsis thaliana obtained from in vitro and in vivo binding experiments, as well as in planta regulation experiments. We specifically chose to use the interactions in connecTF obtained from CHIP-Seq and TARGET experiments that represent the most robust data in order to validate connections.

The second dataset is an experiment on *Escherichia coli*, generated by the authors of the "Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles" [60]. We restricted ourselves to a subset of this compendium of experimental conditions corresponding to a single combinatorial experiment. In the latter, bacteria were exposed to a control treatment or to norfloaxacin for different amounts of time, for a total of 24 experimental conditions. The 4345 genes of the organism provided in the dataset, containing 154 transcription factors, are used for GRN inference followed by edges testing. In order to validate the connections of the networks generated in DIANE, we used RegulonDB [61], a database of regulatory interactions built from classic molecular biology experiments and more recently high throughput genomics such as CHIP-Seq and gSELEX.

For each organism, we compared the validity of network predictions between two strategies. The first one corresponds to a network obtained by applying a hard threshold to GENIE3's weighted regulatory associations, to achieve a desired network connectivity density. The second strategy corresponds to that same network, but after removing the edges deemed spurious by our empirical testing procedure for edges selection. By doing so, we aim at determining weather refining edges with our testing procedure leads to networks of higher quality.

The performance metric we chose to assess our method's performance is the precision. It is computed as the fraction of edges in the final network that are present in the set of validated interactions, among those for which the regulator possesses validation information in the gold standard (for example, not all regulators were studied in CHIP-Seq nor TARGET experiments, thus are not present in the validated pairs from connecTF).

To provide some parameter exploration, we compare the two strategies for two different initial connectivity densities, and three FDR thresholds to remove spurious interactions. For all the following benchmarks, we used Random Forests made of 1000 trees, and grouped regulators correlated over 90%, as discussed in the previous paragraph "Proposed approach". In order to evaluate robustness while giving an overview of the variability inherent to Random Forest inference and statistical testing by permutations, we launched the two strategies 20 times for each set of parameters and performed non parametric tests for group mean comparisons.

The results are gathered in Fig. 6a and b. They demonstrate that a significant increase of precision can be achieved on both datasets when choosing stringent adjusted *p*-values for edges removal, independently of prior density. This finding supports that *p*-values obtained from permutations on Random Forest importance metrics can allow more confidence in the inferred edges than hard thresholding GENIE3's fully connected network. Figure 6a and b also illustrate the order of magnitude of the number of connections removed by the testing strategy.

After using our empirical testing procedure for edges removal, we stored the number of remaining edges. We then applied hard-thresholding to GENIE3's ranking in order to create networks containing those same number edges. We observed that the precision of such networks was not as high as with our empirical testing

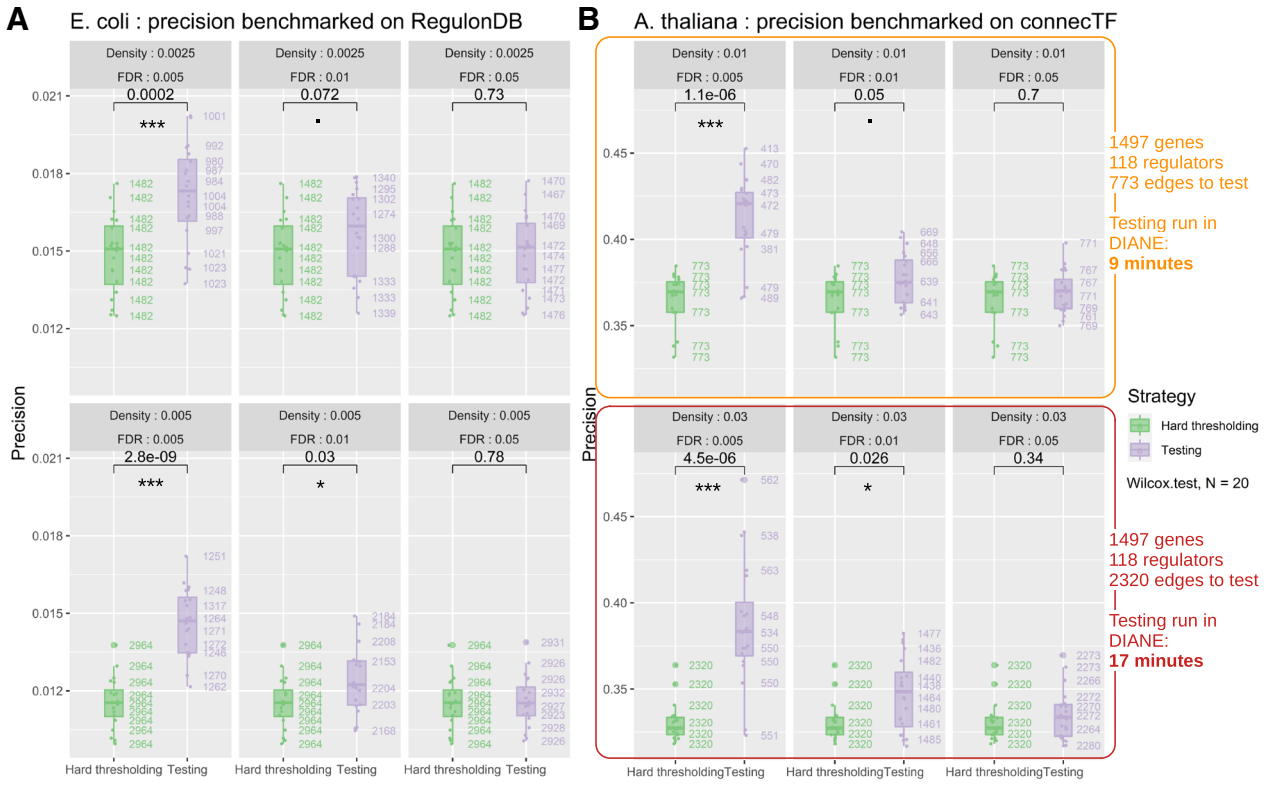


Fig. 6 Benchmark of the proposed testing method on the *E. coli* and *A. thaliana* datasets. Boxplots compare the distributions of precision between hard-thresholding (green) and hard-thresholding followed by the removal of non significant edges as predicted by the testing procedure (purple). The 20 replicates for each configuration provide an estimation of the precision dispersion caused by randomness in GENIE3 and testing by permutations. For each organism, we investigate two appropriate connectivity densities, and three adjusted *p*-value thresholds (FDR). On the right of the boxplots, the number of edges kept in the final network are displayed. *P*-values significance of non parametric mean comparisons between the strategies are encoded as follows : $0 \leq^{***} < 0.001 \leq^{**} < 0.01 \leq^{*} < 0.05 \leq . < 0.1$. The results demonstrate that the proposed testing strategy offers a robust gain in precision when using a stringent adjusted *p*-value threshold for edges removal. **a** Results for the GRN inferred on *E. coli* genes, validated on the regulonDB database. **b** Results for the GRN inferred on *A. thaliana* heat-responsive genes, validated on the connecTF database. Additional metrics about the number of genes, interactions to test, and computation time on DIANE's interface are shown

procedure. This reveals that our adjusted *p*-values bring more information than GENIE3's ranking only, even with a hard-thresholding resulting in the same number of final interactions.

Figure 6b shows computation times required to perform statistical testing on *A. thaliana* dataset, as permitted by DIANE's online interface. DIANE's online version is hosted on a Debian 9.13 server with a 256Go RAM, and 2 Intel(R) Xeon(R) Gold 6130 2.10GHz CPUs. The parallel computing for online use allows up to 16 CPU cores (computation time reported in Fig. 6b uses 16 cores).

Altogether, this benchmarking analysis demonstrates an added-value in terms of network precision when edges selection is performed on the basis of *p*-values rather than by hard thresholding, for a limited time of computation.

Interactive network analysis and community discovery

The last tab of the application is dedicated to network manipulation and exploration. An interactive view of the network is proposed, showing connections between regulatory genes and their predicted targets. By clicking one of the genes, its inward and outward interactions are shown, as well as its annotation and expression profile across samples.

Network-related statistics are automatically generated, delivering topological insights on genes behaviors and network structure. For instance, in and out degree distribution are displayed, and genes can be ranked based on their number of connections. This ranking might then be used for further identification of hub genes and candidate key regulators in the response of interest. In addition, DIANE extracts gene modules, making use of the Louvain algorithm [62]. The experimenter is then free to visualize the results in the network as color-coded communities, while exploring module-specific expression profiles and GO enrichment analyses. At last, it is possible to download edges and node information as csv dataframes, to be further investigated or opened in popular network visualization tools such as Cytoscape.

We used the GRN features of DIANE in order to infer a GRN of the response to heat under osmotic stress, environmental conditions that plants are supposed to face more frequently under climate change circumstances. The input list of genes is obtained in DIANE, by calculating DEG between simple osmotic stress and the double heat-osmotic perturbation (M versus HM, FDR $<$ 0.01, LFC $>$ 2). 640 DEG are detected, among which 363 are up-regulated, 277 are down-regulated, and 45 are transcriptional regulators. Regulators with Spearman correlations over 90% in all available experimental conditions were grouped before network inference, so that a total of 27 regulators are used as predictive variables during inference. For GRN reconstruction, we used Random Forests composed of 4000 trees. A prior network density of 0.03 was defined to select the strongest edges for permutation testing, and edges under a 0.01 FDR were kept in the final network. This network, presented in Fig. 7a, is composed of 289 nodes and 438 edges.

The M versus MH GRN provided by DIANE revealed two interesting groups of regulators, acting as central nodes in their topological modules, and being connected to a large number of target genes.

The most connected regulator of the network is composed by the WRKY47-WRKY8 grouping. Along with other top-ranked WRKY transcription factors (WRKY30, WRKY6, WRKY55), they belong to the topological community of genes that exhibit antagonistic behavior between heat and osmotic stress. The expression values of WRKY8 and WRKY47 in the experiment are presented in Fig. 7b. As already pointed out by our clustering analysis in Fig. 4, those genes undergo a strong induction after mannitol treatment while being repressed by all high temperature conditions. This behavior can also be observed in the intra-module expression profiles in Fig. 7c. Such a module is of high biological interest, as these opposite interactions between drought and high temperature might explain the increased damages observed in the combination of those perturbations [21], and help to understand how heat can suppress the adaptive response of plants to water deficit. Given that WRKY47 and WRKY8 act as a hub in the inferred network, they would be a relevant choice of candidates for experimental pathway validation. Interestingly, WRKY47 has already been identified in rice as a positive regulator of the response to drought [63], strongly reinforcing the validity of the candidate genes from GRN inference in DIANE.

The second most connected node is formed by the regulators TRFL3-AT5G57150-PRR3-BBX8-PIL2-BT1-DREB2B-FRG5-ASY1-ARR15. Those genes, sharing highly correlated profiles across the 24 experimental samples, respond to heat in a clear manner, as well as the other genes inside their community as shown in Fig. 7c. It is worthy to note that PIL2 is a member of a transcription factor family known to be involved in the response to temperature [64] and that DREB2B is a regulator already characterized to act at the interaction between drought and heat stress [65]. The other mentioned regulators offer thus promising leads to be further explored. Three members of the Heat Stress Transcription Factor family (HSFA2 grouped with HSFB2B, and HSFA3) are also found within the genes of the module.

Inside each module, both correlated and anti-correlated expression patterns coexist, which can indicate negative regulation between their gene members. Such opposite variations are captured by the Random Forest algorithm, and allow to go beyond co-expression analysis provided by a clustering approach alone.

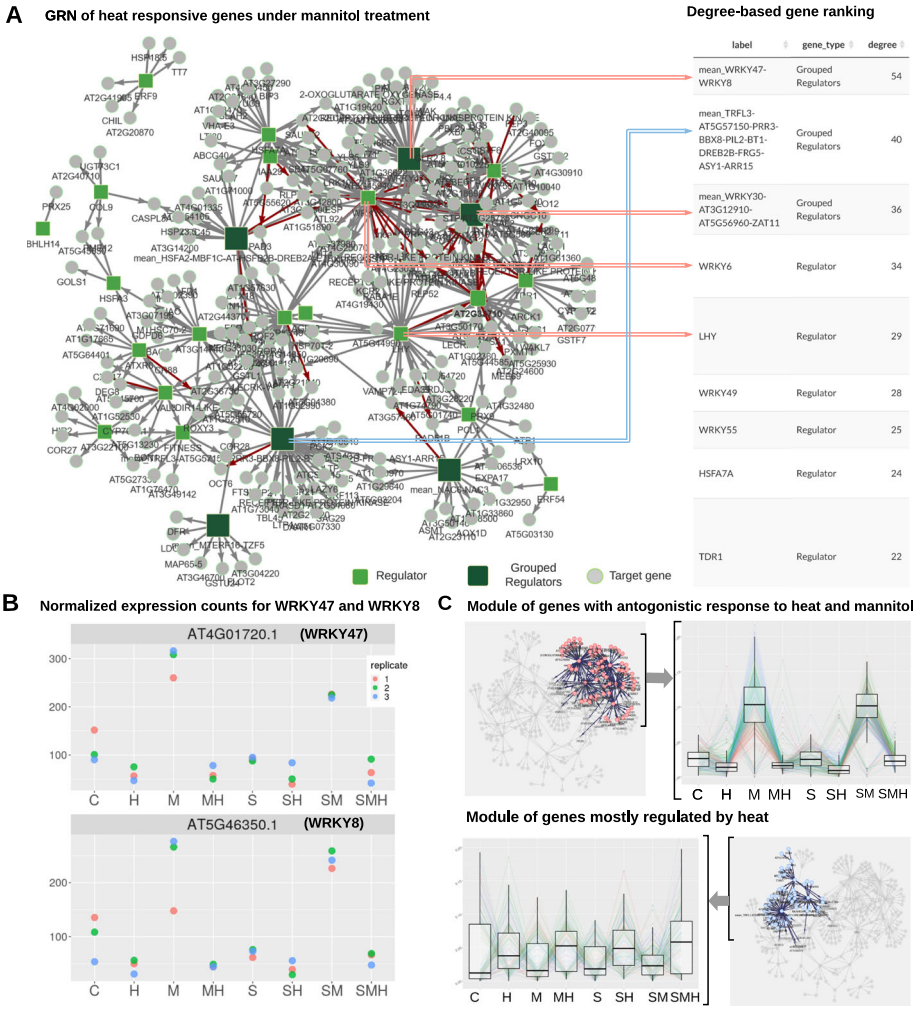


Fig. 7 Network inference and exploration with DIANE. **a** GRN on M versus MH DEG using DIANE's method for GNR inference, and the corresponding degree-based ranking of the nodes. The 11 most connected nodes are presented by order of importance. The regulators mentioned in the network analysis are pointed out by an arrow, the color of the arrow depending on the topological cluster. **b** Expression profiles for WRKY47 and WRKY8, representing the most connected node of the network. **c** Topological modules containing the two most connected groups of regulators are highlighted, juxtaposed to their genes expression profiles

Research reproducibility

For each step of the pipeline, automatically generated reports can be downloaded, rendered on the fly in RMarkdown. They store the users settings, chosen strategies, and display previews of the results. In that way, analysis can be re-run, shared across users, and their settings can be backed-up. The chosen format for those reports is HTML, as it keeps a possibility to interact with data tables, or even manipulate network objects outside of the application. Additional file 2 is an example of report as generated for the network inference described in previous section. Besides, a seed can be set as a global setting of the application, to ensure reproducible runs of the pipeline steps making use of randomness.

Accessibility

DIANE is a tool designed to be as accessible as possible. However, it can be challenging for users with little programming and command line experience to process raw RNA-Seq data into the expression matrix needed in DIANE. Services such as quality control, read mapping and quantification require to handle large files transfers and intensive computations, which are much less easily set up on online applications. However, local programs such as the Tuxedo suite [66], RMTA [67] or GenePattern [68] represent well documented and adequate solutions to most users in order to produce the expression matrices required in DIANE.

Conclusions

To summarise this work, we presented an online graphical user interface to easily conduct in-depth analyses on gene expression data from multi-factorial experiments,

including gene expression profile clustering and GRN inference. It can be downloaded and installed seamlessly as any R package to run the pipeline locally or from R scripts. Given that all other graphical interface tools found in the literature are (i) more oriented toward co-expression rather than regulation and (ii) do not provide recent advanced methodological frameworks for pathway reconstruction, our application positions itself as a tool of first choice to explore regulatory mechanisms.

The demonstration of DIANE on its companion dataset allowed to better understand the effect of combined heat, osmotic and salinity perturbations on Arabidopsis thaliana, consistently with the original analysis [21]. Similar patterns in gene behaviors were highlighted, such as the predominant influence of heat, and its aggravating effect when combined to dehydration. Moreover, DIANE provided new leads through its network inference features : key genes involved in the response to high temperature under drought were pointed out to be promising candidate regulators for improving crops resistance to arid conditions and climate change.

In terms of computational cost, the final step of DIANE's pipeline, i.e. the statistical testing of TF-target edges, could be improved. The R implementations of Random forests and permutations in rfPermute are currently being used, but a C++ version could be envisioned to shorten the method's execution time. Besides, the inference method itself could be subject to improvement in the future. First, combining the results of several inference methods has proven to be as a robust and powerful approach on validated datasets [45, 52]. Second, our strategy is particularly well-suited for multi-factorial and perturbation designs, but is not optimal for time series RNA-Seq. Other inference methods specific to time series RNA-Seq data [69] could be available in DIANE, to bring closer to causality in the inferred transcriptionnal interactions. Lastly, it would be valuable to add further functional features in DIANE, notably in order to integrate external information, such as interaction databases, or data from TF binding or chromatin accessibility experiments.

Availability and requirements

Project name: DIANE
Project home pages: https://oceanecsn.github.io/DIANE https://github.com/OceaneCsn/DIANE
Operating system(s): Platform independent
Programming language: R
Other requirements: Web use : none. Local use: R >4.0.1
License: GNU GPL
Any restrictions to use by non-academics: none

Abbreviations

H: High temperature perturbation M: Mannitol perturbation S: Salinity perturbation SM: Salinity and Mannitol perturbations SH: Salinity and High temperature perturbations MH: Mannitol and High temperature perturbations SMH: Salinity, Mannitol and High temperature perturbations DEA: Differential Expression Analysis DEG: Differentially Expressed Genes DIANE: Dashboard for the Inference and Analysis of Networks from Expression data FDR: False Discovery Rate GENIE3: GEne Network Inference with Ensemble of trees GO: Gene Ontology GRN: Gene Regulatory Network LFC: Log Fold Change NGS: Next-Generation Sequencing PCA: Principal Component Analysis RNA-Seq: Sequencing of RNA TF: Transcription Factors TMM: Trimmed Mean of M values WGCNA: WeiGhted Correlation Network Analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07659-2.

Additional file 1: Full description of the procedure of importance measures empirical testing. the files gives more details about the methodological choices for the procedure.

Additional file 2: Network inference report from the M versus MH GRN. Interactive report generated after network inference and edges testing in DIANE. Slight changes might be observed from the textual description of the network because of the stochasticity inherent to the Louvain, Random Forest, and permutations procedures.

Acknowledgements

We thank Alexandre Soriano, Cécile Fizames, Adrien Jarretier-Yuste for help, comments and suggestions during the development of this application. We thank Benjamin Péret for his support in the initial web deployment of DIANE.

Authors' contributions

SL, AM, OC defined the application concepts, searched scientific literature for appropriate methods, tools, biological findings, and redacted the article. SL, OC developed the empirical testing procedure on edges importance measures. AM, OC chose the demonstration dataset, used DIANE on it, and performed biological interpretations. OC carried out the programming and benchmarking of DIANE. All authors have read and approved the manuscript.

Funding

OC, SL and AM are supported by a 80 Prime fellowship from the National Center of Scientific Research (CNRS, France).

Availability of data and materials

The RNA-Seq experiment we included to DIANE for demonstration purposes corresponds to the GEO accession GSE146206 and can be found at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE146206.
The code and benchmark scripts of DIANE are available in the github repositories https://github.com/OceaneCsn/DIANE and https://github.com/OceaneCsn/Benchmarking_DIANE.
DIANE largely relies on the CRAN https://cran.r-project.org/ and Bioconductor https://bioconductor.org/ packages repositories.
The datasets queried to retrieve validated regulatory interactions are connecTF https://connectf.org/ and RegulonDB http://regulondb.ccg.unam.mx/.
The expression data used to infer regulatory networks on *Escherichia coli* were taken from the Many Microbe Microarrays Database at http://m3d.mssm.edu/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] BPMP, CNRS, INRAE, Institut Agro, Univ Montpellier, 34060 Montpellier, France. [2] IMAG, Univ. Montpellier, CNRS, Montpellier, France. [3] Université Paul-Valéry-Montpellier 3, Montpellier, France.

Received: 17 November 2020 Accepted: 28 April 2021
Published online: 26 May 2021

References

1. Kucukural A, Yukselen O, Ozata DM, Moore MJ, Garber M. DEBrowser: Interactive differential expression analysis and visualization tool for count data 06 Biological Sciences 0604 Genetics 08 Information and Computing Sciences 0806 Information Systems. BMC Genomics. 2019;20(1):6. https://doi.org/10.1186/s12864-018-5362-x.
2. Li Y, Andrade J. DEApp: An interactive web interface for differential expression analysis of next generation sequence data. Source Code Biol Med. 2017;12(1):10–3. https://doi.org/10.1186/s13029-017-0063-4.
3. Choi K, Ratner N. IGEAK: An interactive gene expression analysis kit for seamless workflow using the R/shiny platform. BMC Genomics. 2019;20(1):177. https://doi.org/10.1186/s12864-019-5548-x.
4. Harshbarger J, Kratz A, Carninci P. DEIVA: A web application for interactive visual analysis of differential gene expression profiles. BMC Genomics. 2017;18(1):47. https://doi.org/10.1186/s12864-016-3396-5.
5. Sundararajan Z, Knoll R, Hombach P, Becker M, Schultze JL, Ulas T. Shiny-Seq: advanced guided transcriptome analysis. BMC Res Notes. 2019;12(1):432. https://doi.org/10.1186/s13104-019-4471-1.
6. Monier B, McDermaid A, Wang C, Zhao J, Miller A, Fennell A, Ma Q. IRIS-EDA: An integrated RNA-seq interpretation system for gene expression data analysis. PLoS Comput Biol. 2019;15(2):. https://doi.org/10.1371/journal.pcbi.1006792.
7. Ge SX, Son EW, Yao R. iDEP: An integrated text application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics. 2018;19(1):1–24. https://doi.org/10.1186/s12859-018-2486-6.
8. Su W, Sun J, Shimizu K, Kadota K. TCC-GUI: A Shiny-based application for differential expression analysis of RNA-Seq count data. BMC Res Notes. 2019;12(1):133. https://doi.org/10.1186/s13104-019-4179-2.
9. Rau A, Celeux G, Martin-Magniette M-L, Maugis-Rabusseau C. Clustering high-throughput sequencing data with poisson mixture models. [Research Report] RR-7786, INRIA. 2011, p. 36. hal-01193758v2.
10. Rau A, Maugis-Rabusseau C, Martin-Magniette M-L, Celeux G. Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models. Bioinformatics. 2015;31(9):1420–7.
11. Rau A, Maugis-Rabusseau C. Transformation and model choice for RNA-seq co-expression analysis. Brief Bioinforma. 2018;19(3):425–36. https://doi.org/10.1093/bib/bbw128.
12. Langfelder P, Horvath S. Wgcna: an r package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559.
13. Sanguinetti G, Huynh-Thu VA. Gene regulatory networks. New York: Springer, Humana Press; 2019.
14. Zhang M, Li Q, Yu D, Yao B, Guo W, Xie Y, Xiao G. Geneck: a web server for gene network construction and visualization. BMC Bioinformatics. 2019;20(1):1–7.
15. Chen J, Zhang R, Dong X, Lin L, Zhu Y, He J, Christiani DC, Wei Y, Chen F. shinybn: an online application for interactive bayesian network inference and visualization. BMC Bioinformatics. 2019;20(1):711.
16. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE. 2010;5(9):12776. https://doi.org/10.1371/journal.pone.0012776.
17. Haury A-C, Mordelet F, Vera-Licona P, Vert J-P. Tigress: trustful inference of gene regulation using stability selection. BMC Syst Biology. 2012;6(1):145.
18. Chiquet J, Robin S, Mariadassou M. Variational inference for sparse network reconstruction from count data. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97. PMLR; 2019. p. 1162–71.
19. Mochida K, Koda S, Inoue K, Nishii R. Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. Front Plant Sci. 2018;9:1770.
20. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
21. Sewelam N, Brilhaus D, Bräutigam A, Alseekh S, Fernie AR, Maurino VG. Molecular plant responses to combined abiotic stresses put a spotlight on unknown and abundant genes. J Exp Bot. 2020.
22. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.
23. Chang W, Cheng J, Allaire J, Xie Y, McPherson J, et al. Shiny: web application framework for r. R package version 1(5). 2017.
24. Guyader V, Fay C, Rochette S, Girard C. Golem: A Framework for Robust Shiny Applications. 2020. R package version 0.2.1. https://CRAN.R-project.org/package=golem. Accessed 04 May 2021.
25. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014;2014(239):2.
26. Carlson M. org.At.tair.db: Genome Wide Annotation for Arabidopsis. 2020. R package version 3.11.4.
27. Carlson M. org.Ce.eg.db: Genome Wide Annotation for Worm. 2020. R package version 3.11.4.
28. Carlson M. org.Dm.eg.db: Genome Wide Annotation for Fly. 2020. R package version 3.11.4.
29. Carlson M. org.EcK12.eg.db: Genome Wide Annotation for E Coli Strain K12. 2020. R package version 3.11.4.
30. Carlson M. org.Hs.eg.db: Genome Wide Annotation for Human. 2020. R package version 3.11.4.
31. Carlson M. org.Mm.eg.db: Genome Wide Annotation for Mouse. 2020. R package version 3.11.4.
32. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):25. https://doi.org/10.1186/gb-2010-11-3-r25.
33. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
34. Sun J, Nishiyama T, Shimizu K, Kadota K. TCC: An R package for comparing tag count data with robust normalization strategies. BMC Bioinformatics. 2013;14(1):219. https://doi.org/10.1186/1471-2105-14-219.
35. Sha Y, Phan JH, Wang MD. Effect of low-expression gene filtering on detection of differentially expressed genes in rna-seq data. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). New York: IEEE; 2015. p. 6461–4.
36. Kruskal JB. Multidimensional Scaling, vol. 11. Thousands Oaks, California: Sage; 1978.
37. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97. https://doi.org/10.1093/nar/gks042.
38. Yu G, Wang L-G, Han Y, He Q-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. Omics J Integr Biol. 2012;16(5):284–7.
39. Wang W, Vinocur B, Shoseyov O, Altman A. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends Plant Sci. 2004;9(5):244–52.
40. Ko Y, Kim J, Rodriguez-Zas SL. Markov chain monte carlo simulation of a bayesian mixture model for gene network inference. Genes Genomics. 2019;41(5):547–55.
41. Omranian N, Eloundou-Mbebi JM, Mueller-Roeber B, Nikoloski Z. Gene regulatory network inference using fused lasso on multiple data sets. Sci Rep. 2016;6:20533.
42. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):8.
43. Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A. Reverse engineering cellular networks. Nat Protoc. 2006;1(2):662.
44. Greenfield A, Madar A, Ostrer H, Bonneau R. DREAM4: Combining genetic and dynamic information to identify biological networks and Dynamical Models. PLoS ONE. 2010;5(10):. https://doi.org/10.1371/journal.pone.0013397.
45. Marbach D, Costello JC, Küffner R, Vega N, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G, Performed GSM. Wisdom of crowds for robust gene network inference the DREAM5 Consortium HHS Public Access. Nat Methods. 2016;9(8):796–804. https://doi.org/10.1038/nmeth.2016.
46. Aghdam R, Ganjali M, Zhang X, Eslahchi C. Cn: a consensus algorithm for inferring gene regulatory networks using the sorder algorithm and conditional mutual information test. Mol BioSyst. 2015;11(3):942–9.

47. Zhang X, Zhao J, Hao J-K, Zhao X-M, Chen L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic Acids Res. 2015;43(5):31–31.
48. Anwar M, Tambalo M, Ranganathan R, Grocott T, Streit A. A gene network regulated by FGF signalling during ear development. Sci Rep. 2017;7(1):. https://doi.org/10.1038/s41598-017-05472-0.
49. Shibata M, Breuer C, Kawamura A, Clark NM, Rymen B, Braidwood L, Morohashi K, Busch W, Benfey PN, Sozzani R, Sugimoto K. GTL1 and DF1 regulate root hair growth through transcriptional repression of ROOT HAIR DEFECTIVE 6-LIKE 4 in Arabidopsis. Development (Cambridge). 2018;145(3):. https://doi.org/10.1242/dev.159707.
50. Brooks MD, Cirrone J, Pasquino AV, Alvarez JM, Swift J, Mittal S, Juang C-L, Varala K, Gutiérrez RA, Krouk G, et al. Network walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. Nat Commun. 2019;10(1):1–13.
51. Coscia M, Neffke FMH. Network backboning with noisy data; 2017. p. 425–436. https://doi.org/10.1109/ICDE.2017.100.
52. Schiffthaler B, Serrano A, Delhomme N, Street NR. Seidr: A toolkit for calculation of crowd networks: Cold Spring Harbor Laboratory; 2018, p. 250696. https://doi.org/10.1101/250696.
53. Pirayre A, Couprie C, Bidard F, Duval L, Pesquet JC. BRANE Cut: Biologically-related a priori network enhancement with graph cuts for gene regulatory network inference. BMC Bioinformatics. 2015;16(1):368. https://doi.org/10.1186/s12859-015-0754-2.
54. Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA. A Guide to Conquer the Biological Network Era Using Graph Theory. Front Media S.A. 2020. https://doi.org/10.3389/fbioe.2020.00034.
55. Leclerc RD. Survival of the sparsest: Robust gene networks are parsimonious. Mol Syst Biol. 2008;4:. https://doi.org/10.1038/msb.2008.52.
56. Hayes W, Sun K, Pržulj N. Graphlet-based measures are suitable for biological network comparison. Bioinformatics. 2013;29(4):483–91. https://doi.org/10.1093/bioinformatics/bts729.
57. Archer E. rfPermute: Estimate Permutation *p*-values for Random Forest Importance Metrics. 2020. R package version 2.1.81. https://CRAN.R-project.org/package=rfPermute.
58. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol). 1995;57(1):289–300.
59. Brooks MD, Juang C-L, Katari MS, Alvarez JM, Pasquino A, Shih H-J, Huang J, Shanks C, Cirrone J, Coruzzi GM. Connectf: A platform to integrate transcription factor-gene interactions and validate regulatory networks. Plant Physiol. 2020;185(1):49–66. https://doi.org/10.1093/plphys/kiaa012. https://academic.oup.com/plphys/article-pdf/185/1/49/36389080/kiaa012.pdf.
60. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):8.
61. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, García-Sotelo JS, Alquicira-Hernández K, Muñiz-Rascado LJ, Peña-Loredo P, et al. Regulondb v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli k-12. Nucleic Acids Res. 2019;47(D1):212–20.
62. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):10008.
63. Raineri J, Wang S, Peleg Z, Blumwald E, Chan RL. The rice transcription factor oswrky47 is a positive regulator of the response to water deficit stress. Plant Molecular Biol. 2015;88(4-5):401–13.
64. Lin L, Liu X, Yin R. Pif3 integrates light and low temperature signaling. Trends Plant Sci. 2018;23(2):93–5.
65. Lata C, Prasad M. Role of drebs in regulation of abiotic stress responses in plants. J Exp Bot. 2011;62(14):4731–48.
66. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nat Protocol. 2012;7(3):562–78.
67. Peri S, Roberts S, Kreko IR, McHan LB, Naron A, Ram A, Murphy RL, Lyons E, Gregory BD, Devisetty UK, Nelson ADL. Read mapping and transcript assembly: A scalable and high-throughput workflow for the processing and analysis of ribonucleic acid sequencing data. Front Genet. 2020;10:1361. https://doi.org/10.3389/fgene.2019.01361.
68. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. Genepattern 2.0. Nat Genet. 2006;38(5):500–1.
69. Geurts P, et al. dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. Sci Rep. 2018;8(1):1–12.

Publisher's Note

Supplementary File 1

Selecting meaningful importance values from Random Forests inference

Océane Cassan, Sophie Lèbre, Antoine Martin

October 12, 2022

1 Statistical procedure

To assess weather an importance value is significant or not, the `rfPermute` package [1] fits Random Forests and repeatedly shuffles the target gene expression profile so that the null distribution of each regulator influence is estimated. Hence, the empirical p-value of a regulator-gene pair is given by the extremeness of its importance as compared to the estimated null distribution.

As biological networks are known for their pronounced sparsity [2, 3, 4], testing all possible regulator-target pairs would be of very little interest, as well as a waste of computation time. Besides, our preliminary analysis showed that corrections for multiple testing were made unreasonably conservative by the very large number of edges. We therefore propose to create a first graph, topologically consistent with biological network standards, which will be further refined by statistical testing.

More precisely, the steps of the method are :

1. **Inference of the importance values for all regulator-target gene pairs using GENIE3**. The importance metric returned by GENIE3's Random Forests is the total decrease in node impurities from splitting on the variable, averaged over all trees [5]. It requires the target gene expressions to be normalized to a unit variance, so that their regulatory importance measures can be compared without bias. In the GENIE3 framework, it was shown faster and equivalent to another importance metric, the prediction error on the out-of-bag permuted data. Although both can be used for this step, we recommend the use of the second one for consistency reasons regarding the third step.

2. **Selection of the number E of edges based on the inferred regulatory ranking**. The value of E is such as it gives a superior limit to the network density. The total number of possible edges in an oriented regulatory network being $E_{max} = N_{regulators}(N_{genes} - 1)$, and the density being defined as $d = \frac{E}{E_{max}}$, we deduce $E = dN_{regulators}(N_{genes} - 1)$.

Studies such as [4] on state of the art protein-protein interaction structure found that the typical values of density in biological networks lie approximately between 0.1 and 0.001, guiding the user's choice for this parameter.

3. **Empirical p-values are computed for the selected regulatory weights** with the `rfPermute` package. For each gene involved in the selected edges, Random Forests are fitted using its connected regulators as variables, as defined in the network resulting from the first step. The response variable is permuted $nShuffle$ times to build the null distributions. The empirical p-value for an edge is consequently the proportion of the null importance values above the observed importance. We propose a default value $nShuffle = 1000$, but it can be increased for more precise p-value estimations. The importance metric to use in the Random Forests for this step is the prediction error on Out-of-bag examples. Indeed, we observed (data not shown) that, unlike the node impurity measure, prediction error on OOB examples was robust to the reduced number of regulators caused by the selection of E edges only and to over-fitting as well. Moreover, it does not require any expression normalisation, as it is already dealt with within the metric definition.
4. **FDR adjustment** [6] for multiple testing is applied to the set of p-values.
5. Only the **edges above a certain FDR threshold** are kept to be part of the final network.

In brief, the main user-defined parameters are the estimated network density, and the FDR cut-off. Together, they bring much more biological meaning and decision help than an arbitrary importance threshold.

2 Implementation

For the implementation of this method for edges selection, the source code of GENIE3 was modified in order to use the R implementation for Random Forests and allow to change the importance metric. The testing procedure was implemented in a function that benefits from CPU multi-threading to reduce computation time, but it stays the more time consuming step. Graphics that show the p-values distribution and the final number of edges depending on the FDR choice are displayed, providing the user with additional decision guidance.

The method is embedded in DIANE, available either through its user interface, or via functions to run from R scripts, as detailed in the package vignette.

References

[1] Archer, E.: rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics. (2020). R package version 2.1.81. https://CRAN.R-project.org/package=rfPermute

[2] Koutrouli, M., Karatzas, E., Paez-Espino, D., Pavlopoulos, G.A.: A Guide to Conquer the Biological Network Era Using Graph Theory. Frontiers Media S.A. (2020). doi:10.3389/fbioe.2020.00034

[3] Leclerc, R.D.: Survival of the sparsest: Robust gene networks are parsimonious. Molecular Systems Biology **4** (2008). doi:10.1038/msb.2008.52

[4] Hayes, W., Sun, K., Pržulj, N.: Graphlet-based measures are suitable for biological network comparison. Bioinformatics **29**(4), 483–491 (2013). doi:10.1093/bioinformatics/bts729

[5] Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)

[6] Benjamini, Y., Hochberg, Y.: 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 (1995)

2.1.3 Afterword

More than a year after the publication of DIANE, its use in the research community can be debriefed. DIANE has been updated several times in order to add a few functional or ergonomic features. It was also updated in order to comply with the new versions of base R and of the packages it depends on, but also after the reports of minor bugs from several users. DIANE received external code contributions and was added to online portals from SouthGreen like the Rice Genome Hub or the Banana Genome Hub. It has also been used in teaching for plant biology students and in research formations.

An internal tool we developed allows to visualize the logs of the online use of DIANE. It discloses that since its publication in May 2021, it has been increasingly used. In octobre 2022, between 80 and 100 connections are reported weekly. Some features are also more commonly used than others: pre-processing steps and differential expression are almost always used, while tasks in the end of the pipeline like the statistical testing of inferred edges or GO enrichments in GRN modules are less frequent.

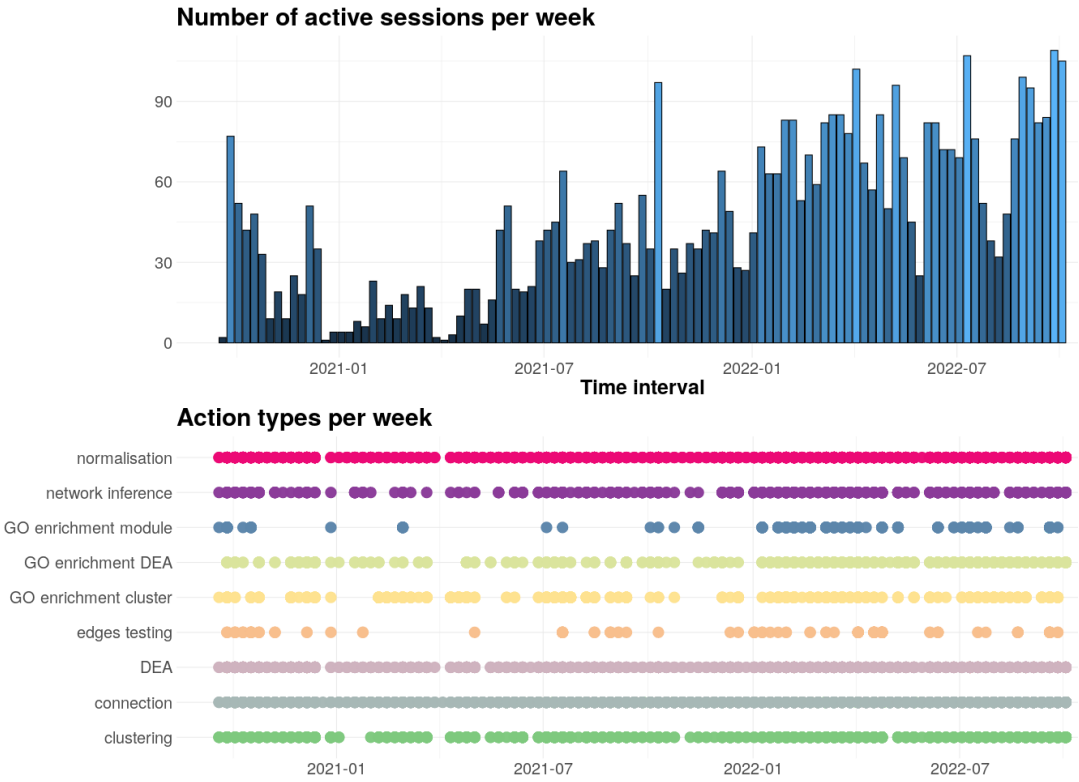


FIGURE 2.1: **DIANE online usage from its initial web deployment before publication in 2020, until the 8^{th} of october, 2022.** The first subplot displays the number of connections opened on a weekly basis. The second subplot shows the use of the different analyses proposed in DIANE along time, where a dot is present if the feature was used at least once in the considered week.

2.2 An inferred GRN identifies candidate genes in the root response to elevated CO_2 under limiting nitrate

2.2.1 Preamble

As presented in the introduction Publication #1, the deleterious impact of eCO_2 in C3 plants is linked to mineral nutrition in several ways. Firstly, mineral nutrition appears to be negatively altered, as the composition of the plant leaves declines under eCO_2 with N being especially affected [Loladze, 2014]. Our review puts forward the potential existence of transcriptional networks in the plant roots governing nitrate acquisition and metabolism genes, but such networks were not studied in whole transcriptomes yet. Second, the effect of eCO_2 on biomass gain and shoot composition depends on the nutritional status of the plant. For instance, those phenotypic responses seem to be modulated by the quantity of nutrients brought to the plant during its growth, or the form of those nutrients. It is thus of great interest to study the effect of eCO_2 in combination with different nutritional conditions for the plant. This motivated the design of a combinatorial transcriptomic dataset to elucidate eCO_2 response under sufficient and low supply of two main nutrients for the plant: nitrate and iron (Fe).

Because we use our combinatorial transcriptomic dataset to test the hypothesis of N acquisition pathway alteration by eCO_2, **this preamble starts by establishing some background knowledge about N nutrition in *Arabidopsis thaliana*, and the key genes it involves**. The macronutrient N is essential to plant growth and function, as it is a component of amino acids, proteins, nucleic acids and chlorophyll. It can represent more than 5% of the dry mass of some plant tissue. With the exception of atmospheric N_2 fixation cases by symbiotic bacteria, N is predominantly acquired by plant roots directly from the soil environment. In soils, N is mainly available to the roots in the form of nitrate (the NO_3^- molecule). Even in fertilized soils where N can be brought in the form of NH_4^+, this ion is rapidly turned into nitrate by bacterial activity. Nitrate nutrition in *Arabidopsis thaliana* can be broken down into three major categories :

- **Nitrate uptake**. Nitrate is absorbed in roots by multigenic families of transporters. In particular, root nitrate transporters belong to the NRT1 (Nitrate Root Transporter 1) or NRT2 (Nitrate Root Transporter 2) families, later renamed NRT1/NPF. Depending on the external nitrate concentration, two separate classes of transporters are active [Bellegarde, Gojon, and Martin, 2017; Crawford and Glass, 1998; O'Brien et al., 2016]. From 0.2 to 0.5 mM of NO_3^-, high-affinity transporters like NRT2.1, NRT2.2 or NRT1.1 are involved, while the even higher affinity transporters NRT2.4 and NRT2.5 come into play under starvation conditions. Some transporters may work with co-factors, like NRT2.1 with NAR2.1. In contrast, under higher nitrate concentrations (> 1 mM) low affinity transporters carry nitrate into the root cells, like NRT1.1. Interestingly, NRT1.1 belongs to both high and low affinity systems [Maghiaoui, Gojon, and Bach, 2020].

- **Nitrate assimilation**. Once absorbed by the roots, nitrate either stays in the roots or is transported to the foliar tissues in the xylem. At this step, the Nitrate Reductase (NR) and its two isoforms NIA1 and NIA2 reduce NO_3^- into NO_2^-. Then the NIR enzyme uses NO_2^- to produce $NH4_4^+$ [Vidal et al., 2020], which can be assimilated in the course of amino acids synthesis by the Glutamine Synthetase pathway.

- **Nitrate signalling**. As N availability in natural conditions displays strong spatial and temporal heterogeneity, plants have evolved mechanisms to locally and systemically tune the rate of N acquisition and assimilation. As a consequence, environmental conditions and the plant internal N status can both trigger signals controlling nitrate transport and metabolism [Bellegarde, Gojon, and Martin, 2017; O'Brien et al., 2016; Vidal et al., 2020]. Not only phosphorylation and calcium-modulated signals were observed [O'Brien et al., 2016], but transcriptional regulations have been extensively documented as well. Transcription factors from the TGA [Alvarez et al., 2014] and NLP [Marchive et al., 2013] families were shown to enhance the expression of several nitrate transporters along with *NIA1*, *NIA2* and *NIR*, while TFs of the LBD family, namely LBD37/LBD38/LBD39, are known to repress those genes [Rubin et al., 2009]. Other instances of negative regulators are HRS1 and its homologs HHO1/HHO2/HHO3, that were identified as repressing the expression of nitrate transport genes *NRT2.1*, *NRT2.4* et *NRT1.1* [Kiba et al., 2018; Safi et al., 2021]. Finally, BT1/BT2 also act as repressors of Nitrogen use efficiency genes [Araus et al., 2016], although they are not transcription factors but E3 ligases. The existence of systemic, long distance signals to regulate nitrate nutrition was also linked to mobile peptids from the CEP family. For example, low nitrate conditions in the roots can be signalled to shoot tissues by such peptids transported in the xylem, and shoot-to-root signals can be sent back by CEPD peptids of the glutaredoxin family to induce the expression of *NRT2.1* and *NRT1.1* [**ohkubo2017shoot**; Ota et al., 2020].

In Publication #3, we study the behavior of key nitrate and iron nutrition genes in plants grown under eCO_2. We apply the statistical pipeline and methodology provided by DIANE [Cassan, Lèbre, and Martin, 2021] (Publication #2) to infer the root transcriptomic response to eCO_2 combined with low nitrate conditions, in order to identify candidate genes controlling this response.

2.2.2 Publication #3 (Published)

Note : This section has its own reference system. Citation numbers refer to bibliography items included in the present article, and not at the end of the PhD manuscript. ***This article has been published in the New Phytologist, in january 2023.*** *Code and data to reproduce the results presented in this section are available in the github repository* `https://github.com/OceaneCsn/CO2_root_networks_inference`.

New Phytologist

Research

A gene regulatory network in Arabidopsis roots reveals features and regulators of the plant response to elevated CO_2

Océane Cassan[1], Léa-Lou Pimparé[1], Christian Dubos[1], Alain Gojon[1], Liên Bach[1], Sophie Lèbre[2,3] and Antoine Martin[1]

[1]IPSiM, Univ. Montpellier, CNRS, INRAE, Institut Agro, 34000 Montpellier, France; [2]IMAG, Univ. Montpellier, CNRS, 34000 Montpellier, France; [3]Université Paul-Valéry-Montpellier 3, 34000 Montpellier, France

Author for correspondence:
Antoine Martin
Email: antoine.martin@cnrs.fr

Received: *19 December 2022*
Accepted: *29 January 2023*

New Phytologist (2023)
doi: 10.1111/nph.18788

Key words: Arabidopsis, elevated CO_2, gene regulatory network, growth stimulation, mineral nutrition.

Summary

- The elevation of CO_2 in the atmosphere increases plant biomass but decreases their mineral content. The genetic and molecular bases of these effects remain mostly unknown, in particular in the root system, which is responsible for plant nutrient uptake.
- To gain knowledge about the effect of elevated CO_2 on plant growth and physiology, and to identify its regulatory in the roots, we analyzed genome expression in Arabidopsis roots through a combinatorial design with contrasted levels of CO_2, nitrate, and iron.
- We demonstrated that elevated CO_2 has a modest effect on root genome expression under nutrient sufficiency, but by contrast leads to massive expression changes under nitrate or iron deficiencies. We demonstrated that elevated CO_2 negatively targets nitrate and iron starvation modules at the transcriptional level, associated with a reduction in high-affinity nitrate uptake. Finally, we inferred a gene regulatory network governing the root response to elevated CO_2. This network allowed us to identify candidate transcription factors including MYB15, WOX11, and EDF3 which we experimentally validated for their role in the stimulation of growth by elevated CO_2.
- Our approach identified key features and regulators of the plant response to elevated CO_2, with the objective of developing crops resilient to climate change.

Introduction

The atmospheric concentration of carbon dioxide (CO_2) is expected to reach between 750 ppm and > 1000 ppm at the end of the century (IPCC, 2021). Elevated atmospheric CO_2 (eCO_2) will profoundly modify plant physiology, as CO_2 is the primary substrate of photosynthesis. This is illustrated by the eCO_2 fertilization effect, which leads in C_3 plants to an enhanced photosynthesis, a stimulation of growth, and an accumulation of biomass for plants grown under eCO_2 condition (Ainsworth & Long, 2021). The stimulation of plant growth by eCO_2 has significant implications, as an augmentation of green biomass and yield is required for satisfying the increasing demand for food, and for mitigating the rise of the CO_2 concentration in the atmosphere. Nevertheless, a large number of CO_2-enrichment experiments, conducted both in fields and in laboratories, have returned that the gain of biomass for plants grown under eCO_2 is much lower than theoretically expected due to plant acclimation to eCO_2 (Tausz-Posch *et al.*, 2020). Acclimation of plants to eCO_2 is usually associated with a negative feedback of photosynthesis due to the accumulation of sugars, and to a decrease in leaf Rubisco content (Thompson *et al.*, 2017; Tausz-Posch *et al.*, 2020; Ainsworth & Long, 2021), but the genetic basis remains poorly understood. In addition to this, growing C_3 plants under eCO_2 leads to an unexpected decline of their mineral composition (Loladze, 2014; Myers *et al.*, 2014; Gojon *et al.*, 2022). Indeed, plants grown under eCO_2 show a decrease in the tissue concentrations of most mineral nutrients compared with those grown under ambient CO_2 (aCO_2), especially concerning nitrogen (N) and essential micronutrients like iron (Fe). The acclimation of plants to eCO_2 and the negative effect of eCO_2 on plant mineral content are concerning for food security, and present a serious threat of increasing starvations in the coming decades, especially for populations already at risk (Smith & Myers, 2018). Several physiological hypotheses have been proposed to explain this negative effect of eCO_2 on plant mineral composition. Among them, general hypotheses have been highlighted, such as a dilution of nutrients in higher biomass or a reduced root-to-shoot translocation of nutrients due to a lowered transpiration rate and lower stomatal conductance under eCO_2 (Tausz-Posch *et al.*, 2020). For instance, it was shown recently that increasing transpiration in the *aca7* mutant partially restores the content of Fe in seeds under eCO_2 (Sun *et al.*, 2022). In Arabidopsis and wheat, it has been demonstrated that eCO_2 negatively affects the uptake and reduction in nitrate, highlighting a close link between eCO_2 and N nutrition (Bloom *et al.*, 2010, 2014). Strikingly, not much is known about the regulatory mechanisms that are associated with the plant acclimation to

eCO_2 and to the negative effect of eCO_2 on plant mineral composition. Only a handful of transcriptomic experiments analyzing plants under eCO_2 have been performed. These experiments strongly suggested that eCO_2 has a minor effect on genome expression (Miyazaki *et al.*, 2004; Taylor *et al.*, 2005; Ainsworth *et al.*, 2006; Li *et al.*, 2006, 2008; Tallis *et al.*, 2010; Vicente *et al.*, 2019). In Arabidopsis leaves, transcriptomic markers similar to those found under N deficiency have been found under eCO_2, for example, the expression of several genes usually downregulated by N limitation was also lower under eCO_2 condition (Li *et al.*, 2006). Several genes coding for major actors of Fe acquisition, such as the transporter *IRT1*, have been also identified as downregulated by eCO_2 in rice leaves (Yang *et al.*, 2020). On the other hand, despite the crucial importance of roots for the homeostasis of nutrients and for growth, the effect of eCO_2 on root genome expression has been very poorly investigated. Several studies, nevertheless, suggested that eCO_2 leads to a misregulation of some genes associated with nitrate or Fe homeostasis, including nitrate transporter genes from the *NRT1* and *NRT2* families (Jauregui *et al.*, 2015; Vicente *et al.*, 2015, 2016; Bencke-Malato *et al.*, 2019). Collectively, these data suggest that eCO_2 has a significant impact on the regulation of signaling pathways associated with growth and mineral nutrition. However, there is no clear view concerning the effect of eCO_2 on these signaling pathways in the roots yet, and so far, no regulators of the response to eCO_2 in the roots have been identified. The objective of this work was to contribute to the understanding of the effect of eCO_2 on plants through the study of regulatory mechanisms in the roots, and how they affect two main phenotypes induced by eCO_2: the stimulation of biomass production and the alteration of mineral content.

To do so, we performed a combinatorial analysis of the effect of eCO_2 on root genome expression under contrasting provision of nitrate and/or Fe, to reveal features by which eCO_2 disrupts the regulation of major root function, and to identify regulators of this response through the inference of gene regulatory networks (GRNs). The targeted analysis of genome expression data demonstrated that eCO_2 severely disrupts the expression of regulatory modules associated with nutrient limitation, including the negative regulators of nitrate signaling and of Fe starvation. In addition, the inference of GRNs reveals several candidate genes for the regulation of the response to eCO_2 in the roots. The analysis of these candidate regulators notably demonstrated that *MYB15*, *WOX11*, and *EDF3* transcription factors are required to reach the full potential of growth stimulation and biomass accumulation in a CO_2-rich atmosphere under nitrate limiting condition, without penalizing the mineral composition of plants.

Materials and Methods

Biological material

Plant growth conditions and material *Arabidopsis thaliana* plants were grown in hydroponics using nutrient solution as described by Gansel *et al.* (2001). Nitrate concentration was either 10 mM (high nitrate) or 0.5 mM (low nitrate) KNO_3 during all the experiments. Fe was supplied at a concentration of 50 µM during all the experiments for plants under Fe supply, or was removed from the medium during the last week of growth for plants undergoing Fe starvation. CO_2 conditions in the chambers were constantly maintained at air (*c.* 420 ppm, aCO_2) or 900 ppm (eCO_2), under 200 µmol m^{-2} s^{-1} light intensity, and 8 h : 16 h, light : dark, 22°C : 20°C photoperiod. Plant shoots and roots were sampled after 5 wk of growth. Accessions and mutant alleles used in this study are Columbia (Col), Wassilewskija (WS), *rsl5* (WiscDsLox384E5), *myb15* (SALK 151976), *myb85* (SALK 052089), *wrky59* (SALK 102984), *wox11-2* (SALK 004777), and *edf3* (FLAG 606H09). Mutants were obtained from the Nottingham and the Versailles Arabidopsis Stock Centers.

RNA extraction, quantification of transcripts, and RNA-sequencing

Five plant roots from identical conditions were pooled together into one biological replicate, flash frozen in liquid nitrogen, and stored at −80°C. RNAs were extracted from root tissues using Trizol (Invitrogen), and DNAse treated using RQ1 (Promega). Reverse transcription was achieved from 1 µg of total RNA with M-MLV reverse transcriptase (RNase H minus, Point Mutant; Promega) using an anchored oligo(dT)20 primer. Accumulation of transcripts was measured by qRT-PCR (LightCycler 480; Roche Diagnostics, Bâle, Switzerland) using the SYBR Premix Ex Taq™ (TaKaRa, Kusatsu, Japan). Gene expression was normalized using *UBQ10* and *ACT2* as internal standards. Results are presented as the expression relative to *UBQ10*. Sequences of primers used in RT-qPCR for gene expression analysis are listed in Supporting Information Table S1. RNA-sequencing libraries were done from root total RNA using standard RNA-Seq protocol method (Poly-A selection for mRNA species) by the Novogene Co. (Cambridge, UK). RNA-sequencing was performed using Illumina technology on a NovaSeq6000 system providing PE150 reads.

Biomass and nutrient-related measurements

In all, 15–20 rosettes were dried in an oven at 70°C for 72 h, and plant shoot biomass was measured using a precision weighing scale. Nitrogen and carbon composition of shoots was obtained using an Elementar Pyrocube analyzer. Fe content was measured using acidic digestion and a microwave plasma atomic emission spectrometer (MP-AES; Agilent, Santa Clara, CA, USA). Nitrate uptake was measured by supplying $^{15}NO_3^-$ (1 atom% excess (^{15}N)) in hydroponic solution for 72 h. Roots and shoots were then dried at 70°C for 72 h, and the samples were analyzed for total N and atom% ^{15}N using a continuous flow isotope ratio mass spectrometer coupled with a C : N elemental analyzer (model Euroflash; Eurovector, Pavia, Italy).

Processing of raw RNASeq files

The quality control and adapter trimming of raw paired-end FASTQ files was done with FASTP and its default parameters. Mapping to the TAIR10 reference genome was performed with STAR, and the options following:

--outSAMtype BAM SortedByCoordinate
--outFilterMismatchNmax 1
--outFilterMismatchNoverLmax 0.15
--alignIntronMin 30
--alignIntronMax 5000

Quantification of the bam files against the TAIR10 GFF3 annotation file was done using *htseq-count* with options:

-f bam --type gene -r pos
--idattr=Name --stranded=no

Statistical analyses of phenotypic data

We fit linear models to quantitative traits using categorical predictors via the *lm()* R function. We focused on the interpretation of interaction terms, of which we assessed the significance based on the *t*-test performed on each regression coefficients given by the *summary()* R function.

Transcriptomic analyses

Transcriptomes normalization, principal component analysis (PCA), and differential expression: the raw expression matrix was normalized using the TMM method. Lowly expressed genes with an average value across conditions under 10 were excluded from the analysis. PCA was carried out on normalized transcriptomes via the ADE4 R package. Differential expression was tested using the EDGER R package as proposed in the DIANE R package (Cassan *et al.*, 2021), with no fold change constraint, and an adjusted *P*-value threshold (FDR) of 0.05.

Multivariate expression-based gene clustering

The COSEQ (Rau & Maugis-Rabusseau, 2018) package embedded in DIANE was used to partition genes based on their expression changes across conditions. The underlying framework is the framework of mixture models: Gaussian mixtures were fit to each cluster after applying a prior arcsin transformation to the normalized counts. The model parameters, that is, the mixing proportions and the cluster-specific distributions parameters are estimated through an Expectation–Maximization algorithm. A global quality score can be computed to evaluate a given clustering model: we used the Integrated Complete Likelihood. The final number of clusters (9) was chosen where the Integrated Complete Likelihood reached a plateau (elbow method).

GRN inference

To reconstruct the transcriptional dependencies between genes, we relied on the network inference method GENIE3 (Huynh-Thu *et al.*, 2010), extended by a permutation-based approach to sparsify its output as implemented in DIANE (Cassan *et al.*, 2021). GENIE3 was shown to be among the best performers in benchmark studies such as the DREAM challenges (Marbach *et al.*, 2012), and allows to quantify the strength of regulatory influences between regulator genes and their targets. This influence is extracted from a regression framework: random forests are fit to predict the expression of target genes using the expression of regulator genes as predictors. In the process of fitting the regression trees, the importance of the predictors can be extracted so that a ranking of all regulator-target pairs is obtained. To select the strongest regulatory interactions among that ranking, we first built a biologically relevant network with a connectivity density of 0.03, made of the regulatory pairs with the strongest importance values. Then, we used a permutation-based procedure to estimate null distributions of random forest importance values against which we tested the observed importance, and selected the interactions with an adjusted *P*-value (FDR method) below 5%. Network handling and the extraction of network-related metrics were allowed by the *igraph* library.

GRN validation

To validate the inferred network, we made use of the R package ARANETBENCH. In the network evaluation process, the inferred network is first transformed so that grouped regulators are ungrouped, duplicating their interactions with target genes. Then, each one-to-one regulator-to-target link is compared to DAPSEQ, CHIPSEQ, or TARGET databases. The validation rate of a network is computed as the number of links supported by at least one experiment, divided by the total number of links for which the regulator was experimentally studied. The statistical significance of this validation rate is then assessed by comparing it to the validation rates of a large population of networks with randomly swapped edges. To avoid confusing biases, those random networks and the inferred network are composed of the same nodes, and the regulator's degrees remain unchanged so that the overall connectivity distribution is preserved.

Community discovery

The Stochastic Block Model (SBM) partitioning was determined via the SBM R package. The optimal number of communities was determined automatically, as the number of communities maximizing the inferred Block Model's quality criteria: the Integrated Complete Likelihood.

Gene ontology enrichment analyses

DIANE's wrapper of CLUSTERPROFILER was used to detect significantly overrepresented ontologies, which relies on fisher's exact test with an adjusted *P*-value threshold of 0.05. The gene background used to assess enrichments was the list of all Arabidopsis genes.

Results

eCO_2 leads to profound reprogramming of genome expression under nutrient starvation conditions

To explore the effect of eCO_2 on the root regulatory responses under nutrient limitation, Arabidopsis plants were subjected to a combination of treatments including CO_2 (ambient, 420 ppm or

elevated, 900 ppm), nitrate (high provision, 10 mM or low provision, 0.5 mM), and Fe (sufficient provision, 50 μM or starvation for 1 wk) (Fig. 1a). First, we looked at the expression in the roots of known marker genes of nitrate or Fe nutrition. We observed that eCO_2 leads to a decrease in the expression of genes involved in nitrate uptake and assimilation such as the high-affinity root nitrate transporter genes *NRT2.1* and *NRT1.1*, and the nitrate reductase gene *NIA1*, especially under nitrate limitation (Fig. 1a). We also showed that eCO_2 inhibits the induction of the expression of major markers involved in the Fe starvation response such as the Fe transporter gene *IRT1* and the Fe chelate reductase gene *FRO2* (Fig. 1b). Therefore, eCO_2 seems to disrupt the expression of key actors involved in the response to nitrate or Fe deficiencies. To fully explore the genome expression changes in the roots induced by CO_2 elevation and how they affect responses to nutrient starvation, we performed root RNA-seq for the eight combinations of CO_2 levels, nitrate supply, and Fe supply.

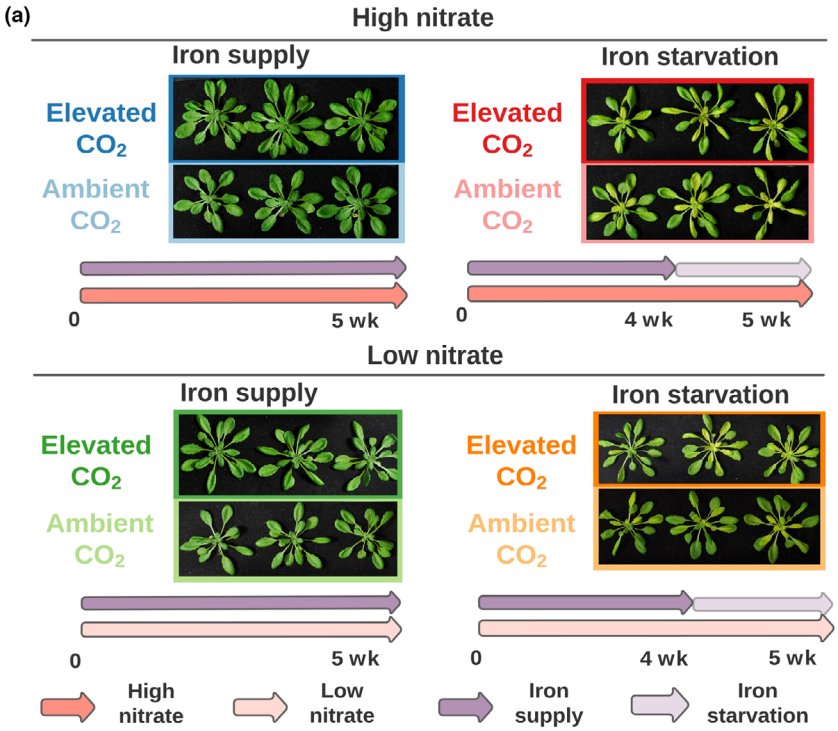
First, a PCA on the whole dataset revealed that the first principal component, which explains > 50% of gene expression variation, mainly discriminated genes differentially expressed in response to Fe starvation (Fig. 2a,b). The second principal component, explaining 12.9% of gene expression variation, separated genes differentially expressed in response to nitrate provision, especially under Fe starvation. The effect of nitrate provision under Fe supply on the root transcriptome, mainly visible through the third principal component, explained < 10% of gene expression variation in the dataset. Lastly, the effect of eCO_2 on the root transcriptome was explained by further principal components 4, 5, and 6, carrying together 11.2% of gene expression change. Therefore, we concluded that the effect of eCO_2 on the root transcriptome was modest in comparison to those of nutrient starvation.

However, a more striking observation was made when looking at the number of genes differentially expressed by eCO_2 depending on the provision of nitrate and Fe. Indeed, very few genes were found to be differentially expressed by eCO_2 when plants were grown under sufficient nitrate and Fe provision (124 genes upregulated or downregulated; FDR ≤ 0.05) (Fig. 2c; Tables S2–S5). On the contrary, many more genes were found to be differentially expressed when eCO_2 was combined to at least one nutrient limitation. Growth under eCO_2 condition leads to 1550 differentially expressed genes under nitrate limitation, to 3524 genes under Fe starvation, and to 2429 genes under the combination of nitrate limitation and Fe starvation (Fig. 2c; Tables S2–S5). Therefore, we concluded that eCO_2 has a limited effect on root transcriptome when plants grow under sufficient nutrient conditions, but leads to profound reprogramming of genome expression when plants grow under nutrient starvation conditions.

eCO_2 disrupts gene expression associated with nitrate and Fe starvation signaling pathways

First, we adopted a targeted approach by exploring the expression profile of marker genes involved in nitrate and Fe responses, and their regulation. In the case of nitrate response, we observed that an important number of genes involved in nitrate transport and assimilation were significantly downregulated by eCO_2, especially under nitrate limitation (Fig. 3a). In line with the qRT-PCR data shown earlier (Fig. 1b), this was the case for nitrate transporter genes *NRT2.1*, *NAR2.1*, and *NRT1.1*, nitrate and ammonium assimilation genes *NIR1* and *GLN1.2*, and nitrate-responsive genes like *G6PD3*. Other genes like *NRT2.2*, *NIA1*, or *NIA2* were also downregulated but to a lesser extent (Tables S2–S5). In parallel to this, we found that genes involved in the positive regulation of the nitrate starvation response, such as *NLP2*, *TGA4*, or *CEP9* were also downregulated by eCO_2. In opposition, we strikingly observed that numerous genes involved in the negative regulation of nitrate transport and assimilation were upregulated by eCO_2, specifically under nitrate limitation. This was the case of *BT1* and *BT2*, known to downregulate the expression of high-affinity nitrate transporters such as *NRT2.1*, but also for members of the *NIGT* transcription factor family that repress nitrate transporter genes under satiety conditions (Araus *et al.*, 2016; Kiba *et al.*, 2018) (Fig. 3a). In addition, we also found that the expression of *LBD41*, a close homolog of the *LBD* transcription factors sub-clade that repress nitrate transport and assimilation (Rubin *et al.*, 2009), was also induced by eCO_2. Altogether, these observations show that under nitrate limitation, eCO_2 markedly affects the expression of nitrate signaling modules by upregulating the expression of negative regulators of nitrate uptake and assimilation, and accordingly, by downregulating the expression of nitrate uptake and assimilation genes. In the case of Fe-related gene expression, we observed a similar deregulation of signaling modules under eCO_2. Indeed, several genes that are induced by Fe starvation under ambient CO_2, such as *IRT1*, *FRO2*, or the coumarin transporter *PDR9*, were much less induced or even not induced anymore under eCO_2 (Fig. 3b). In addition to this, we found that the regulators of Fe starvation response such as the transcription factors *FIT* or *BHLH39*, that are induced by Fe starvation under ambient CO_2, were much less induced or even not induced anymore under eCO_2 (Fig. 3b). These observations show that eCO_2 also disrupts the Fe starvation response, and then, more generally, that eCO_2 has a strong negative effect on the expression of signaling modules associated with deficiencies in major nutrients.

To see whether these observations coincide with the well-known effects of eCO_2 on physiological parameters in shoots, we measured biomass accumulation, N and Fe concentrations in shoots under each condition. We observed that eCO_2 significantly led to increased biomass, and this regardless of nutrients availability (Fig. 4a). While biomass is consistently increased by eCO_2, we observed contrasted effects on the mineral composition of plants. Growth under eCO_2 led to a strong and significant decline in shoot N concentration when plants were grown under low nitrate conditions, but not under high nitrate conditions (Fig. 4b). From aCO_2 to eCO_2, N concentration dropped by 30% under low nitrate and Fe supply, and by 15% under low nitrate and Fe starvation (Fig. 4b). Surprisingly, the growth under eCO_2 did not lead to a decrease in Fe content, regardless of the nutritional condition (Fig. 4c), suggesting that Fe



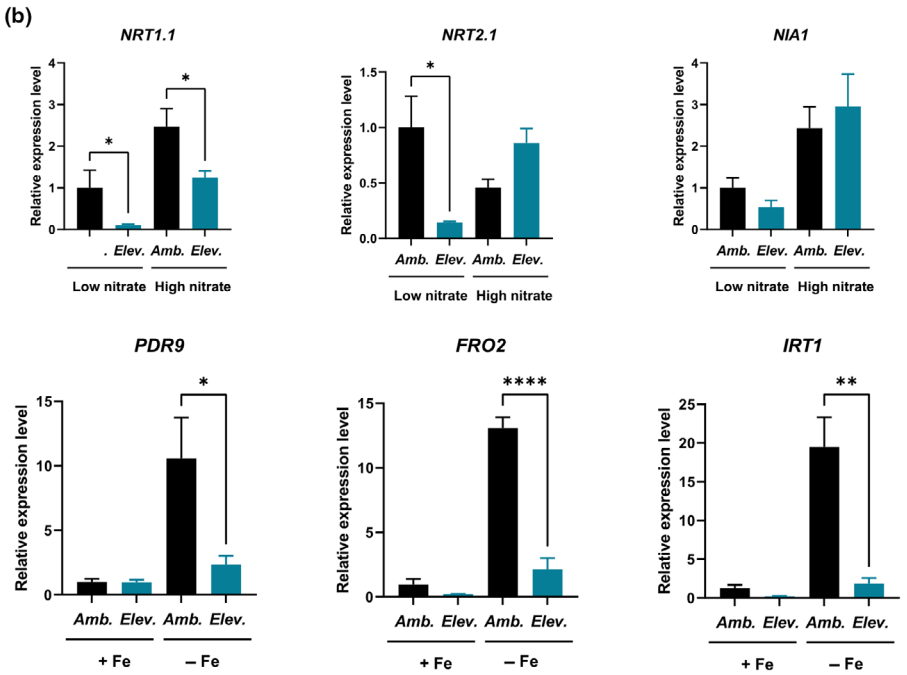


Fig. 1 Combinatorial analysis of the effects of eCO_2, nitrate limitation, and Fe starvation. (a) Design of the combinatorial experiment combining contrasting levels of CO_2, nitrate and Fe. Arabidopsis plants were grown in hydroponics for 5 wk under contrasted levels of CO_2, nitrate, and Fe. High nitrate (10 mm), low nitrate (0.5 mM), +Fe (50 μM), −Fe (0 μM), Amb. (ambient CO_2, 420 ppm), and Elev. (elevated CO_2, 900 ppm). For each combination of nitrate and Fe supply levels, three representative rosettes are shown under aCO_2 and eCO_2. (b) Quantitative RT-PCR showing the relative expression of marker genes of nitrate and Fe nutrition in the roots, in response to the combination of CO_2 and nitrate, or CO_2 and Fe supply. High nitrate (10 mM), low nitrate (0.5 mM), +Fe (50 μM), −Fe (0 μM), Amb. (ambient CO_2, 420 ppm), and Elev. (elevated CO_2, 900 ppm). Data represent mean ± SD of five biological replicates from a representative experiment. Statistical significance was computed using an unpaired two-tailed Student's *t* test (*, $P \leq 0.05$; **, $P \leq 0.01$; ****, $P \leq 0.001$).

provision or Fe accumulation in plants was too high to be affected by eCO_2, even after 1 wk of Fe starvation. Finally, we investigated whether the alterations in the expression of nitrate signaling modules under eCO_2 translated into altered nitrate uptake capacities. Indeed, in accordance with the observations made at the level of gene expression and total N concentration levels, we clearly observed that the nitrate uptake rate was significantly and negatively affected by eCO_2 under limiting nitrate condition, but not affected under high nitrate condition (Fig. 4c).

eCO_2 largely disrupts the response to nitrate limitation

Considering that the major effect of eCO_2 on shoot biomass and N concentration were observed under low nitrate conditions, we investigated in a broader way the impact of eCO_2 on the

New Phytologist

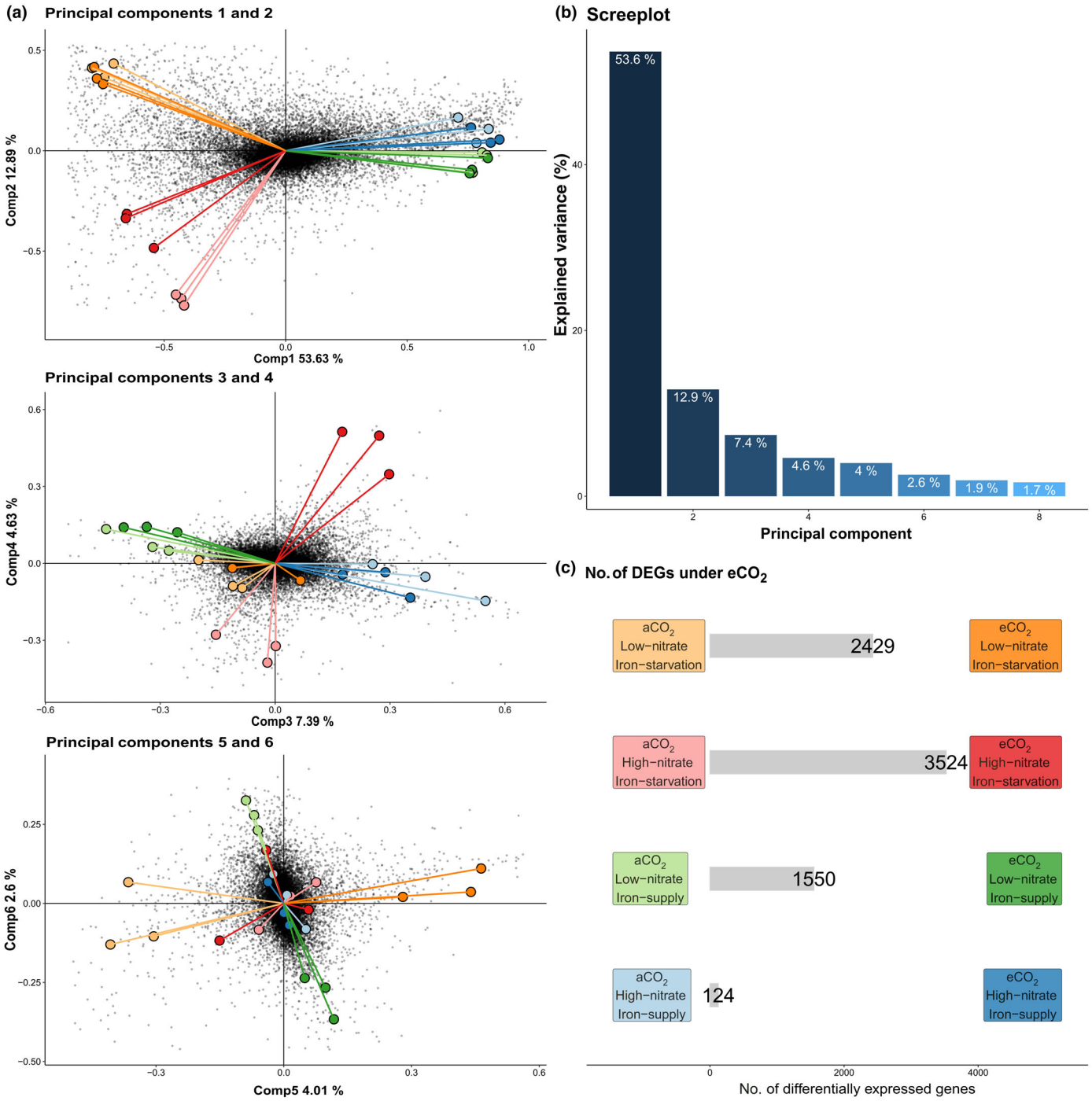


Fig. 2 Elevated CO_2 reprograms root genome expression under nutrient deficiency. (a) Principal component analysis of the normalized root transcriptomes under the eCO_2 and nutrient limitations combinatorial design. Contributions (i.e. correlation) of each experimental condition to the six first principal components. Each color stands for an experimental condition. Genes are shown as dots. (b) Percentage of variance explained by each of the principal components determined by the analysis. (c) Number of genes differentially expressed by eCO_2 under different conditions of nitrate and Fe provision.

reprogramming of genome expression by nitrate limitation. To do so, we performed a clustering approach, focusing our analysis on the interaction between eCO_2 and nitrate limitation response. The 1550 differentially expressed genes in response to high CO_2 under nitrate limitation (Fig. 2c) were partitioned based on their co-expression in the four (triplicated) experimental conditions with reference and perturbation levels of CO_2 and nitrate supply (in all cases with sufficient Fe). The mixture models-based approach we employed (Rau & Maugis-Rabusseau, 2018) led to an organization in nine clusters (Fig. 5). In line with the profile of nitrate-responsive marker genes (Fig. 3a), we observed that the regulation of the expression of a large number of genes by nitrate is strongly modified under eCO_2 condition. In clusters 5 and 9, we observed >300 genes that are downregulated by nitrate

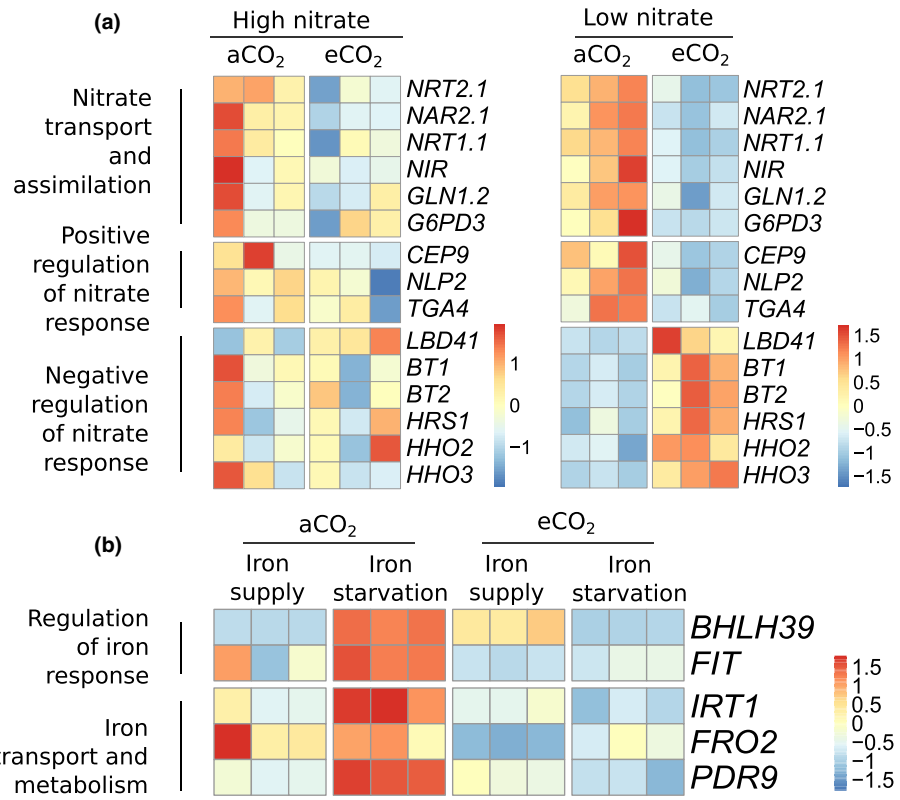


Fig. 3 Elevated CO_2 disrupts the expression of key nitrate and Fe marker genes specifically under nitrate limitation and Fe starvation. Heatmap representation showing the *Z*-score of normalized expression levels of genes important for nutrition in the transcriptomic dataset. (a) Regulation by eCO_2 of genes involved in nitrate transport and assimilation and in their regulation, under high or limiting nitrate conditions. (b) Regulation by eCO_2 of genes involved in Fe transport and in their regulation, under Fe supply or Fe starvation.

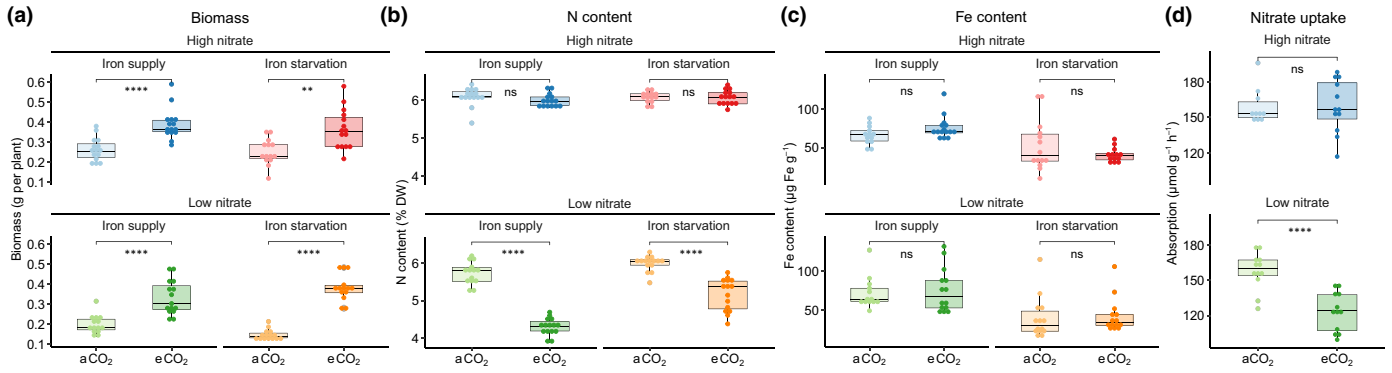


Fig. 4 Elevated CO_2 modifies plant physiology depending on nutrient supply. Shoot biomass (a), N concentration (b), and Fe content (c) were measured on individual rosettes from plants grown for 5 wk in hydroponics under contrasted levels of CO_2, nitrate, and Fe. (d) Nitrate uptake measured on plants grown in hydroponics under contrasted levels of CO_2 and nitrate. Statistical tests for comparison between aCO_2 and eCO_2 were performed with the Wilcoxon test. High nitrate (10 mM), low nitrate (0.5 mM), Fe supply (50 µM), Fe starvation (0 µM), aCO_2 (*c.* 420 ppm), and eCO_2 (900 ppm). In each panel, horizontal lines, whiskers and circles correspond to medians, first quartiles and individual samples, respectively.

limitation under aCO_2 condition, but not under eCO_2 condition, as observed for *HRS1*, *BT1*, and *BT2*. This suppression of regulation by eCO_2 was also observed for genes induced under nitrate limitation; in clusters 7 and 8, > 200 genes have their expression induced by nitrate limitation under aCO_2 condition but not under eCO_2 condition, as observed for *CEP9*, *NLP2*, and *NRT1.1*. In an even more pronounced way, we observed that growth under eCO_2 reversed the regulation by nitrate availability of several hundreds of genes, found in clusters 2, 3, and 6. Finally, cluster 4 contains almost 200 genes showing modest or no expression changes induced by nitrate limitation, but that are considerably repressed under the combination of nitrate limitation and eCO_2. Several nitrate transport or assimilation genes, such as *NRT2.1*, *NAR2.1*, *NIR*, or *GLN1.2*, belong to this cluster. Based on this co-expression study, we show that the growth under eCO_2 either lessens, cancels, or even reverses nitrate-induced regulation of the expression of a large number of genes. This strengthens the hypothesis that eCO_2 severely affects regulatory pathways involved in the adaptation to nitrate limitation.

GRN inference yields insightful genes communities and candidate regulators of the response to eCO_2 under low nitrate

Second, we adopted a non-targeted approach to identify in the roots the regulators that drive the response mediated by eCO_2,

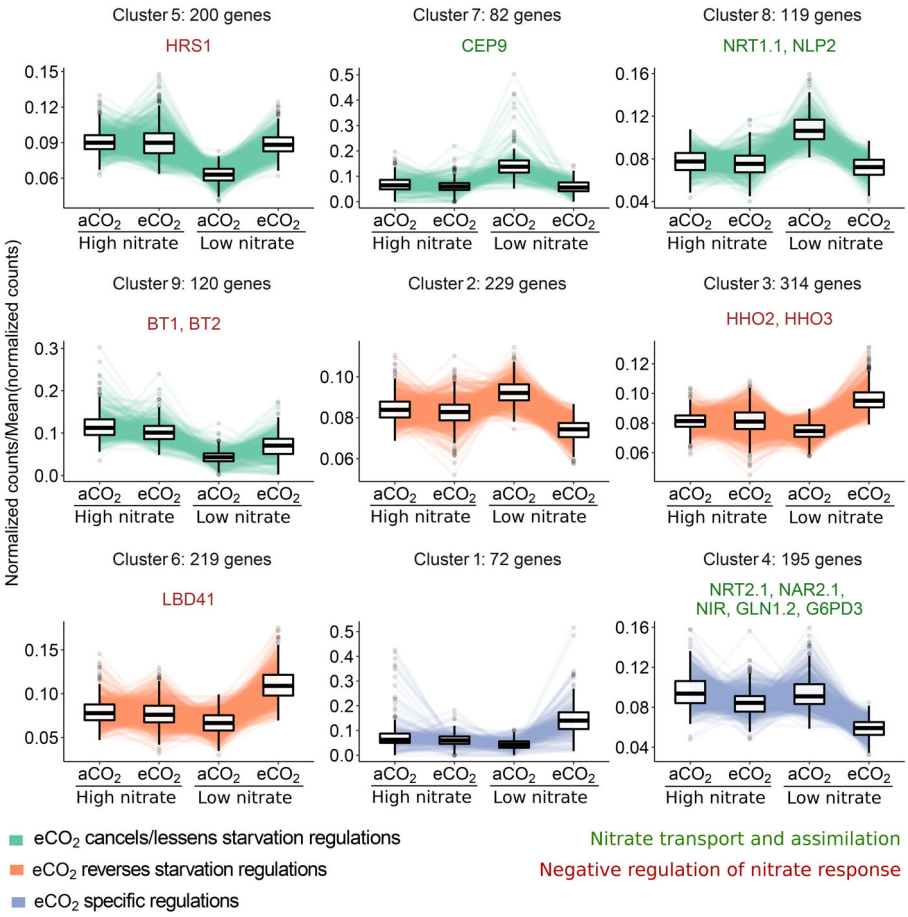
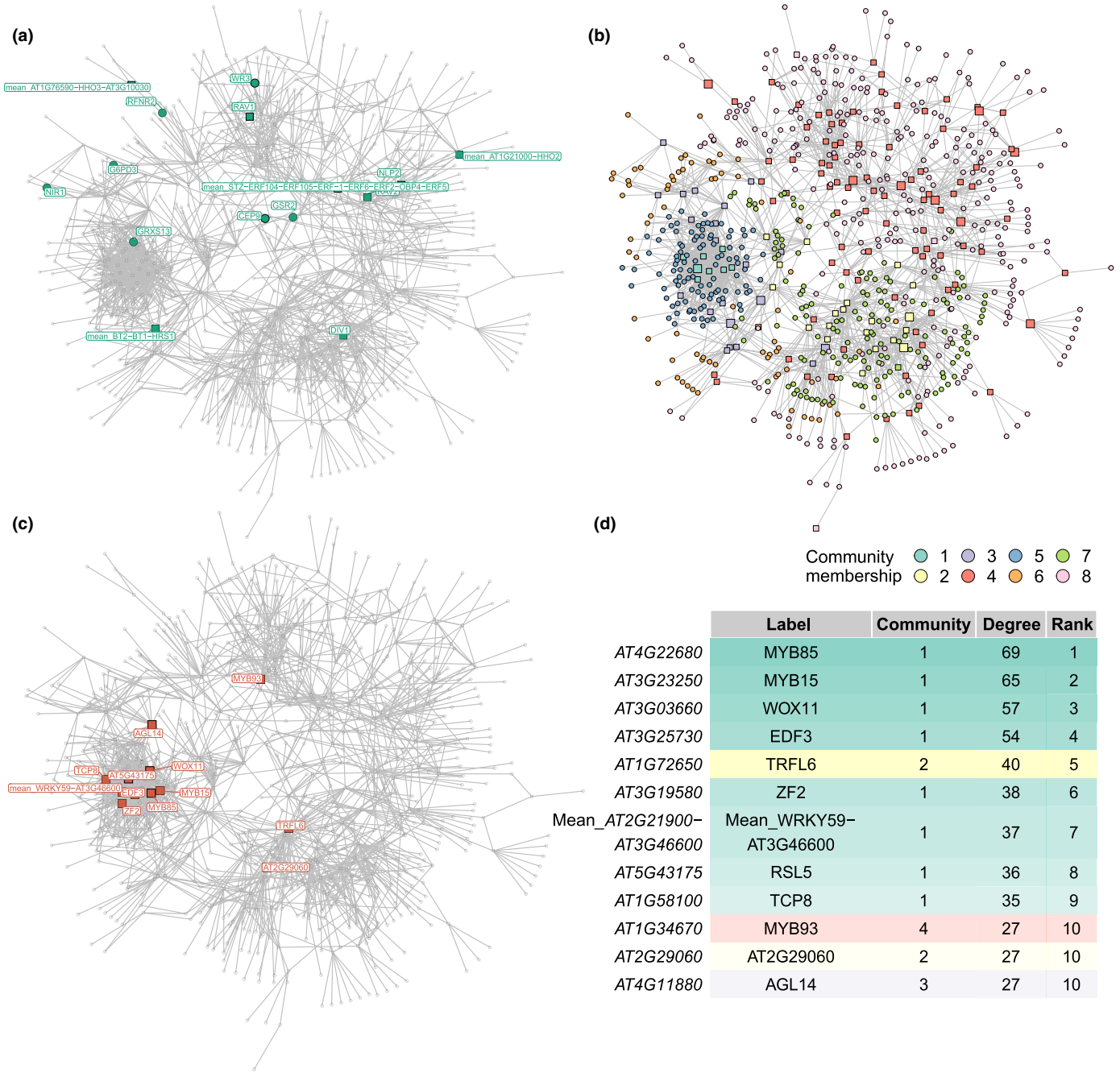


Fig. 5 Clustering analysis of the 1550 DEGs by eCO_2 under low nitrate, on the four combinations of aCO_2 or eCO_2, and high or low nitrate. Expression profiles are defined as the normalized expression, divided by the mean normalized expression in all conditions. Genes were partitioned between nine clusters. In some clusters, gene names previously identified as key actors of nitrate nutrition are highlighted, either in green (for actors or positive regulators) or red (for negative regulators). We identified three cluster categories based on the interpretation of their expression changes, as represented by the color of the expression profiles. Green, cancels/lessens starvation regulations; orange, reverses starvation regulations; blue, specific regulations. In each cluster, horizontal lines, whiskers and circles correspond to medians, first quartiles and expression of individual genes, respectively.

and that would be able to modulate the stimulation of growth or the alteration of N concentration under eCO_2. To do so, we inferred a GRN using the 1550 genes found differentially expressed by eCO_2 under low nitrate supply (Table S5). Using as input the expression values obtained under aCO_2, eCO_2, limiting or abundant nitrate supply (12 samples in total), we employed random forest regressions combined with permutation-based statistics (Huynh-Thu *et al.*, 2010; Cassan *et al.*, 2021) to draw connections between regulator and target genes. To do so, we used a list of regulators composed of Arabidopsis transcription factors, enriched with indirect regulators of gene expression (i.e. chromatin regulators) and previously known indirect regulators of nitrate-related gene expression such as *BT1*/*BT2* (Table S6). In this approach, a regulator was linked to a target if its expression was a robust predictor of the expression of the target in the input samples (see the Materials and Methods section). The resulting GRN is made of 796 nodes (647 target genes and 149 regulators) and 1700 connections (Fig. S1).

To validate the inferred edges against known regulatory interactions, we leveraged the ConnecTF database (Brooks *et al.*, 2021), composed of *in vitro* (DAPSeq) and *in vivo* (CHIP-Seq) binding experiments, as well as in planta regulation assays. Considering all the regulators for which validation information is available, 31.3% of the predicted interactions in the inferred network were supported by at least one experiment in ConnecTF (Fig. S2). To determine the significance of this validation percentage, we compared it to the validation percentages of a population of 200 shuffled networks, in which the interactions between the same genes were randomly unmatched. No random network had a higher validation rate than our inferred network, leading us to conclude that our inference approach captured consistent biological information linked to the regulation of gene expression in the roots by eCO_2 under low nitrate conditions (Fig. S3).

We tested the presence of the known regulators of the nitrate response signaling pathway and their targets on this GRN, to appreciate their influence on the effect of eCO_2 on root genome expression under low nitrate conditions (Fig. 6a). Surprisingly, most of the important actors of nitrate uptake and assimilation such as *NRT2.1*, *NRT1.1*, or *NIA* genes were not present in the network, with the exception of *NAR2.1* and *NIR1*. By contrast, an important number of known regulators of the nitrate response signaling pathway were found in the GRN. Notably, *BT1*, *BT2*, and *HRS1* were grouped together because of the high correlation value of their expression profiles in the 12 samples, which supports their similar function in the repression of nitrate limitation signaling pathways. To continue the general analysis of this GRN, we used the SBM framework to analyze the topological structure of the GRN and identify gene communities. This revealed a modular and disassortative topology, typical of biological networks (Fig. 6b). In each of the eight gene communities



	Label	Community	Degree	Rank
AT4G22680	MYB85	1	69	1
AT3G23250	MYB15	1	65	2
AT3G03660	WOX11	1	57	3
AT3G25730	EDF3	1	54	4
AT1G72650	TRFL6	2	40	5
AT3G19580	ZF2	1	38	6
Mean_*AT2G21900*−*AT3G46600*	Mean_WRKY59−AT3G46600	1	37	7
AT5G43175	RSL5	1	36	8
AT1G58100	TCP8	1	35	9
AT1G34670	MYB93	4	27	10
AT2G29060	AT2G29060	2	27	10
AT4G11880	AGL14	3	27	10

Fig. 6 Inferred gene regulatory network (GRN) of the root response to eCO_2 under low nitrate. In each network panel, round nodes are target genes, square nodes are regulator genes, and large square nodes are regulator genes grouped together because of a high correlation. (a) Network view with a highlight on actors of nitrate acquisition, signaling and metabolism and their regulators. (b) Topological clustering in the inferred GRN. Each color represents a community of highly connected genes. (c) Network view with a highlight on the 12 most influencing candidate regulators identified on the basis of their overall degree. (d) Ranking of these the 12 most influencing candidate regulators in the inferred GRN, with their TAIR AGI, common name, topological community, and overall degree.

identified, we performed gene ontology (GO) enrichment. Notably, genes included in the communities 1 and 3 displayed significant enrichment for GO related to N like 'response to N compound', 'response to nutrient levels', or 'response to nitrate' (Fig. S4). This supports again the hypothesis according to which eCO_2 severely affects the regulation of genome expression associated with N and with nitrate in particular.

MYB15, *EDF3*, and *WOX11* regulate the stimulation of plant growth by eCO_2 under low nitrate

To identify candidate regulators of the response to eCO_2, we hypothesized that the regulators displaying the highest degree of connectivity to target genes are the most relevant to control a large part of the GRN, and thus of the response to eCO_2.

Therefore, we ranked the regulators present in the GRN by their degree (Table S7). We analyzed the topological distribution of the 12 most connected regulators (Fig. 6c,d). Most of them belonged to the topological cluster 1, which is enriched for genes associated with the response to N compounds, nutrients, or nitrate. We used knockout mutant lines for several of these most connected regulators to test their involvement in the plant's response to eCO_2. Mutants for *MYB15*, *MYB85*, *WOX11*, *EDF3*, *WRKY59*, and *RSL5* transcription factors were grown under eCO_2 and limiting nitrate conditions, and phenotyped for their shoot biomass and N concentration. We further analyzed whether biomass and N concentration changes caused by eCO_2 in each genotype were statistically different from the change observed in the wild-type (WT). First, there was no evidence in our analysis that any of these transcription factors were involved in the control of N concentration in the shoot under eCO_2, as no single mutation in these regulators led to a significant change in the decrease in N concentration induced by high CO_2 compared with that of WT (Fig. 7a,b). By contrast, we observed that the stimulation of growth induced by eCO_2 was significantly lower in myb15, wox11, and edf3 than in the WT (Fig. 7c,d). The stimulation of growth by eCO_2 was reduced by 40%, 25%, and 49% in myb15, wox11, and edf3, respectively. In *myb15* and *wox11*, the biomass appears already affected under aCO_2. However, the effect of *myb15* and *wox11* mutations was significantly stronger on the biomass stimulation by eCO_2 than on the biomass under aCO_2 (Table S8). This shows that these transcription factors have a general role in biomass production, which is significantly exacerbated by the elevation of CO_2. In *edf3*, the biomass under aCO_2 condition biomass remains unaffected in comparison to the WT (Table S8). Therefore, the strong reduction of biomass observed in *edf3* under eCO_2 was strictly associated with a defect of the stimulation of growth induced by high CO_2. This shows that EDF3 is a major regulator of the gain of biomass that is observed under eCO_2.

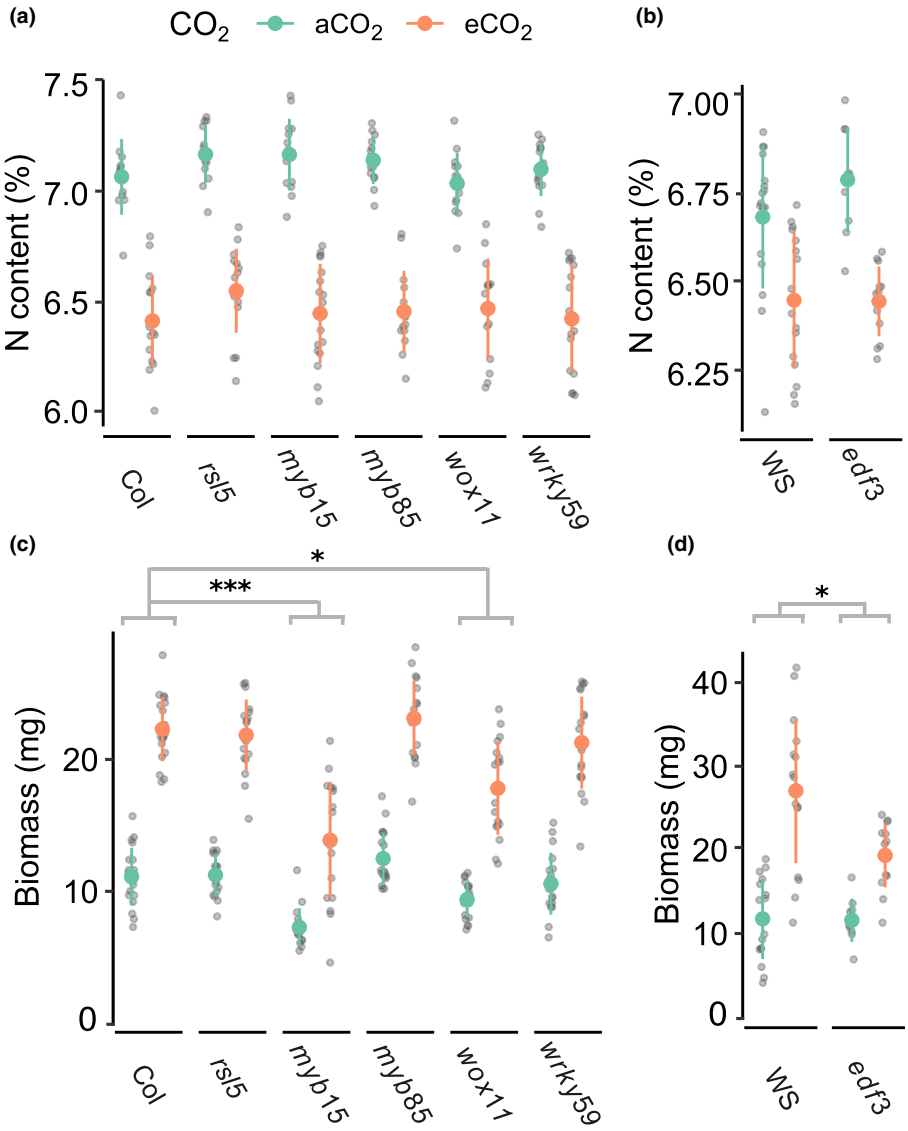


Fig. 7 *MYB15*, *WOX11*, and *EDF3* control the stimulation of biomass production by eCO_2 under nitrate limitation. Phenotypic response to eCO_2 of plants with mutations for the candidate regulators *rsl5*, *myb15*, *myb85*, *wox11*, *wrky59*, and *edf3*, as compared with their relative wild-type (WT; *Col*, Columbia; *WS*, Wassilewskija). Plants were grown for 5 wk in hydroponics under contrasted levels of CO_2. Error bars represent SD. The significance levels refer to the *P*-values relative to the genotype by environment interactions (i.e. the effect of the mutation on the effect of eCO_2), using a linear model (*, $P \leq 0.05$; ***, $P \leq 0.001$). (a, b) Leaf N concentration measured in WT and in candidate mutant lines. (c, d) Shoot biomass measured in WT and in the candidate mutant lines. aCO_2, ambient CO_2; eCO_2, 900 ppm.

Discussion

In this study, we analyzed the effect of eCO_2 on plant growth and on nitrate and Fe nutrition in Arabidopsis. Several studies have explored the effect of eCO_2 on plant genome expression. Most of them concluded to a minor effect of eCO_2 on genome expression, with a relatively limited number of differentially expressed genes (Taylor *et al.*, 2005; Ainsworth *et al.*, 2006; Fukayama *et al.*, 2009; Tallis *et al.*, 2010; Jauregui *et al.*, 2015; Bencke-Malato *et al.*, 2019). Our results show that eCO_2 actually leads to modest changes in gene expression under sufficient nutrient conditions, but in opposition leads to a profound reprogramming of genome expression as soon as plants grow under limiting nitrate or under Fe starvation condition. More precisely, we demonstrated that eCO_2 has a strong effect on the expression of genes associated with nitrate or Fe signaling. A striking contrast was observed between the repressive effect of eCO_2 on the expression of nitrate transport and assimilation genes such as *NRT2.1*, *NRT1.1*, or *NIR1*, and the opposite inductive effect on the expression of genes such as *BT1*/*BT2* or *NIGT* transcription factors, known to be among the main negative regulators of nitrate transport and assimilation (Araus *et al.*, 2016; Kiba *et al.*, 2018). The same observation made on Fe signaling genes led us to postulate that eCO_2 may broadly affect the response to nutrient limitation at the gene expression level, by altering the expression of major signaling modules.

Concerning nitrate, our results provide highly consistent lines of evidence supporting the original hypothesis that the repressive effect of eCO_2 on root nitrate uptake specifically targets the high-affinity transport system (HATS), and not the low-affinity transport system (LATS). Indeed, the nitrate transport genes shown to be markedly downregulated by eCO_2 (*NRT2.1*, *NAR2.1*, *NRT1.1*) all encode major contributors to the HATS activity (O'Brien *et al.*, 2016; Jacquot *et al.*, 2020). In line with this, root nitrate uptake rate is repressed by eCO_2 at 0.5 mM external nitrate (at which the HATS is largely predominant over the LATS), but not at 10 mM external nitrate (at which the HATS plays a negligible role as compared to the LATS). Accordingly, total N concentration in shoots is reduced by eCO_2 at 0.5 mM, but not at 10 mM nitrate. To our knowledge, such a differential response of the nitrate HATS and LATS to eCO_2 was not previously reported, and may provide some explanation to contradictory conclusions found in the literature. Indeed, the response of root nitrate acquisition to eCO_2 was found to be highly variable, ranging from strong inhibition to no effect or even stimulation, therefore preventing any clear conclusion on this point (BassiriRad *et al.*, 2001; Coskun *et al.*, 2016; Gojon *et al.*, 2022). However, most studies did not precisely monitor external nitrate availability, leaving the possibility that part of the above discrepancies may result from the fact that the plants in these studies either relied on the HATS or the LATS for their N nutrition. In opposition to root nitrate uptake, the response of shoot growth to eCO_2 was similar regardless of nutrient supply, with a large stimulation in all cases. Our observations show that the responses of the growth and mineral composition of shoots exposed to eCO_2 are not strictly correlated, which is evidence against the hypothesis of a growth dilution of plant N under eCO_2. This supports recent reports (Myers *et al.*, 2014; Feng *et al.*, 2015; Wujeska-Klause *et al.*, 2019), and thus suggests direct negative effects of eCO_2 on plant N acquisition.

We used a machine-learning approach to infer the GRN in the roots of Arabidopsis in response to eCO_2 under nitrate limitation, and to identify putative regulatory factors involved in the plant's response to eCO_2. This revealed that large communities of genes with a function associated with the response to nitrate or to N were affected by the growth under eCO_2, suggesting again that rising CO_2 in the atmosphere will markedly alter the physiological mechanisms of plant nutrition. In addition, ranking the regulators by their degree of connection to target genes in the GRN led to the identification of potential actors of the response to eCO_2 in plants. By phenotyping loss-of-function mutants for a subset of these highly connected regulators, we found that three of them, *MYB15*, *WOX11*, and *EDF3*, were involved in the stimulation of growth by eCO_2. To date, only a handful of genes involved in the stimulation of growth by eCO_2 have been identified (Bouain *et al.*, 2022; Oguchi *et al.*, 2022). To our knowledge, our work provides the first identification of root related genes involved in the regulation of the stimulation of plant growth by eCO_2. We have demonstrated here that *MYB15*, *WOX11*, and *EDF3* function are essential to fully reach the potential of eCO_2 fertilization under nitrate limitation, with a reduction in the growth stimulation by almost half in *myb15* or *edf3* mutants. *MYB15* has been previously identified as a regulator of the expression of the *PHO1;H3* phosphate transporter in the roots (Pal *et al.*, 2017). Interestingly, phosphate accumulation in the shoot has been recently proposed to regulate plant growth under eCO_2, through the expression of the *PHT4;3* phosphate transporter in shoots (Bouain *et al.*, 2022). In our data, the expression of these two phosphate transporters was not deregulated by eCO_2. Two other phosphate transporter genes were actually deregulated by eCO_2, but were not predicted as *MYB15* target genes in the GRN. Nevertheless, the link between the regulation of growth under eCO_2 by *MYB15* and phosphate accumulation will deserve further attention. Much less is known about the function of *EDF3*, apart from its role in the transcriptional network of N-associated growth inferred in Arabidopsis, were *EDF3* regulates essential N-associated genes such as *NIR1*, *NAR2.1*, or *GLN1.2* (Gaudinier *et al.*, 2018). Finally, the absence of variation in the decrease in N concentration led by the growth under eCO_2 in *myb15* or *edf3* mutants supports the physiological observations earlier mentioned, both strongly suggesting a decoupling between the growth stimulation and the decrease in N concentration established under eCO_2, at the physiological and genetic levels. Altogether, the identification of these genes as important components of the response to eCO_2 paves the way for the optimization of plant biomass production under future CO_2 atmosphere without penalizing nutrient content, and nutritional value of crops.

Acknowledgements

This work was supported by the I-Site Montpellier Université d'Excellence (MUSE; project ECO2TREATS), the CNRS

through the Mission for Transversal and Interdisciplinary Initiatives (MITI) 80 PRIME program, and the Biologie et Amélioration des Plantes (BAP) department of the INRAE. OC was recipient of a PhD fellowship from the CNRS. We acknowledge Hugues Baudot for the implementation and the monitoring of growth chambers with CO_2 enrichment at IPSiM. We thank all members of our groups for constructive discussion that contributed to this work.

Competing interests

None declared.

Author contributions

CD, AG, and AM designed the project. OC, L-LP, CD, AG, LB, and AM planned the experiments. OC and L-LP performed most of the experiments. OC, CD, AG, LB, SL, and AM analyzed and interpreted the data. OC and AM wrote the manuscript with help from every authors.

ORCID

Liên Bach https://orcid.org/0000-0003-4614-3800
Océane Cassan https://orcid.org/0000-0002-4595-2457
Christian Dubos https://orcid.org/0000-0001-5486-3643
Alain Gojon https://orcid.org/0000-0001-5412-8606
Sophie Lèbre https://orcid.org/0000-0003-3444-2416
Antoine Martin https://orcid.org/0000-0002-6956-2904

Data availability

Data and R notebooks containing the analyses performed in this article can be found at https://github.com/OceaneCsn/CO2_root_networks_inference. RNA-seq data are available at https://www.ebi.ac.uk/biostudies/arrayexpress/studies using the accession no.: E-MTAB-12483.

References

Ainsworth EA, Long SP. 2021. 30 years of free-air carbon dioxide enrichment (FACE): what have we learned about future crop productivity and its potential for adaptation? *Global Change Biology* **27**: 27–49.

Ainsworth EA, Rogers A, Vodkin LO, Walter A, Schurr U. 2006. The effects of elevated CO_2 concentration on soybean gene expression. An analysis of growing and mature leaves. *Plant Physiology* **142**: 135–147.

Araus V, Vidal EA, Puelma T, Alamos S, Mieulet D, Guiderdoni E, Gutierrez RA. 2016. Members of BTB gene family of scaffold proteins suppress nitrate uptake and nitrogen use efficiency. *Plant Physiology* **171**: 1523–1532.

BassiriRad H, Gutschick VP, Lussenhop J. 2001. Root system adjustments: regulation of plant nutrient uptake and growth responses to elevated CO_2. *Oecologia* **126**: 305–320.

Bencke-Malato M, De Souza AP, Ribeiro-Alves M, Schmitz JF, Buckeridge MS, Alves-Ferreira M. 2019. Short-term responses of soybean roots to individual and combinatorial effects of elevated [CO_2] and water deficit. *Plant Science* **280**: 283–296.

Bloom A, Burger M, A. Kimball B, J. Pinter JP. 2014. Nitrate assimilation is inhibited by elevated CO_2 in field-grown wheat. *Nature Climate Change* **4**: 477–480.

Bloom AJ, Burger M, Rubio Asensio JS, Cousins AB. 2010. Carbon dioxide enrichment inhibits nitrate assimilation in wheat and Arabidopsis. *Science* **328**: 899–903.

Bouain N, Cho H, Sandhu J, Tuiwong P, Prom UTC, Zheng L, Shahzad Z, Rouached H. 2022. Plant growth stimulation by high CO_2 depends on phosphorus homeostasis in chloroplasts. *Current Biology* **32**: 4493–4500.

Brooks MD, Juang CL, Katari MS, Alvarez JM, Pasquino A, Shih HJ, Huang J, Shanks C, Cirrone J, Coruzzi GM. 2021. ConnecTF: a platform to integrate transcription factor–gene interactions and validate regulatory networks. *Plant Physiology* **185**: 49–66.

Cassan O, Lebre S, Martin A. 2021. Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. *BMC Genomics* **22**: 387.

Coskun D, Britto DT, Kronzucker HJ. 2016. Nutrient constraints on terrestrial carbon fixation: the role of nitrogen. *Journal of Plant Physiology* **203**: 95–109.

Feng Z, Rutting T, Pleijel H, Wallin G, Reich PB, Kammann CI, Newton PC, Kobayashi K, Luo Y, Uddling J. 2015. Constraints to nitrogen acquisition of terrestrial plants under elevated CO_2. *Global Change Biology* **21**: 3152–3168.

Fukayama H, Fukuda T, Masumoto C, Taniguchi Y, Sakai H, Cheng W, Hasegawa T, Miyao M. 2009. Rice plant response to long term CO_2 enrichment: gene expression profiling. *Plant Science* **177**: 203–210.

Gansel X, Munos S, Tillard P, Gojon A. 2001. Differential regulation of the NO_3^- and NH_4^+ transporter genes *AtNrt2.1* and *AtAmt1.1* in *Arabidopsis*: relation with long-distance and local controls by N status of the plant. *The Plant Journal* **26**: 143–155.

Gaudinier A, Rodriguez-Medina J, Zhang L, Olson A, Liseron-Monfils C, Bågman A-M, Foret J, Abbitt S, Tang M, Li B *et al.* 2018. Transcriptional regulation of nitrogen-associated metabolism and growth. *Nature* **563**: 259–264.

Gojon A, Cassan O, Bach L, Lejay L, Martin A. 2022. The decline of plant mineral nutrition under rising CO_2: physiological and molecular aspects of a bad deal. *Trends in Plant Science* **28**: 185–198.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5**: e12776.

IPCC. 2021. Climate change 2021: the physical science basis. In: Masson-Delmotte V, Zhai P, Pirani A, Connors SL, Péan C, Berger S, Caud N, Chen Y, Goldfarb L, Gomis MI *et al.*, eds. *Contribution of working group I to the sixth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY, USA: Cambridge University Press.

Jacquot A, Chaput V, Mauries A, Li Z, Tillard P, Fizames C, Bonillo P, Bellegarde F, Laugier E, Santoni V *et al.* 2020. NRT2.1 C-terminus phosphorylation prevents root high affinity nitrate uptake activity in *Arabidopsis thaliana*. *New Phytologist* **228**: 1038–1054.

Jauregui I, Aparicio-Tejo PM, Avila C, Rueda-Lopez M, Aranjuelo I. 2015. Root and shoot performance of *Arabidopsis thaliana* exposed to elevated CO_2: a physiologic, metabolic and transcriptomic response. *Journal of Plant Physiology* **189**: 65–76.

Kiba T, Inaba J, Kudo T, Ueda N, Konishi M, Mitsuda N, Takiguchi Y, Kondou Y, Yoshizumi T, Ohme-Takagi M *et al.* 2018. Repression of nitrogen starvation responses by members of the Arabidopsis GARP-type transcription factor NIGT1/HRS1 subfamily. *Plant Cell* **30**: 925–945.

Li P, Ainsworth EA, Leakey AD, Ulanov A, Lozovaya V, Ort DR, Bohnert HJ. 2008. Arabidopsis transcript and metabolite profiles: ecotype-specific responses to open-air elevated [CO_2]. *Plant, Cell & Environment* **31**: 1673–1687.

Li P, Sioson A, Mane SP, Ulanov A, Grothaus G, Heath LS, Murali TM, Bohnert HJ, Grene R. 2006. Response diversity of *Arabidopsis thaliana* ecotypes in elevated [CO_2] in the field. *Plant Molecular Biology* **62**: 593–609.

Loladze I. 2014. Hidden shift of the ionome of plants exposed to elevated CO_2 depletes minerals at the base of human nutrition. *eLife* **3**: e02245.

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Consortium D, Kellis M, Collins JJ *et al.* 2012. Wisdom of crowds for robust gene network inference. *Nature Methods* **9**: 796–804.

Miyazaki S, Fredricksen M, Hollis KC, Poroyko V, Shepley D, Galbraith DW, Long SP, Bohnert HJ. 2004. Transcript expression profiles of *Arabidopsis thaliana* grown under controlled conditions and open-air elevated concentrations of CO_2 and of O_3. *Field Crops Research* **90**: 47–59.

Myers SS, Zanobetti A, Kloog I, Huybers P, Leakey AD, Bloom AJ, Carlisle E, Dietterich LH, Fitzgerald G, Hasegawa T *et al.* 2014. Increasing CO_2 threatens human nutrition. *Nature* **510**: 139–142.

O'Brien JA, Vega A, Bouguyon E, Krouk G, Gojon A, Coruzzi G, Gutierrez RA. 2016. Nitrate transport, sensing, and responses in plants. *Molecular Plant* **9**: 837–856.

Oguchi R, Hanada K, Shimizu M, Mishio M, Ozaki H, Hikosaka K. 2022. Enhanced growth rate under elevated CO_2 conditions was observed for transgenic lines of genes identified by intraspecific variation analyses in *Arabidopsis thaliana*. *Plant Molecular Biology* **110**: 333–345.

Pal S, Kisko M, Dubos C, Lacombe B, Berthomieu P, Krouk G, Rouached H. 2017. TransDetect identifies a new regulatory module controlling phosphate accumulation. *Plant Physiology* **175**: 916–926.

Rau A, Maugis-Rabusseau C. 2018. Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics* **19**: 425–436.

Rubin G, Tohge T, Matsuda F, Saito K, Scheible WR. 2009. Members of the LBD family of transcription factors repress anthocyanin synthesis and affect additional nitrogen responses in Arabidopsis. *Plant Cell* **21**: 3567–3584.

Smith MR, Myers SS. 2018. Impact of anthropogenic CO_2 emissions on global human nutrition. *Nature Climate Change* **8**: 834–839.

Sun P, Isner JC, Coupel-Ledru A, Zhang Q, Pridgeon AJ, He Y, Menguer PK, Miller AJ, Sanders D, McGrath SP *et al.* 2022. Countering elevated CO_2 induced Fe and Zn reduction in Arabidopsis seeds. *New Phytologist* **235**: 1796–1806.

Tallis MJ, Lin Y, Rogers A, Zhang J, Street NR, Miglietta F, Karnosky DF, De Angelis P, Calfapietra C, Taylor G. 2010. The transcriptome of *Populus* in elevated CO reveals increased anthocyanin biosynthesis during delayed autumnal senescence. *New Phytologist* **186**: 415–428.

Tausz-Posch S, Tausz M, Bourgault M. 2020. Elevated [CO_2] effects on crops: advances in understanding acclimation, nitrogen dynamics and interactions with drought and other organisms. *Plant Biology* **22**: 38–51.

Taylor G, Street NR, Tricker PJ, Sjodin A, Graham L, Skogstrom O, Calfapietra C, Scarascia-Mugnozza G, Jansson S. 2005. The transcriptome of *Populus* in elevated CO_2. *New Phytologist* **167**: 143–154.

Thompson M, Gamage D, Hirotsu N, Martin A, Seneweera S. 2017. Effects of elevated carbon dioxide on photosynthesis and carbon partitioning: a perspective on root sugar sensing and hormonal crosstalk. *Frontiers in Physiology* **8**: 578.

Vicente R, Bolger AM, Martinez-Carrasco R, Perez P, Gutierrez E, Usadel B, Morcuende R. 2019. *De novo* transcriptome analysis of durum wheat flag leaves provides new insights into the regulatory response to elevated CO_2 and high temperature. *Frontiers in Plant Science* **10**: 1605.

Vicente R, Perez P, Martinez-Carrasco R, Feil R, Lunn JE, Watanabe M, Arrivault S, Stitt M, Hoefgen R, Morcuende R. 2016. Metabolic and transcriptional analysis of durum wheat responses to elevated CO_2 at low and high nitrate supply. *Plant & Cell Physiology* **57**: 2133–2146.

Vicente R, Pérez P, Martínez-Carrasco R, Gutiérrez E, Morcuende R. 2015. Nitrate supply and plant development influence nitrogen uptake and allocation under elevated CO_2 in durum wheat grown hydroponically. *Acta Physiologiae Plantarum* **37**: 114.

Wujeska-Klause A, Crous KY, Ghannoum O, Ellsworth DS. 2019. Lower photorespiration in elevated CO_2 reduces leaf N concentrations in mature Eucalyptus trees in the field. *Global Change Biology* **25**: 1282–1295.

Yang A, Li Q, Chen L, Zhang W-H. 2020. A rice small GTPase, Rab6a, is involved in the regulation of grain yield and iron nutrition in response to CO_2 enrichment. *Journal of Experimental Botany* **71**: 5680–5688.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Metrics of the GRN inferred from the effect of elevated CO_2 under low nitrate condition.

Fig. S2 Validation of edges in the GRN.

Fig. S3 Statistical significance of the comparison between validated edges in the GRN and in random networks made by permutation from DEGs included in the GRN.

Fig. S4 Gene ontology enrichment in the communities of the GRN.

Table S1 List and sequence of primers used in this study.

Table S2 Differentially expressed genes by eCO_2 under high nitrate and iron starvation.

Table S3 Differentially expressed genes by eCO_2 under high nitrate and iron supply.

Table S4 Differentially expressed genes by eCO_2 under low nitrate and iron starvation.

Table S5 Differentially expressed genes by eCO_2 under low nitrate and iron supply.

Table S6 List of regulators used in the inference of GRN.

Table S7 Full ranking of regulators from their degree in the GRN.

Table S8 Statistics of $eCO_2 \times$ genotype effects using a linear model.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

2.3 Integration of transcription factor binding sites to gene expression data improves regression-based Gene Regulatory Network inference in *Arabidopsis thaliana*

2.3.1 Preamble

After the identification of candidate genes under a combinatorial design based on expression measures, we reflected on possible improvements of our GRN inference strategy. A first research direction that came to mind was the exploration of other regression functions. We were especially interested in exploring different complexities of regression models to test weather linearity assumptions or the lack of interaction effects would still provide valuable GRN models as compared to those provided by Random Forests. Second, we wished to complement expression data with another kind of information describing gene regulation by integrating gene expression with sequence data. We thus implemented two GRN inference methods based on sparse linear regression and Random Forests making use of gene expression data and guided by the presence of TFBSs of regulators in the promoters of their target. Publication #4 describes those methods in detail, our validation procedures, and their application to a transcriptomic dataset of the root response to nitrate induction in *Arabidopsis thaliana*.

2.3.2 Publication #4 (In preparation)

Note : This section has its own reference system. Citation numbers refer to bibliography items included in the present article, and not at the end of the PhD manuscript. ***This article is in preparation, and is representative of the results in october, 2022. Before submission in 2023, changes to the methods and results are planned by the authors. In particular, those additions will improve the inference pipelines and provide a better understanding of the gene-specific added value of data integration.*** *Code and data to reproduce the results presented in this section are available in the github repository* `https://github.com/OceaneCsn/integrative_GRN_N_induction`

Unpublished manuscript, 2022
PhD thesis contribution article

Integration of Transcription Factor Binding Sites to gene expression data improves regression-based Gene Regulatory Network inference in *Arabidopsis thaliana*

Océane Cassan[1], Charles Lecellier [2], Laurent Bréhélin[3], Antoine Martin[1], Sophie Lèbre[4,5]

Abstract

Gene Regulatory Networks (GRNs) are abstract models of the control of transcription. In complex organisms, the inference of GRNs is an ultimate goal but also an unsolved challenge because it involves multiple intertwined molecular layers. Regression based-methods are popular and powerful statistical approaches traditionally applied to expression data. More recently, integrative regression-based strategies have emerged to guide GRN inference with complementary data, but could be extended to fit the data more accurately or model causality. In addition, it is possible that depending on the nature of the regression model the advantages of data integration advantages will vary.
Based on the temporal response to nitrate induction in the roots of *Arabidopsis thaliana*, we propose to jointly study the impact of model choice for the regression function in GRN inference, and the benefits of integrating Transcription Factor Binding Sites (TFBS) information to expression data. To do so, we improve upon two promising regression strategies: an integrative version of the Random Forest algorithm (bRF), and a LASSO generalized linear model with differential shrinkage and stability selection (LASSO-D3S). We benchmark their predictive capabilities, and accuracy against experimentally validated Transcription Factor(TF)-target interactions. This evaluation is carried out for a range of biologically relevant network densities and through a parameter finely tuning the contribution of TFBS to GRN inference.
We conclude that TFBS integration improves the biological relevance of inferred GRNs for both bRF and LASSO-D3S, and we discuss the features of inferred GRNs depending on model choice. In addition, pathways relevant to nitrate nutrition expected in this response are realistically modeled by bRF and LASSO-D3S, and functionally validated regulations of the nitrate transporter *NRT2.1* are progressively retrieved as TFBS contribution is strengthened. This outlines the importance of further developments in multi-omics data integration for regression-based GRN inference. All R scripts for the bRF and LASSO-D3S functions are made available.

Keywords
Data integration — Feature selection — Random forest — LASSO — Differential shrinkage — Stability selection — Gene Regulatory Network inference — Arabidopsis thaliana

[1] *BPMP, CNRS, INRAE, Institut Agro, Univ Montpellier, 34060, Montpellier, France*
[2] *IGMM, Univ. Montpellier, CNRS, Montpellier, France*
[3] *LIRMM, Univ. Montpellier, CNRS, Montpellier, France*
[4] *IMAG, Univ. Montpellier, CNRS, Montpellier, France*
[5] *Université Paul-Valéry-Montpellier 3, Montpellier, France*

Introduction

Principles of GRN Inference from transcriptomic data
Gene Regulatory Network (GRN) inference has the objective of deciphering the relationships between genes in the context of transcription. Statistical inference methods usually leverage high-throughput genomics to reconstruct those networks, in which nodes represent genes, and edges represent a relation of regulation between those genes. Because transcriptomic data are increasingly common and less costly, they are the input of choice for most statistical approaches to GRN inference. Gene expression profiles across environmental conditions or developmental stages can be interrogated to infer links between genes influencing each other.

In order to model regulation, it is possible to infer oriented edges from regulator genes to other genes. Regression-based inference often decomposes the problem of network infer-

ence into as many regression problems as there are genes. In those regressions, the response variable, e.g. the expression of a target gene, is approximated by a function of the expression levels of the regulator genes. Regression-based techniques mainly differ in their choice of regression function to link the expression of a target gene to the expression of its regulators. For example, TIGRESS [Haury et al., 2012] or LASSO [Tibshirani, 1996, Tjärnberg et al., 2013] techniques chose a linear parametric models for this task, while GENIE3 [Huynh-Thu et al., 2010] and inspired works [Geurts et al., 2018, Cirrone et al., 2020] model non-linear relations via Random Forests (RFs) or more broadly, ensembles of trees or predictors. Another advantage of RFs is that they do not make any assumption on the distribution of the input data.

Once regression models are fit, they allow the extraction of the influence of each regulator over each gene. This is usually followed by the ranking of all possible regulator-gene interactions on the basis of this influence score, and then by the selection of the strongest pairs to build a final network. Biological networks such as GRNs exhibit standard topological features, such as low densities [Koutrouli et al., 2020, Leclerc, 2008, Hayes et al., 2013]. Under this constraint, only a restricted set of regulators usually regulate a target gene [Campos and Freyre-González, 2019]: the notion of sparsity is crucial to the accuracy and interpretability of models describing the control of regulation. In models such as RFs, a threshold on the influence metric can be defined so that the final network contains a number of interactions providing a biologically relevant network density. In simple or generalized linear models, the high dimensional setting (when the number of expression measure per gene far exceeds the number of candidate regulators), also requires some form of regularization for feature selection. In the case of TIGRESS, Least Angle Regression is used to select the most influential regulators, while penalizations such as the LASSO, ridge or elastic-net are popular choices to regularize network inference [Qin et al., 2014, Miraldi et al., 2019].

Integration of other omics Given the underdetermined nature of GRN inference from expression alone, using additional sources of data can guide the choice between several models explaining expression data equally well. Binding experiments, protein-protein interactions, chromatin accessibility or regulatory sequence information are very valuable in the description of transcriptional events. They are however costly, often limited to a small number of TFs or scarce, especially when one is interested in rare environmental conditions or in non-model organisms. This is why in the majority of cases, data-driven methods can hardly rely on these omics alone for a genome-wide inference task. They have, however, already been used to guide network inference in combination with expression data. For example, GRACE [Banf and Rhee, 2017] and SCENIC [Aibar et al., 2017] methods proceed in two steps. First, a GRN is inferred via RFs on transcriptomic data and then refined by additional knowledge about TF binding motifs. In other works, the complementary sources of data are used to build a consensus. IGRN [Clercq et al., 2021] combines several types of input networks, obtained from expression data, CHIP-Seq, chromatin accessibility and regulatory sequences into one predictive regulatory network, trained on interactions documented in the literature.

Integrative inference, *i.e.* including prior data in the GRN estimation process, has already been adapted for different regression-based methods, either linear or non-linear. In one of the first attempts, a linear system of gene regulation was solved globally with several types of regularization including the LASSO [Qin et al., 2014]. Data integration was realised with CHIP-X priors being integrated at the solution initialisation step. More recently, the latest version of the Inferelator [Gibbs et al., 2022], a suite for GRN inference, includes three methods for GRN inference based on regression and with prior incorporation. Those three methods are BBSR, a bayesian approach [Greenfield et al., 2013], AMuSR, a multitask approach [Castro et al., 2019], and LASSO-StARS, a LASSO approach [Miraldi et al., 2019]. They integrate prior knowledge in the Inferelator by using TF activities (TFAs) as predictors.

In LASSO-StARS [Miraldi et al., 2019] prior integration is not only possible through the use of TFAs, but also via modulating the penalty strength for each TF during the estimation of the LASSO models. In this setting, regulators with prior support are less penalized than the others, so that they are favorably included in the final GRN. This modulation of the penalty strength will be further referred to as 'differential shrinkage'. LASSO-StARS optimizes a model penalized by differential shrinkage using using StARS [Liu et al., 2010], a stability selection method that finds the sparsest network while guarantying edges with an acceptable robustness to sub-sampling. The idea of encoding prior knowledge via differential shrinkage was already proposed in previous works on gene regulation [Christley et al., 2009, Studham et al., 2014]. Those two works, however, do not infer oriented edges between regulators and targets, but can link all genes to each other, which is not ideal for GRN modelling. Besides, none of the methods employing differential shrinkage in LASSO regression, including the Inferelator, model RNA-Seq data as count data, which seems to be a limitation. Indeed, many existing tool that work with RNA-Seq data to perform genome-wide statistical analyses such as differential expression or co-expression clustering successfully modeled RNA-Seq data with Poisson negative-binomial distributions [McCarthy et al., 2012, Rau and Maugis-Rabusseau, 2018].

Regarding non-linear regression, iRafNet [Petralia et al., 2015] introduced a way to encode prior information into RFs, such as knock-out experiments, protein-protein interactions, or time series gene-expression. In iRafNet, regression trees predicting the target gene expression from the expression of other genes are elongated in a way that increases the chance of predictor genes supported by prior knowledge to get chosen in each decision nodes. This has the effect of inflating the importance metric of interactions supported by the chosen prior. It

was further adapted to time series expression data in the Outpredict method [Cirrone et al., 2020]. However, iRafNet does not restrict predictive variables in the regressions to regulator genes only, which makes its outputs closer to co-expression network than GRNs. In addition, it is limited to the node purity importance metric to estimate the influence of the regulators. This metric was proven to be ill-designed to interpret variable importance when variables are dependent and are interacting [Scornet, 2020, Nicodemus and Malley, 2009]. Other importance metrics less vulnerable to this issue and more robust to over-fitting exist, namely the Mean Decrease Accuracy (MDA) [Breiman, 2001], but are not implemented in GENIE3 nor iRafNet.

Objectives Regression-based GRN inference techniques historically estimated regulatory links between regulators and target genes solely on the basis of their expression profiles. We propose to extend them in the direction of attributing stronger regulatory influences to interactions in which the TFBS of the regulator is known to be present in the promoter of its target. Indeed, they are largely available in model organisms in the form of Position Weight Matrices (PWMs). They have the advantage of being computationally searchable in the regulatory regions of a genome for a very low cost, and do not require additional experimental work. Besides, this is a valuable source of information because it is relative to physical interactions between TFs and DNA, which is complementary to expression data. This has the potential to shed light on the complexity of regulation, in the course of which binding and regulation of expression often co-occur. Combining binding motifs and expression has already been proven useful for network inference, especially in complex organisms [Marbach et al., 2012, Kundaje et al., 2007, Gibbs et al., 2022].

During data integration, it is reasonable to believe that, depending on the characteristics of the statistical models like linearity or their parametric nature, the performance and benefits of data integration will vary. In this work, we jointly study the impact of model choice for the regression function in GRN inference, and the benefit of integrating TFBS information to expression data. More precisely, we propose to improve two existing GRN inference methods for integrative regressions in the linear and non-linear cases.

1. An integrative version of the Random Forest algorithm, similar to iRafNet [Petralia et al., 2015], but adapted toward GRN inference and implementing the MDA. This method will be referred to as **biased Random Forests: bRF**
2. A LASSO-penalized generalized linear model with Stability Selection, in which we apply differential shrinkage to encode prior knowledge. This is similar to LASSO-StARS [Miraldi et al., 2019], but with major differences: we model RNA-Seq data as count data, we do not choose the same weighting of regulatory edges, we do not use the same Stability Selection approach, and we propose the first R implementation of this method. This method will be referred to as **LASSO with Differential Shrinkage and Stability Selection: LASSO-D3S**

This work leverages bRF and LASSO-D3S to model the response to nitrate (N) induction in the model plant *Arabidopsis thaliana* as a GRN. We investigate the aspects and performance of TFBS integration on this real dataset for different parametrizations of bRF and LASSO-D3S, a range of biologically relevant densities, and through a parameter gradually influencing the contribution of TFBSs to GRN inference, α. To do so, we evaluate the predictive abilities of the inferred GRNs against experimental databases of regulatory interactions, and state of the art knowledge about nitrate nutrition pathways.

We conclude that TFBS integration improves the biological relevance of inferred GRNs for both bRF and LASSO-D3S in benchmarks against experimental gold standards, and that they realistically model nitrate nutrition pathways in this case study.

Methods

Datasets for GRN inference

Binding motifs dataset

The TFBSs, encoded by PWMs, were retrieved from the JASPAR database [Castro-Mondragon et al., 2021] and The Plant Cistrome Database [O'Malley et al., 2016]. Among the 2547 known regulators in Arabidopsis, a collection of 631 non redundant PWMs was formed by the union of these two sources.

We define the promoter region of Arabidopsis genes as the sequence from -1000 bp to +200 around the Transcription Start Site (TSS), as this interval has been estimated to contain 86% of binding sites in plants [Yu et al., 2016]. The TSS of genes were defined as the start of the messenger RNA in the TAIR10 GFF reference annotation, and there are in total 31824 promoter regions in Arabidopsis.

Among the 201 nitrate-responsive regulators, 70 regulators were associated to a known PWM. All the promoters were scanned to find occurrences of the 631 PWMs using FIMO [Grant et al., 2011]. FIMO reported all occurrences of a PWM in a given promoter if its p-value was under the default value of $1e^{-4}$. In the case of several significant occurrences, the one with minimum p-value was kept. Descriptive statistics of the TFBS search results, forming a TFBS network between PWMs and promoters, are presented in Figure 1.

Gene expression normalization

The RNA-Seq raw counts were normalized via the TMM method [Robinson and Oshlack, 2010], and lowly expressed genes were removed prior to differential expression analysis.

Validation data and metrics

Validation data To evaluate the correctness of the inferred networks from both methods, we interrogated the ConnecTF [Brooks et al., 2020] database. ConnecTF gathers experimental regulatory interactions for a large number of TFs:

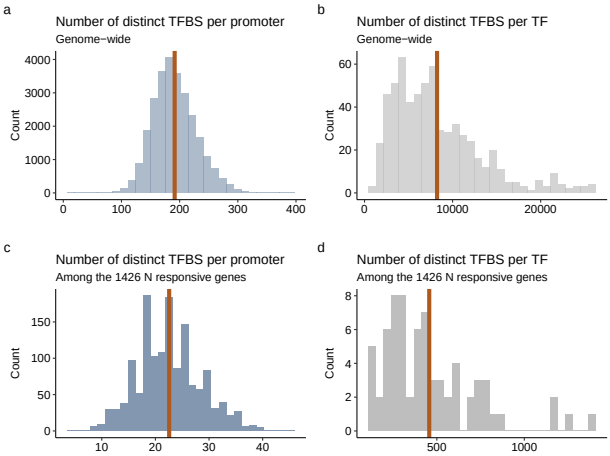


Figure 1. Distributions of the number of distinct PWM hits per promoter, and the number of distinct promoter hits per PWM. Subplots a and b are genome-wide (31824 promoters and 631 PWMs), while subplots c and d are restricted to nitrate-responsive genes (1426 promoters and 70 PWMs). The average value of each distribution is overlaid in orange.

- Direct regulation in Arabidopsis, via the TARGET assay. TARGET measures the gene expression changes after importing a specific TF from the cytoplasm into the nucleus of modified cells [Bargmann et al., 2013].
- *In vivo* binding via CHIP-Seq experiments.
- *In vitro* binding via a DAP-Seq assay [O'Malley et al., 2016].

In total, 60 TFs of the 201 nitrate-responsive TFs are studied in at least one type of validation (Supp. Table S1).

Validation metrics In this study, we consider four criteria to evaluate the relevance of the inferred oriented interactions $\mathscr{E}$. Such interactions can be either from a regulator to a target gene, but also to another regulator.

Two of them rely on the experimental data contained in ConnecTF. Here, we denote the set of validated regulatory interactions in ConnecTF as $\mathscr{C}$.

1. **Precision** is the percentage of edges validated by at least one source of validation in ConnecTF. To compute precision, only the interactions involving one of the 70 TFs studied in ConnecTF are considered, forming the new set of edges $\mathscr{E}'$, as the other interactions can be neither confirmed nor falsified.

$$\text{Precision} = \frac{|\mathscr{E}' \cap \mathscr{C}|}{|\mathscr{E}'|}$$

2. **Recall** is the percentage of ConnecTF interactions retrieved by GRN inference. The validation set $\mathscr{C}$ is restricted to interactions involving on both side genes from the 1426 input nitrate-responsive genes, forming the new set of validated edges $\mathscr{C}'$.

$$\text{Recall} = \frac{|\mathscr{E}' \cap \mathscr{C}'|}{|\mathscr{C}'|}$$

To assess the performance of GRN inference, precision and recall can be computed for all possible density values, ranging from no edge to the entire weighted GRN of size $R(T-1)$. The area under the curve formed by all possible density thresholds is the Area Under the Precision and Recall curve (AUPR). It is common to use this curve for GRN inference benchmark. However, we did not use this metric because LASSO-D3S inference involves feature selection, and is thus not capable of providing a weight for all regulator-target pairs. It only permits to weight the regulators selected for a target, and gives the same score to all non-selected regulators. Instead of relying on an incomplete AUPR for one of the methods, we rather explore a range of common and biologically relevant densities (from 0.005 to 0.1) for which we measure precision and recall.

The other validation metrics we employ are statistics about the efficiency of data integration, and the predictive capabilities of GRN models:

1. The **PWM support** of an inferred network controls the outcome of integrative GRN inference by measuring the average binding site scores of the predicted edges:

$$\frac{\Sigma_{(t,r)\in\mathscr{E}}\Pi_{r,t}}{|\mathscr{E}|}$$

2. **The ability of an inferred GRN to accurately predict the target genes expression** using the selected regulators as predictors on test data. To evaluate bRF inference for a target gene t, we fit a RF with the incoming regulators of t in the GRN as predictive variables. MSE_t is computed as the mean MSE of this RF model on OOB examples, i.e the experimental conditions not seen while adjusting the trees. To evaluate LASSO-D3S inference for a target gene t, we fit a Poisson regression with the incoming regulators of t in the GRN as predictive variables. MSE_t is computed as the mean MSE in the test folds of a 5-folds cross-validation.

   The global MSE of an inferred GRN is the average MSE for all target genes in the GRN, taken to the log scale.

Results

Integrative GRN inference methods

We give here a detailed overview of two novel integrative GRN inference procedures based on TFBS prior information (Figure 2). The information of TFBS occurrences in the gene promoters was derived from the binding motifs dataset described

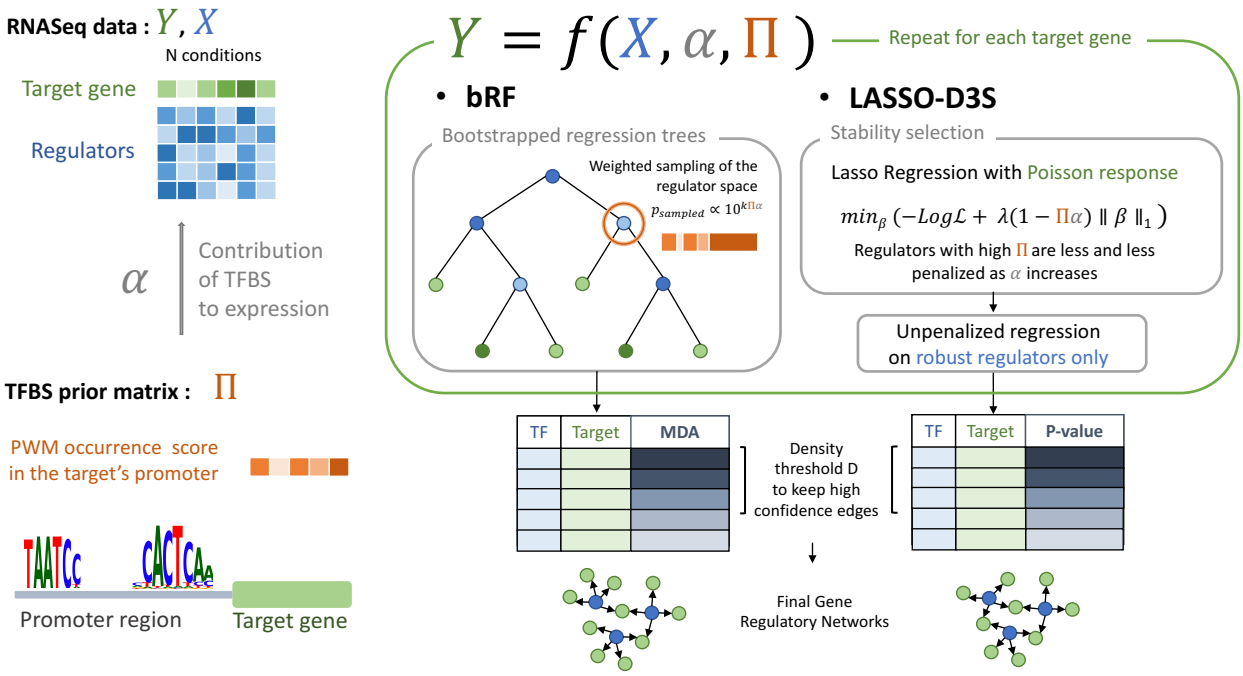


Figure 2. Illustration the bRF and LASSO-D3S integrative GRN inference methods. Input data for a target gene Y is in the form of an expression matrix for the regulators X, and a TFBS scoring matrix Π. Π contains information about the presence or absence of the regulators PWM in target gene's promoter. α is a parameter controlling the force of TFBS integration expression. For each target gene, a regression model is fit to Y with bRF or LASSO-D3S using the expression levels of the regulators as predictive variables, and favoring the attribution of high influences to regulators with their TFBS in the target's promoter. For bRF, this prioritization of provided by a weighted subsampling of the regulators space when elongating the regression trees, while it is achieved by differential shrinkage combined to stability selection is LASSO-D3S. Once all regulator-gene pairs were ranked based on their influence in the regression models, final GRNs are built by selecting the number of strongest interactions providing a desired network density.

in the Methods section. The TFBS prior matrix Π gives, for each regulator-target pair (r,t) a prior value $\Pi_{r,t} \in [0,1]$ defined as:

$$\Pi_{r,t} = \begin{cases} 0 : \text{if the PWM of } r \text{ is not in the promoter of } t \\ 1 : \text{if the PWM of } r \text{ is in the promoter of } t \\ \frac{1}{2} : \text{if the PWM of } r \text{ is missing} \end{cases} \quad (1)$$

Biased Random Forests

Non-linear regressors such as ensembles of regression trees model complex relations between a target gene and the expression of its regulators, and combinatorics of regulators as well. As inspired by iRafNet [Petralia et al., 2015], we model data integration in RF by increasing the use of regulators supported by a binding site in the decision nodes of the regression trees. A biased RF is inferred for each target gene t. At each decision node, the most discriminating regulator is chosen among a subset of $\sqrt{R}$ regulators. This subset, traditionally equiprobably sampled from all the regulators, is submitted here to a weighted sampling where the weights encode prior knowledge about the regulators. With a high weight, a regulator is more likely to get chosen to compete with others among the $\sqrt{R}$ regulators tested to create a decision node. As a result, the importance metric of this regulator is expected to be inflated.

The chance of a regulator r to get picked in the decision node of a regression tree adjusted to the target gene t is the weight $w_{r,t}$. To bias the RFs toward attributing high importances to TFBS-supported variables, we define a weight $w_{r,t}$ that increases with the TFBS prior $\Pi_{r,t}$. No link function was proposed for TFBS prior integration in the form of discrete scores in $[0,1]$ in the iRafNet publication. Defining $w_{r,t}$ through a linear relation with $\Pi_{r,t}$ did not strongly increase the importance of prior-supported regulators in our preliminary experiments (not shown). Consequently, we designed a novel function to link the weight of each TF $w_{r,t}$ during biased subsampling in RFs estimation with their value in the TFBS prior matrix $\Pi_{r,t}$:

$$w_{r,t}^{RF} = 10^{k\Pi_{r,t}\alpha} \quad (2)$$

where α is chosen in $[0,1]$ to tune the prior strength and k is a small integer that allows to intensify the prior strength de-

pending on the dataset. The value of k is further discussed and chosen in the Parameter Choice section of the Experimental analysis in *Arabidopsis thaliana*. We built the implementation of this weight function upon the iRafNet code.

We also added an alternative importance metric, the MDA. The MDA is computed as the relative increase of prediction error induced by the permutation of a given variable, as compared to the Mean Squared Error (MSE) when it is not permuted. The prediction error for a target gene t, measured by the MSE, is evaluated on the Out-Of-Bag (OOB) examples and defined as:

$$MSE_t = \frac{1}{N_{OOB}} \Sigma_{i \in OOB} (y_{t,i} - \hat{y}_{t,i})^2$$

When the expression values of the regulator r are randomized to make the prediction $\hat{y}_{t,i}$, the MSE of gene t is named $MSE_{t,\text{rand}(r)}$. The MDA between regulator r and gene t is then

$$MDA_{t,r} = \frac{MSE_{t,\text{rand}(r)} - MSE_t}{MSE_t} * 100$$

The addition of the MDA metric was implemented by modifying the C++ dependency of iRafNet, based on the dependencies found in the randomForest R package [Liaw and Wiener, 2002].

In addition to being less sensible to over-fitting, the MDA has the advantage of not requiring any scaling of the response variable to get comparable importance metrics between target genes. This is not the case in the default node purity metric in iRafNet and GENIE3 (i.e average variance decrease in the response due to a split using this regulator) because it depends on the variance of the response.

The third modification of iRafNet was to allow the restriction to a subset of candidate regulator genes as variables in the regression, that was initially performed using all genes as variables in iRafNet.

LASSO with Differential Shrinkage and Stability Selection

As RNA-Seq experiments generate count data, we model the expression of a target gene t in the condition i as a Poisson-distributed variable $Y_{t,i} \sim \mathscr{P}(\mu_{t,i})$. The parameter of the Poisson distribution $\mu_{t,i}$ is estimated on the log scale as a linear combination of the expression values of the regulator genes, with $x_{r,i}$ the expression level of regulator r in the condition i:

$$\ln(\mu_{t,i}) = \beta_{t,0} + \Sigma_{r=1}^{R} \beta_{t,r} x_{r,i} \quad (3)$$

We employ a LASSO penalty in order to overcome the high-dimensional setting and to select the most predictive regulators. In addition, we propose to use differential shrinkage in order to favor the selection of TFBS-supported variables. Differential shrinkage allows to modulate the penalty strength of each variable individually, in a way that regulators with their binding site in the target's promoter are less penalized during model adjustment. We model differential shrinkage for the LASSO using specific penalty coefficients $w_{t,r} \in [0,1]$ defined as a linear function of the TFBS prior $\Pi_{t,r}$:

$$w_{t,r}^{LASSO} = 1 - \Pi_{t,r}\alpha \quad (4)$$

where α is chosen in $[0,1]$ to tune the prior strength. For each target gene t, the function to optimize for model estimation is:

$$\text{argmin}_{\beta_t} \left\{ -\frac{1}{N} \log \mathscr{L}(\beta_t; X, Y_t) + \lambda \Sigma_{r=1}^{R} w_{t,r}^{LASSO} |\beta_{t,r}| \right\}, \quad (5)$$

where λ controls the overall strength of the penalty, R is the total number of regulator genes, N the total number of measurements and $\log \mathscr{L}(\beta_t; X, Y_t)$ is the log-likelihood function. The value of λ is learned from 5-fold cross validation: we retain the value of λ for which the minimal error (MSE) on the test folds is reached. We relied on the `glmnet` [Friedman et al., 2010] implementation of the LASSO, with the `penalty.factor` argument for specifying differential shrinkage weights.

Feature selection with LASSO can however be sensitive to noise in the data, correlated variables, and to cross validation partitioning. In order to enable the selection of robust and high confidence regulators, we employed Stability Selection, that was shown to reduce these issues [Meinshausen and Bühlmann, 2010, Bach, 2008], and was already used in a GRN inference approach combined with LARS [Haury et al., 2012]. Hence, instead of fitting one generalized linear model, S models are adjusted with, each time, small random perturbations applied to the data. More precisely, at each iteration of the stability selection procedure with differential shrinkage:

1. N observations are sampled with replacement from the N available experimental conditions, which is known as bootstrapping.
2. For each regulator gene r, the differential shrinkage weight is defined by $w_{t,r}$ to which is added a small quantity, uniformly drawn between -0.1 and 0.1. This is meant to slightly disturb the prior information, brought here from the presence of TFBS in the promoter region of the target gene t.
3. The N observations are randomly partitioned into 5 cross validation folds.
4. A model is fitted by minimizing Equation (5) with perturbed differential shrinkage weights and applied to this bootstrapped dataset.

Noteworthily, stability selection makes LASSO regression more comparable to the procedure of RFs, that are also learned from bootstrapped samples. Based on stability selection, we propose two metrics to score the regulatory influence between a target and its regulator:

- Selection frequency, defined as the number of times a regulator is selected in the LASSO regression divided by S.
- The statistical significance (two-tailed p-value corresponding to the coefficient z-ratio based on a Normal reference distribution) of the coefficient $\beta_{t,r}$ associated

with regulator r in an unpenalized Poisson regression, where the predictive variables are limited to 'robust regulators', i.e. regulators with selection frequency greater than a chosen threshold.

Final edges selection

For both integrative GRN inference methods, the final GRN is built from the edges associated with the strongest interactions (highest MDA for bRF, highest frequency selection or lowest p-values for LASSO-D3S) to satisfy a user-specified network density. Network density is defined as $D = \frac{E}{E_{total}}$ with E the number of edges (regulation relationships) in the inferred network, and $E_{total} = R(T-1)$ the total number of edges in a complete oriented GRN containing R regulators and T genes. The types of scoring metric and hyperparameters settings are explored in the experimental analysis.

Experimental analysis in Arabidopsis thaliana

Gene expression dataset

As a case study for GRN inference, we chose the transcriptomic root response to nitrogen induction in the plant *Arabidopsis thaliana* [Varala et al., 2018]. This dynamic response has the advantage of being already well characterized, and has been the basis of other methodological developments to chart regulatory networks [Varala et al., 2018, Brooks et al., 2019, Cirrone et al., 2020]. Continuing efforts to chart the regulatory mechanisms involved in nitrate response is of great agricultural interest, as nitrate is the main nutrient source of plants. Gene expression was measured in the roots at times 0, 5, 20, 30, 45, 60, 90, and 120 minutes after nitrate or control treatments [1]. As each combination of time point and treatment was measured in three replicates, this results in a total of 45 samples. We selected differentially expressed genes responding to nitrate induction in time by testing the interaction terms between nitrate treatment and time modelled as natural splines. A total of 1426 genes had FDR adjusted p-values under 0.05 for those interaction effects. Among those T = 1426 genes, R = 201 are annotated as transcriptional regulators in the TAIR10 genome release. This set of nitrate-responsive genes form the list of genes of interest taken as input for GRN inference.

Parameter choice

The choice of the parameter values may depend on the dataset, in particular on the dimension ratio between the number of regulator genes and the number of measurements. For setting parameters, we checked the strength of prior knowledge integration via the PWM support, *i.e.* the proportion of inferred edges between a regulator and a target gene, that are supported by the presence of the PWM associated with the regulator in the promoter of the target gene. We expect the PWM support to grow smoothly with the integration strength α.

[1] Although samples at 10 and 15 min were also available, we discarded them, as they were extremely different from the rest in the two first axis of a Principal Component Analysis.

Hence, all analyses involving bRF were performed with 1000 trees and the MDA as importance metric (Figure S1a) and $k = 2$ (Figure S2a) to define the TFBS prior weight $w_{r,t}^{RF}$ (Equation (2)). Stability Selection in LASSO-D3S was composed of 100 iterations. The analyses involving LASSO-D3S used a 2-step procedure with a robustness threshold of 70 % in the stability selection step and p-values to rank the selected (stable) interactions (Figure S3a).

TFBS integration improves the biological relevance of inferred GRNs

In both bRF and LASSO-D3S, using TFBS prior information increases the PWM support in the inferred GRNs (Figure 3a)). The sparsest networks, *i.e.* the networks with the lowest densities D and made of the first selected interactions, globally reach a higher PWM support than denser networks, meaning that scoring metrics in the two approaches are higher for TFBS-supported edges. Hence data integration is correctly encoded in the inference processes.

Regarding precision, *i.e* the fraction of edges validated in the ConnecTF experimental database, the sparsest networks tend to perform better, indicating that the edges scores from bRF and LASSO-D3S are informative about the biological relevance of the interactions (Figure 3b). In addition, precision is maximized with TFBS integration for both methods. The highest precision is achieved by LASSO-D3S, with a value of 43.5 %, with the sparsest network and integration strength $\alpha = 0.5$. For higher values of α, precision decreases, showing that GRN inference in the LASSO regression case requires a trade-off regarding TFBS contribution. In contrast, bRF reaches its optimal precision 42.4% for $\alpha = 0.8$ and stays almost constant until $\alpha = 1$.

Recall, that quantifies the percentage of the experimental gold standard retrieved by the inferred networks, is very similar for bRF and LASSO-D3S (Figure 3c). As expected, networks with a high density are able to retrieve more of the ConnecTF interactions than the sparsest networks. The effect of TFBS integration is to increase recall for both methods.

Overall, the precision and recall analysis indicates that TFBS information is useful in the context of predicting regulatory interactions in which the TFs binds to their target *in vivo* or *in vitro* (DAP-Seq or CHIP-Seq), or in which the TF nuclear presence alters its target expression (TARGET).

Binding sites integration brings consensus

With no integration, the sparsest networks share few edges, but the effect of data integration is stronger for them. In those high-confidence LASSO-D3S and bRF networks, increasing values of α from 0 to 0.8 increases the intersection between their edges (Figure 3d). This discloses that prioritizing to a certain degree the prediction of TFBS-supported interactions captures consensual biological signal. For α values past 0.8, bRF and LASSO-D3S contain less common edges, which suggests that they infer different interactions for very strong TFBSs contributions. The denser networks display a stable fraction of shared edges.

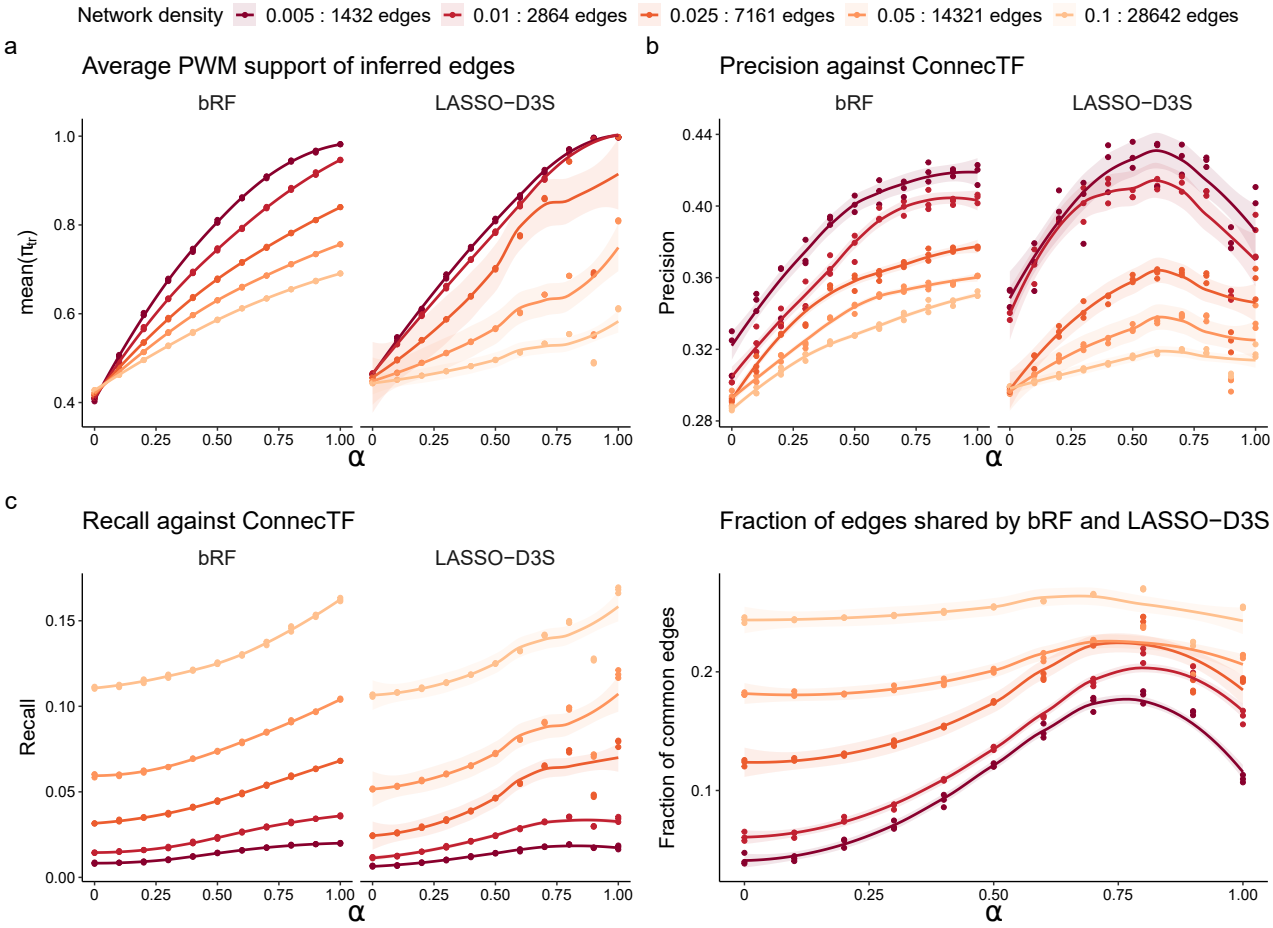


Figure 3. Comparison of bRF and LASSO-D3S performances depending on the integration strength α and for different density thresholds. Each dot represents in inference run, and allow to visualize results variability for the same set of parameters.**a.** PWM support of predicted interactions, i.e the frequency of TFBS supported edges in the GRN **b.** Precision of inferred GRNs against the ConnecTF validation database. **c.** Recall of inferred GRNs against the ConnecTF validation database. **d.** Fraction of inferred interactions in common between bRF and LASSO-D3S at comparable densities.

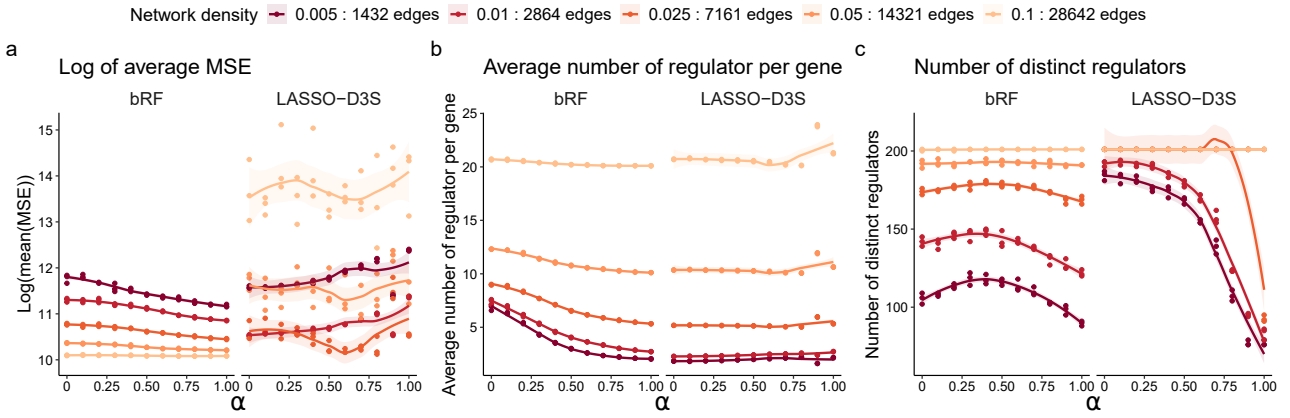


Figure 4. Comparison of bRF and LASSO-D3S predictive capabilities and GRN structures depending on the integration strength α and for different density thresholds. Each dot represents in inference run, and allow to visualize results variability for the same set of parameters. **a.** Average MSE of genes in the inferred GRNs, on the log scale. **b.** Average number of regulators per gene in the inferred GRNs. **c.** Number of distinct regulators in the inferred GRNs.

The predictive performance and topology of inferred GRNs largely vary between bRF and LASSO-D3S

We also evaluated the performance of bRF and LASSO-D3S in the task of predicting the expression levels of the target genes in the inferred GRNs. The prediction error is measured on test data, and is thus an estimator of the generalisation performance in predicting the expression of target genes using their inferred regulators.

Firstly, regardless of sparsity or TFBS integration, LASSO-D3S commits overall more error and shows more dispersed predictions between different runs under the same parameters. It must be noted that the complexity of the models (Poisson regression vs. RFs) strongly differ and impact their predictive power. Still, very different patterns in prediction errors emerge between bRF and LASSO-D3S (Figure 4a). In bRF, the addition of TFBS information slightly decreases prediction error in unseen experimental conditions, suggesting that the prioritization of TFBS-supported regulators allows more robust predictions and biologically relevant subsets of regulators. It is also notable in Figure 4b that the number of regulators per target gene is diminished by TFBS contribution, suggesting that data integration could limit over-fitting in RFs. In LASSO-D3S, error is rather stable with PWM integration. Remarkably, the choice of network density has a strong impact on prediction error. In bRF, the sparsest networks commit the more error, while it is the denser network in LASSO-D3S. In the latter, error is reduced when network density increases from 0.005 to 0.025, and then increases again in denser networks. This indicates that for linear regression, there might be an optimal network density to maximize prediction generalizability, and avoid over-fitting with too many regulators per target gene. These observations on prediction error hint that the complexity of the model (linear regression versus non linear regression) could act upon prediction performance and over fitting.

Network topology between the two approaches have common features : denser networks display an elevated number of regulators per gene while sparser networks have very few (Figure 4b). However, without data integration, bRF results in more regulators per target genes, and thus less target genes in total. For high-confidence GRNs, the number of distinct regulators is also almost half the one of LASSO-D3S, disclosing that network topologies without TFBS contribution have important differences.

As TFBSs contribution is increased, the number of regulators per gene in the inferred GRNs is constant in LASSO-D3S while it slightly decreases in bRF except for the most dense network. The diversity in inferred GRN regulators is also affected by TFBS integration in the sparsest networks. In those GRNs, the increase of α causes the number of distinct regulators to drop especially in LASSO-D3S (Figure 4c), and in a less pronounced way for bRF. This behavior is expected as data integration favors the selection of a restricted subset of regulators that have their PWM in their target's promoter. For high values of α, is it reasonable to assume that regulators lacking a PWM are more often rejected from the GRNs in favor of PWM with higher values in the prior matrix Π.

Integrative GRN inference recovers regulatory mechanisms of nitrate response in Arabidopsis roots

Well characterized regulatory mechanisms describing nitrate nutrition are available in the literature [Vidal et al., 2020, Bellegarde et al., 2017, O'Brien et al., 2016]. Such mechanisms are expected to be retrieved by our integrative GRN inference, which is based on transcriptomic data measuring temporal root response to nitrate treatment [Varala et al., 2018].

Firstly, to evaluate the biological relevance of inferred GRNs regarding nitrate induction, we looked at the position in the degree ranking of TFs known to play a central role in the control of nitrate nutrition: LBD37/38/39 [Rubin et al., 2009], HRS1/HHO2/3 [Kiba et al., 2018], TGA1/4 [Bellegarde et al., 2017], and DIV1 [Brooks et al., 2019]. We define the relative

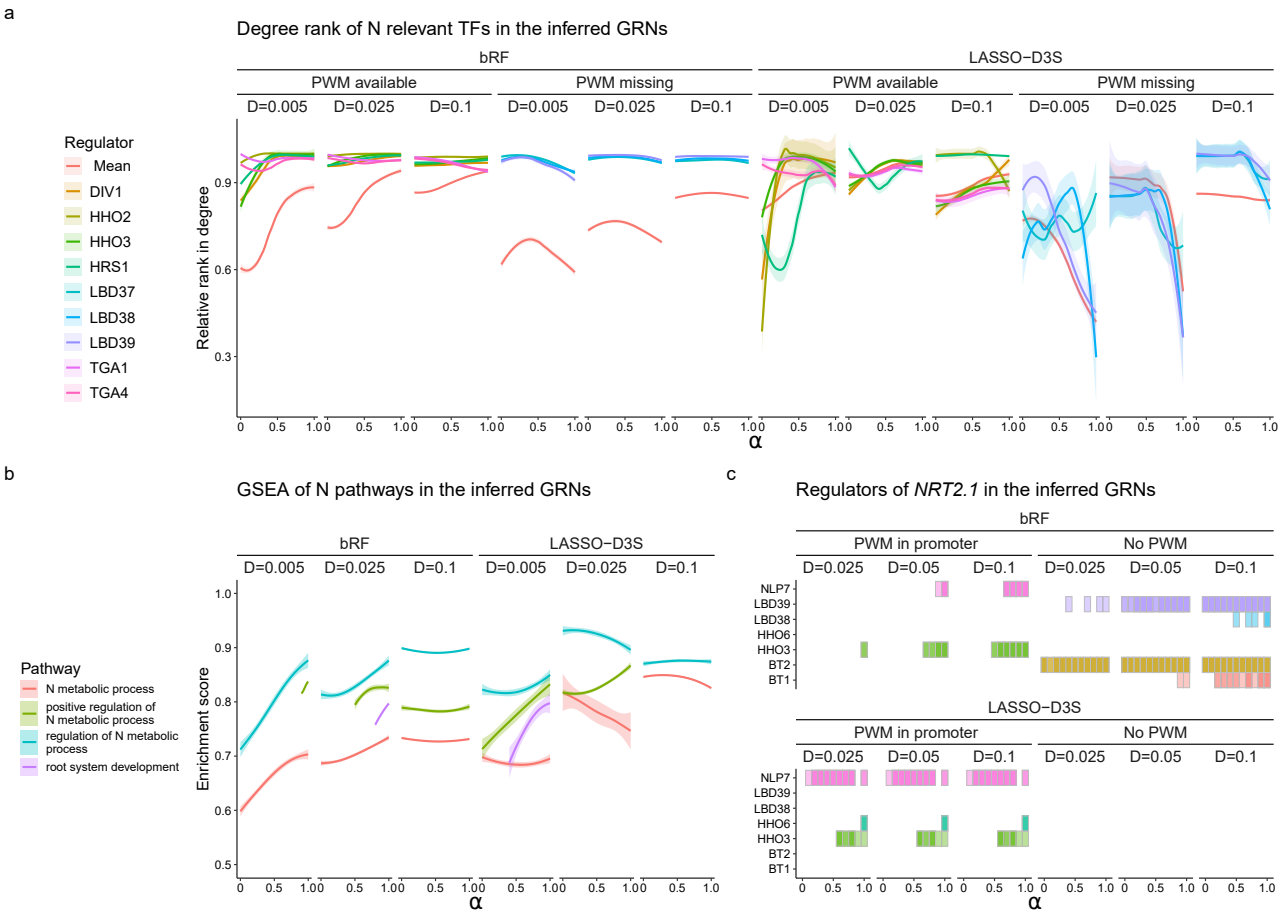


Figure 5. Pathways relevant to nitrate nutrition are realistically modeled by bRF and LASSO-D3S. a. The relative degree rank of TFs known for their role in regulating nitrate acquisition and metabolism is shown in the inferred GRN depending on α and for several densities D. The red lines represent the average relative rank of all regulators for which the PWM is either available or missing. **b.** Gene Set Enrichment Analysis (GSEA) of 5 biological processes associated to nitrate nutrition, metabolism and signalling. The hypothesis that those processes are enriched in the most connected genes of the GRNs is tested by the `gseGO` function of the `clusterprofiler` R package [Yu et al., 2012], taking as input the list of nodes ranked by degree. The enrichment scores relative to node degree that are significantly enriched (FDR ≤ 0.1) are displayed and shown as a function of α and for several densities D. **c.** Predictions of the regulation of *NRT2.1* by NLP7, HHO3/6, LBD38/39, and BT1/2 as a function of α and for several densities D. Tiles are colored if the TF regulates *NRT2.1* in the inferred GRN, with opacity reflecting the regulation robustness in 3 runs of inference.

rank in degree of a TF as its position in the overall ranking of the GRN nodes, divided by the total number of nodes in the GRN. A relative rank in degree of 1 means that a TF is the most connected node in the GRN. Nitrate-related regulators are in most cases among the most connected genes in the GRNs, as their relative degree ranking is often between 0.75 and 1 (Figure 5a). This can be explained by the fact that their expression profiles are good predictors of many target genes responding to nitrate, regardless of TFBS information. This is a first confirmation that the inferred GRNs recover important players of nitrate response. As the strength of TFBS integration increases, the nitrate-relevant TFs for which the PWM is available gain in importance and are even more connected in the GRN for both methods. This behavior is observed for all TFs for which the PWM is available independently of their function, as shown by the mean relative rank in degree of those TFs. To the contrary, the relative rank in degree of nitrate-related TFs for which a PWM is missing is high without integration, and decreases for very high values of α (Figure 5a). The fact that LBD37/38/39 keep a high degree relative to the average of other TFs with unknown PWM in bRF shows that the lack of a PWM is not too severe: if their expression profile is informative, they are still predicted as central nodes for most values of α and densities. The decrease in relative rank in degree of TFs lacking a PWM is more pronounced for low densities, and for LASSO-D3S. Even when the rank of those TFs decreases, they are not rejected from the GRN, which is the desired behavior for such TFs with a neutral prior. Those analyses indicate that important actors of nitrate nutrition are accurately modeled in GRNs, but that the effect TFBS integration on the ranking of those nitrate-relevant TFs stays partly driven by the availability of their PWMs.

We then investigated weather TFBS integration would globally results in more connectivity for genes involved in nitrogen (N) nutrition, metabolism and regulation independently of PWM availability. We thus measured the association between gene overall degree, and gene biological processes relative to N compound (GO:0006807, GO:0051171, GO:0051173), or root system development which is expected to be regulated by nitrate availability (GO:0022622) [Bellegarde et al., 2017]. Gene Set Enrichment Analyses (GSEA) demonstrate that there is a significant enrichment for those pathways among high-degree genes in the GNRs (Figure 5b). The elevation of TFBS integration strength causes some processes to become significantly enriched in top degree nodes, such as the positive regulation of N metabolic process and root system development in bRF as soon as α is sufficiently high. In bRF, N metabolic process and its regulation are even more enriched as TFBS contribution is increased, and so is the positive regulation of N metabolic process in LASSO-D3S. For dense networks, the effect of TFBS integration is lessened. This suggests that the proposed inference methods attribute topological importance at the system level to genes involved in the response to nitrate and and root system plasticity, and even more when TFBS information is utilized.

Finally, we focus on NRT2.1, a key player of the response to nitrate in Arabidopsis root tissues. NRT2.1 belongs to the multigenic family of Nitrate Root Transporters (NRT)[Bellegarde et al., 2017]. This gene transports nitrate under low nitrate conditions from the culture medium into the root cell. Several works have established that HRS1 and its homologs HHO2/3 regulate *NRT2.1* [Kiba et al., 2018, Safi et al., 2021] by directly binding to its promoter. NLP7, an established master regulator of nitrate acquisition, is also a regulator controlling the expression of *NRT2.1*, and other regulators like *HHO3* [Marchive et al., 2013]. With this information, we explored the inferred GRNs to identify predicted regulators of *NRT2.1*.

We found that without TFBS integration, *i.e* $\alpha = 0$, those TFs are never predicted by bRF as incoming regulators for *NRT2.1* (Figure 5c). Still at $\alpha = 0$ but for LASSO-D3S, only NLP7 is predicted as a regulator of *NRT2.1*. In contrast, when TFBS integration strength is pushed to values of α between 0.5 and 1, both LASSO-D3S and bRF predict HHO3 and NLP7 as a regulator of *NRT2.1*. In addition, HH06, a TF from the same family as HHO3, is also predicted to regulate *NRT2.1* in LASSO-D3S at $\alpha = 1$.

Finally, we investigated the effect of the lack of PWMs for known regulators of *NRT2.1* like LBD38/39 [Rubin et al., 2009]. Those genes are TFs but their PWM was not available in current motif databases. However, bRF still predicts LBD39 as regulating *NRT2.1* for all densities, and LBD38 for high values of α and dense networks. Similarly, we studied the inferred role of BT1/2, E3 ligases that do not directly bind to DNA and thus do not possess a PWM, but can indirectly and negatively control the expression of *NRT2.1* [Araus et al., 2016]. bRF retrieves their importance by predicting BT2 for all densities and α values, and BT1 for dense networks and strong TFBS integration. The prediction of LBD38 and BT1 as α is increased is striking as they do not have a strong value in Π, but it could be explained by the ability of bRF to model interactions between regulators. For instance, model complexity in bRF could favor the selection of BT1 or LBD38 based on the relevance of their expression profile conditionally on the expression of other TFBS-supported TFs retrieved as α increases. This might explain the fact that LASSO-D3S does not capture their action upon *NRT2.1*, as its linear framing does not model interactions between regulators.

Discussion

This work shows that TFBS information is beneficial to the accuracy and biological relevance of GRN inference in *Arabidopsis thaliana*. We believe this study is a first step in providing guidance on how to parameterize prior information integration into RF and LASSO-based regressions for GRN inference.

Both in the linear and the non-linear settings, the contribution of TFBS information relative to gene expression results in more inferred edges confirmed in the experimental gold

standard formed by ConnecTF. In particular, both precision and recall can be significantly increased by TFBS integration, which is something very desirable in the output of such classification models. A high recall might however not be as useful as a high precision, as a experimental validations are costly and can reasonably by implemented for a small set of high-confidence edges. In our case study, bRF and LASSO-D3S can both be used to make realistic models of the response to nitrate in Arabidopsis roots. As TFBS integration is gradually increased in bRF, genes from nitrate nutrition pathways are attributed more connections in inferred GRNs, and state of the art regulations of *NRT2.1* can be predicted.

A hypothesis explaining the improvement brought by TFBS integration is that it partly lifts co-linearity issues in gene expression data. Previous work on GENIE3 inference [Cassan et al., 2021] grouped together highly correlated regulators to prevent a biased estimation of their importance. Data integration overcomes this issue by providing **a way to chose between regulators with extremely close expression profiles, based on prior complementary knowledge**.

While similar behaviors between bRF and LASSO-D3S are observed regarding PWM support, recall, network topolgy, and enrichment of nitrate-relevant genes in high degree nodes, their benchmark outcomes differ on a set of criteria. Firstly, LASSO-D3S does not seem to benefit from too strong TFBS input, as shown by a clear optimum on the precision curve (Figure 3b), while bRF performs best at high values of α. This could mean that the two methods have different vulnerabilities to PWM information being missing for a significant part of the regulators. Similarly, LASSO-D3S becomes too severe against regulators with a neutral prior when α is elevated, a limitation that does not seem to affect bRF as strongly. This can be partly explain by the fact that the manner of incorporating prior information varies between the considered approaches because of the fundamental and structural differences between the RF and LASSO models. For example, differential shrinkage can be fully controlled by the definition of a variable-dependant coefficient for the penalty, while prior incorporation in bRF has inherent randomness. Second, predictions made on the expression of target genes based on inferred regulators are overall more accurate in bRF than in LASSO-D3S. This could be attributed to the complexity of the predictive models, and the ability of RFs to model non linear relationships between expression profiles, as well as their ability to combine the expression of several regulators to make use of interaction effects. This could also explain why TFs of the LBD family and BT family, that have a neutral prior value, are inferred as regulating *NRT2.1* when TFBS contribution is strengthened in bRF. **Their selection may be encouraged by the fact that co-regulators are supported by TFBS priors.**

Limitations and perspectives This study has proven that TFBS integration may improve GRN inference and opens new questions to be further investigated. First, the construction of the TFBS prior might be further studied.

There are still many TFs of which the binding motifs are unknown, a problem amplified in non-model organisms. Such TFs, are equally penalized for all target genes, and have a prior value of $\Pi_{t,r} = 0.5$. Even if they are relevant TFs that really bind to their target and influence their expression, they are unlikely to be frequently selected as regulators in GRNs inferred with strong values of α, because TFs with $\Pi_{t,r}$ of 1 will be favored instead. This limitation will be reduced as PWM databases are further completed and maintained by the community.

There is a high chance of **false positives in the scanning of PWMs** in promoter regions: the presence of a binding motif can occur by chance, or may not necessarily cause binding nor regulation in a cellular context. TFBS with low complexity in PWM databases can result in hits in almost all the promoters of an organism. This is, for instance, observed in the heavy tails of the distributions of the number of target promoters per PWM (Figure 1). In this analysis, regulators with such a widespread PWM were included, but their questionable biological relevance could lead to their exclusion or to additionally weighting regulators based on the complexity of their PWM during GRN inference.

Applying TFBS integration to other complex organisms could pose new challenges. In this work, the location of TFBS was assumed to be in the promoter regions of the genes. In organisms where regulation by **distant enhancers** is common, in particular in human, the scanning of the identified enhancers regions would be required as well.

Moreover, the form of prior values relative to TFBSs can be diverse and choosing the optimal one may not be trivial. For instance, instead of using the presence or absence for a PWM hit above a significance threshold in the prior matrix Π, the p-value or the score of the hit could be used to bring more quantitative insights. In addition, instead of using the PWM hit with the maximum score to build Π, more weight could be attributed to the regulators for which the PWM have multiple occurrences in the promoter of a target.

Also, the stability and robustness of GRN inference remains a noticeable direction for further improvements. For instance, it would be interesting to test or adapt the StARS approach [Liu et al., 2010], that also has the objective to select few relevant and robust edges, but differs from our current implementation of stability selection.

Finally, our conclusion that TFBSs combined to expression improve GRN inference would be strengthened when applied to new datasets in Arabidopsis, or even to new organisms in future works.

Funding

This work was supported by a 80 Prime fellowship from the National Center of Scientific Research (CNRS, France)

Data and code availability

Below are the links to the data used in the course of this study.

- The RNA-Seq data for the response to nitrate induction was downloaded from the GEO accession GSE97500
- The PWMs used to build the TFBS dataset were retrieved from JASPAR and the Plant Cistrome Database.
- To identify Arabidopsis TSSs and promoter regions, we relied on the TAIR10 GFF3 file.
- The regulators of Arabidopsis used for GRN inference are the union between PlnTFDB and AtTFDB
- GRN validation were done against the content of the ConnecTF database

All R scripts for the bRF and LASSO-D3S functions and relevant code for this project are available in the github repository:
`https://github.com/OceaneCsn/integrative_GRN_N_induction`

References

[Aibar et al., 2017] Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., and Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086.

[Araus et al., 2016] Araus, V., Vidal, E. A., Puelma, T., Alamos, S., Mieulet, D., Guiderdoni, E., and Gutiérrez, R. A. (2016). Members of BTB gene family regulate negatively nitrate uptake and nitrogen use efficiency in arabidopsis thaliana and oryza sativa. *Plant Physiology*, page pp.01731.2015.

[Bach, 2008] Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40.

[Banf and Rhee, 2017] Banf, M. and Rhee, S. Y. (2017). Enhancing gene regulatory network inference through data integration with markov random fields. *Scientific Reports*, 7(1).

[Bargmann et al., 2013] Bargmann, B. O., Marshall-Colon, A., Efroni, I., Ruffel, S., Birnbaum, K. D., Coruzzi, G. M., and Krouk, G. (2013). TARGET: A transient transformation system for genome-wide transcription factor target discovery. *Molecular Plant*, 6(3):978–980.

[Bellegarde et al., 2017] Bellegarde, F., Gojon, A., and Martin, A. (2017). Signals and players in the transcriptional regulation of root responses by local and systemic n signaling in arabidopsis thaliana. *Journal of Experimental Botany*, 68(10):2553–2565.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Brooks et al., 2019] Brooks, M. D., Cirrone, J., Pasquino, A. V., Alvarez, J. M., Swift, J., Mittal, S., Juang, C.-L., Varala, K., Gutiérrez, R. A., Krouk, G., Shasha, D., and Coruzzi, G. M. (2019). Network walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. *Nature Communications*, 10(1).

[Brooks et al., 2020] Brooks, M. D., Juang, C.-L., Katari, M. S., Alvarez, J. M., Pasquino, A., Shih, H.-J., Huang, J., Shanks, C., Cirrone, J., and Coruzzi, G. M. (2020). ConnecTF: A platform to integrate transcription factor–gene interactions and validate regulatory networks. *Plant Physiology*, 185(1):49–66.

[Campos and Freyre-González, 2019] Campos, A. I. and Freyre-González, J. A. (2019). Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions. *Scientific Reports*, 9(1).

[Cassan et al., 2021] Cassan, O., Lèbre, S., and Martin, A. (2021). Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. *BMC Genomics*, 22(1).

[Castro et al., 2019] Castro, D. M., De Veaux, N. R., Miraldi, E. R., and Bonneau, R. (2019). Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS computational biology*, 15(1):e1006591.

[Castro-Mondragon et al., 2021] Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Pérez, N. M., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F., and Mathelier, A. (2021). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1):D165–D173.

[Christley et al., 2009] Christley, S., Nie, Q., and Xie, X. (2009). Incorporating existing network information into gene network inference. *PLoS ONE*, 4(8):e6799.

[Cirrone et al., 2020] Cirrone, J., Brooks, M. D., Bonneau, R., Coruzzi, G. M., and Shasha, D. E. (2020). OutPredict: multiple datasets can improve prediction of expression and inference of causality. *Scientific Reports*, 10(1).

[Clercq et al., 2021] Clercq, I. D., de Velde, J. V., Luo, X., Liu, L., Storme, V., Bel, M. V., Pottie, R., Vaneechoutte, D., Breusegem, F. V., and Vandepoele, K. (2021). Integrative inference of transcriptional networks in arabidopsis yields novel ROS signalling regulators. *Nature Plants*, 7(4):500–513.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

[Geurts et al., 2018] Geurts, P. et al. (2018). dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. *Scientific reports*, 8(1):1–12.

[Gibbs et al., 2022] Gibbs, C. S., Jackson, C. A., Saldi, G.-A., Tjärnberg, A., Shah, A., Watters, A., Veaux, N. D., Tchourine, K., Yi, R., Hamamsy, T., Castro, D. M., Carriero, N., Gorissen, B. L., Gresham, D., Miraldi, E. R., and Bonneau, R. (2022). High-performance single-cell gene regulatory network inference at scale: the inferelator 3.0. *Bioinformatics*, 38(9):2519–2528.

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

[Greenfield et al., 2013] Greenfield, A., Hafemeister, C., and Bonneau, R. (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060–1067.

[Haury et al., 2012] Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012). TIGRESS: Trustful inference of gene REgulation using stability selection. *BMC Systems Biology*, 6(1).

[Hayes et al., 2013] Hayes, W., Sun, K., and Pržulj, N. (2013). Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, 29(4):483–491.

[Huynh-Thu et al., 2010] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):1–10.

[Kiba et al., 2018] Kiba, T., Inaba, J., Kudo, T., Ueda, N., Konishi, M., Mitsuda, N., Takiguchi, Y., Kondou, Y., Yoshizumi, T., Ohme-Takagi, M., et al. (2018). Repression of nitrogen starvation responses by members of the arabidopsis garp-type transcription factor nigt1/hrs1 subfamily. *The Plant Cell*, 30(4):925–945.

[Koutrouli et al., 2020] Koutrouli, M., Karatzas, E., Paez-Espino, D., and Pavlopoulos, G. A. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory.

[Kundaje et al., 2007] Kundaje, A., LIANOGLOU, S., LI, X., QUIGLEY, D., ARIAS, M., WIGGINS, C. H., ZHANG, L., and LESLIE, C. (2007). Learning regulatory programs that accurately predict differential expression with MEDUSA. *Annals of the New York Academy of Sciences*, 1115(1):178–202.

[Leclerc, 2008] Leclerc, R. D. (2008). Survival of the sparsest: Robust gene networks are parsimonious. *Molecular Systems Biology*, 4.

[Liaw and Wiener, 2002] Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

[Liu et al., 2010] Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in neural information processing systems*, 23.

[Marbach et al., 2012] Marbach, D., Roy, S., Ay, F., Meyer, P. E., Candeias, R., Kahveci, T., Bristow, C. A., and Kellis, M. (2012). Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. *Genome research*, 22(7):1334–1349.

[Marchive et al., 2013] Marchive, C., Roudier, F., Castaings, L., Bréhaut, V., Blondet, E., Colot, V., Meyer, C., and Krapp, A. (2013). Nuclear retention of the transcription factor nlp7 orchestrates the early response to nitrate in plants. *Nature communications*, 4(1):1–9.

[McCarthy et al., 2012] McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

[Meinshausen and Bühlmann, 2010] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

[Miraldi et al., 2019] Miraldi, E. R., Pokrovskii, M., Watters, A., Castro, D. M., De Veaux, N., Hall, J. A., Lee, J.-Y., Ciofani, M., Madar, A., Carriero, N., Littman, D. R., and Bonneau, R. (2019). Leveraging chromatin accessibility for transcriptional regulatory network inference in T helper 17 cells. *Genome Res.*, 29(3):449–463.

[Nicodemus and Malley, 2009] Nicodemus, K. K. and Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25(15):1884–1890.

[O'Brien et al., 2016] O'Brien, J. A., Vega, A., Bouguyon, E., Krouk, G., Gojon, A., Coruzzi, G., and Gutiérrez, R. A. (2016). Nitrate transport, sensing, and responses in plants. *Molecular Plant*, 9(6):837–856.

[O'Malley et al., 2016] O'Malley, R. C., shan Carol Huang, S., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., and Ecker, J. R. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, 165(5):1280–1292.

[Petralia et al., 2015] Petralia, F., Wang, P., Yang, J., and Tu, Z. (2015). Integrative random forest for gene regulatory network inference. *Bioinformatics*, 31(12):i197–i205.

[Qin et al., 2014] Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K., and Wang, J. (2014). Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*, 67(3):294–303.

[Rau and Maugis-Rabusseau, 2018] Rau, A. and Maugis-Rabusseau, C. (2018). Transformation and model choice for co-expression analysis of rna-seq data. *Briefings in Bioinformatics*, 19(3):425–436.

[Robinson and Oshlack, 2010] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

[Rubin et al., 2009] Rubin, G., Tohge, T., Matsuda, F., Saito, K., and Scheible, W.-R. (2009). Members of the lbd family of transcription factors repress anthocyanin synthesis and affect additional nitrogen responses in arabidopsis. *The plant cell*, 21(11):3567–3584.

[Safi et al., 2021] Safi, A., Medici, A., Szponarski, W., Martin, F., Clément-Vidal, A., Marshall-Colon, A., Ruffel, S., Gaymard, F., Rouached, H., Leclercq, J., Coruzzi, G., Lacombe, B., and Krouk, G. (2021). GARP transcription factors repress arabidopsis nitrogen starvation response via ROS-dependent and -independent pathways. *Journal of Experimental Botany*, 72(10):3881–3901.

[Scornet, 2020] Scornet, E. (2020). Trees, forests, and impurity-based variable importance. *arXiv preprint arXiv:2001.04295*.

[Studham et al., 2014] Studham, M. E., Tjärnberg, A., Nordling, T. E., Nelander, S., and Sonnhammer, E. L. L. (2014). Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, 30(12):i130–i138.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Tjärnberg et al., 2013] Tjärnberg, A., Nordling, T. E., Studham, M., and Sonnhammer, E. L. (2013). Optimal sparsity criteria for network inference. *Journal of Computational Biology*, 20(5):398–408.

[Varala et al., 2018] Varala, K., Marshall-Colón, A., Cirrone, J., Brooks, M. D., Pasquino, A. V., Léran, S., Mittal, S., Rock, T. M., Edwards, M. B., Kim, G. J., Ruffel, S., McCombie, W. R., Shasha, D., and Coruzzi, G. M. (2018). Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants. *Proceedings of the National Academy of Sciences*, 115(25):6494–6499.

[Vidal et al., 2020] Vidal, E. A., Alvarez, J. M., Araus, V., Riveras, E., Brooks, M. D., Krouk, G., Ruffel, S., Lejay, L., Crawford, N. M., Coruzzi, G. M., and Gutiérrez, R. A. (2020). Nitrate in 2020: Thirty years from transport to signaling networks. *The Plant Cell*, 32(7):2094–2119.

[Yu et al., 2016] Yu, C.-P., Lin, J.-J., and Li, W.-H. (2016). Positional distribution of transcription factor binding sites in arabidopsis thaliana. *Scientific Reports*, 6(1).

[Yu et al., 2012] Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287.

Validation type	Number of TFs
TARGET	33
CHIP-Seq	5
DAP-Seq	40
Total	60

Table S1. Representation of the 201 nitrate-responsive TFs in each type of data in ConnecTF

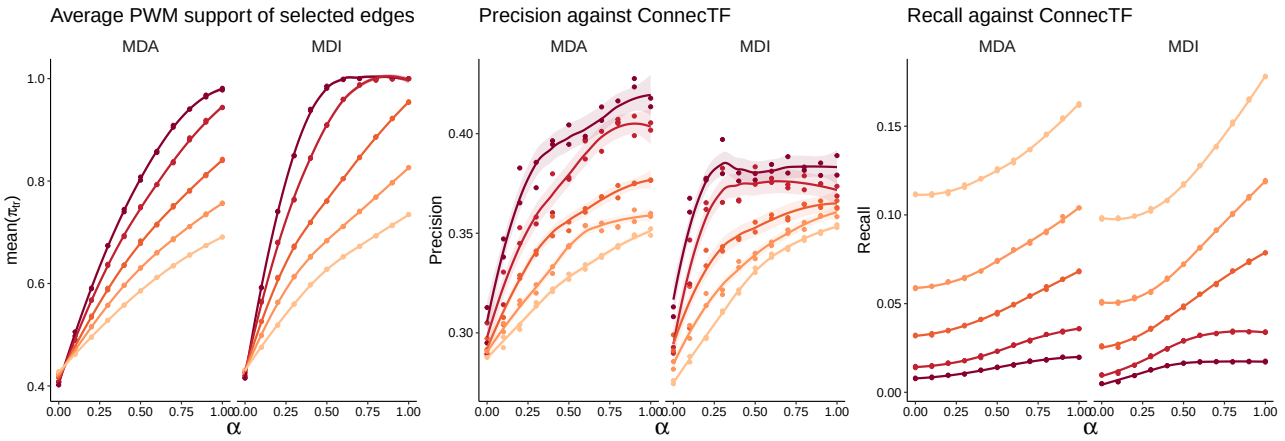


Figure S1. Comparison of importance metrics in bRF depending on the integration strength α and for different density thresholds : the MDA (Mean Decrease in Accuracy) and the MDI (Mean Decrease in Impurity).a. PWM support in inferred edges. **b**. Precision on ConnecTF of inferred edges. **c**. Recall on ConnecTF of inferred edges. While both importance metrics allow an improvement of GRN predictions as more prior information is used, the MDA performs better in terms of precision.

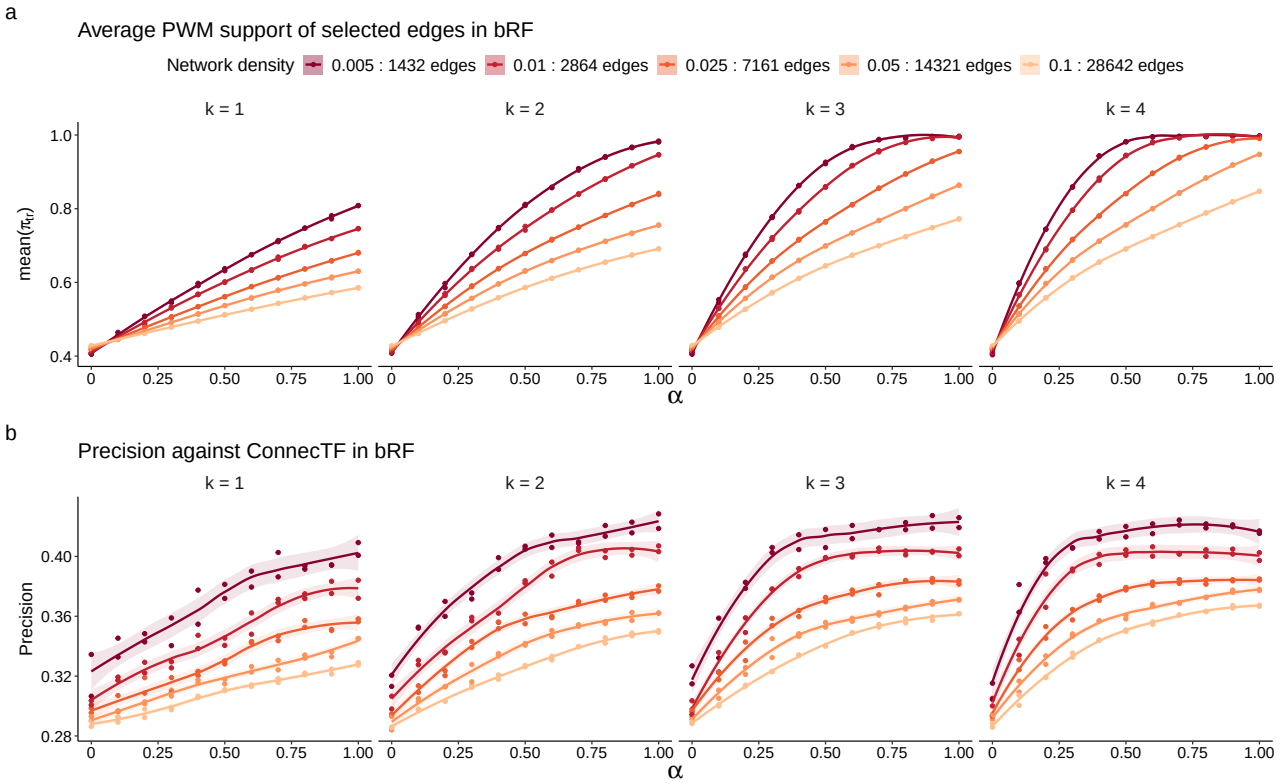


Figure S2. The parameter k tunes the link function $w_{r,t} = 10^{k\Pi_{r,t}\alpha}$ for prior integration in bRF, depending on the integration strength α and for different density thresholds. a. Mean PWM support in inferred edges for different values of k. **b**. Precision on ConnecTF of inferred edges for different values of k. Overall, these results suggest that $k = 2$ is sufficient to permit efficient data integration and produce the highest strongest values.

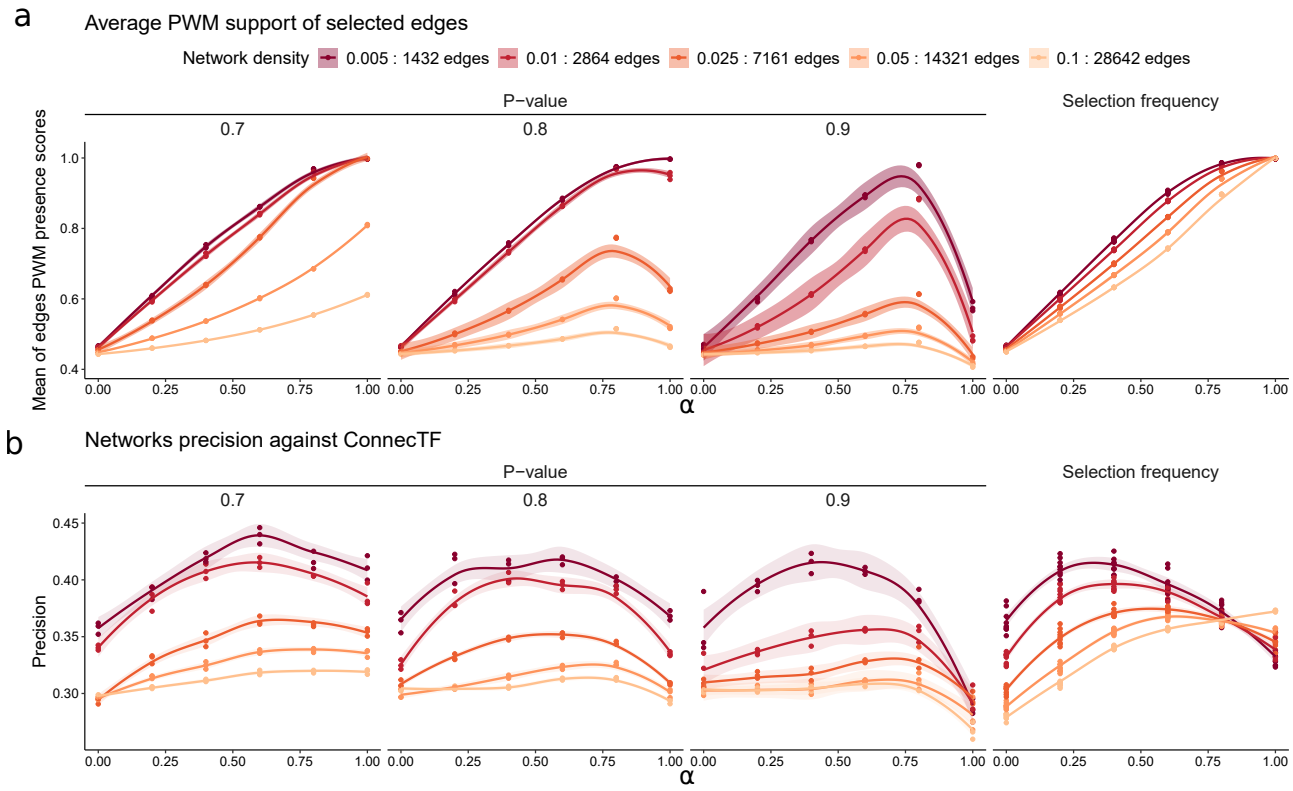


Figure S3. PWM support and precision for different scoring metrics (p-value of selection frequency), and robustness thresholds in LASSO-D3S. a. Mean PWM support of inferred edges. **b**. Precision on ConnecTF of inferred edges. Selection frequency is not the best scoring method to quantify regulatory influence. Indeed, with increasing values of α, the sparsest and supposedly higher confidence networks do not perform better than lower confidence networks. This means that even though this scoring metric permits to increase the PWM support of inferred GNRs, it does not rank inferred edges properly when benchmarked on against known regulatory interactions. We thus propose to work with the second metric : the p-value of robust regulators in unpenalized Poisson regressions. The selection frequency required at this step should not be increased over 0.7 : indeed, it deteriorates at the same time the PWM support of the inferred GRNs, but also the biological relevance of inferred edges for high values of α.

2.4 Understanding the gradual gene expression reprogramming under CO_2 gradients and two N regimes

Note: Code and data to reproduce the results presented in this section are available in the github repository `https://github.com/OceaneCsn/gradientCO2`

The GRN inference results presented so far are based on steady state transcriptomes under contrasted CO_2 and nutrition conditions. There is, however, substantial evidence that regulatory mechanisms occur in an adaptive manner, in time or along the gradual change of an environmental variable. The variation of a continuous environmental variable offers a new dimension to GRN inference by providing expression profiles at a better resolution, and allowing to model finer relations between them.

To measure relevant expression changes in the context of rising CO_2, we generated a transcriptomic dataset of the root response to gradually increasing CO_2 concentrations. In addition to allowing a better resolution for GRN inference, measuring gene expression under a full range of CO_2 concentrations has the potential to shed light on the dynamic of CO_2 response: is gene expression linearly reprogrammed as CO_2 rises, or are there step functions and abrupt changes at specific CO_2 levels? Furthermore, we made the decision to investigate different types of N sources for the plant: nitrate and ammonium nitrate nutrition. This was motivated by the observation in the literature that nitrate nutrition elicits more severe phenotypic responses than ammonium nitrate nutrition in the face of CO_2 elevation. In particular, ammonium nutrition appears to be less penalized in terms of acclimation of photosynthesis to eCO_2 [Asensio, Rachmilevitch, and Bloom, 2015].

Arabidopsis Columbia ecotypes were hydroponically grown in 5 different controlled chambers, differing only in their CO_2 concentrations: 400, 525, 650, 775 and 900ppm. Inside a chamber, plants were separated in two groups, one receiving nitrate (KNO_3) and one receiving an ammonium nitrate (NH_4NO_3) mix, both resulting in a N concentration equal to 0.5 mM. Plants were sampled 5 weeks after the beginning of the experiment, and inside each combination of N nutrition and CO_2 concentration, plants were separated for downstream analyses :

1. 12 plants were used for shoot biomass and N content analyses (2.4.1)
2. 4 samples were formed to be the 4 technical replicates in root transcriptomic analyses (2.4.2). Each sample contains the root systems of approximately 5 plants pooled together.

In the following statistical analyses, we chose natural cubic splines to flexibly model phenotypes or gene expression as a continuous non linear function of CO_2. Spline regression fits a set of piece-wise polynomials to experimental points. Natural cubic splines are characterized by their degree of freedom (DF), which is the number of knots delimiting the intervals of each polynomial. Increasing the number of DF permits to model more complex response behaviors. In the case of natural cubic splines, continuity assumptions and minimal curvature at the knots are enforced to obtain maximal smoothness and goodness of fit. In practice, a natural cubic splines basis for the CO_2 variable is first computed for a given DF, and then taken as input by a desired regression function like the linear model. This results in one coefficient per DF relative to CO_2 in the regression output, which can then be classically analysed in terms of effect size, sign, and significance.

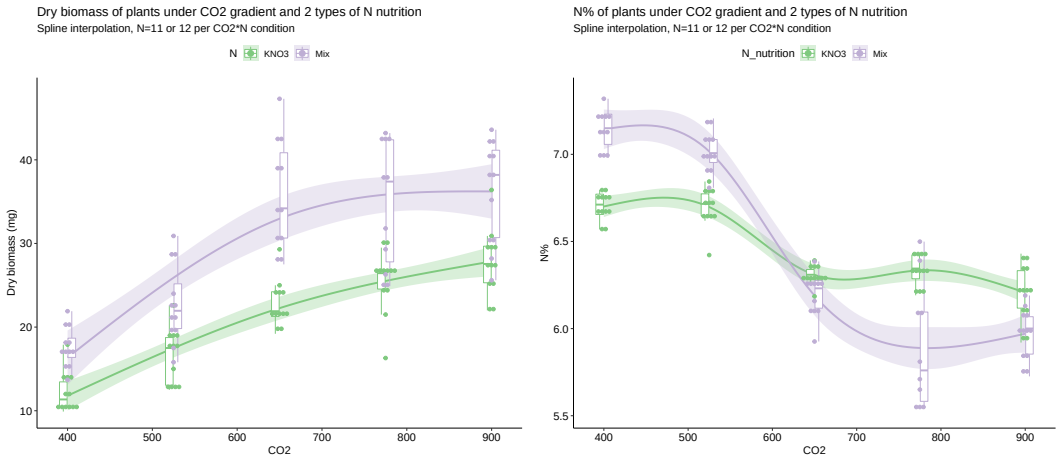


FIGURE 2.2: Dry biomass (mg) and N% of dry shoot matter are plotted as a function of CO_2 level and N nutrition (KNO_3: nitrate, Mix = ammonium nitrate). Each point is one plant, with $N \sim 12$ plants per condition. Splines interpolations are shown with 2 DF for biomass and 4 DF for N content.

2.4.1 Phenotypic response to a CO_2 gradient

Biomass production and N content responded accordingly to what was already documented in the literature and observed in our previous combinatorial study. Biomass production is stimulated by the rise of CO_2 concentration (Figure 2.2). In accordance with previous findings [Asensio, Rachmilevitch, and Bloom, 2015], ammonium nitrate nutrition seems favored: as compared to nitrate nutrition, not only plants grown with ammonium nitrate have a higher average biomass, but CO_2 increases even more their biomass (Table 2.1).

Second, increasing CO_2 levels depletes the N content of plants from both types of N nutrition (Figure 2.2). Interestingly, plants receiving ammonium nitrate contain more N under ambient CO_2 than plants receiving nitrate, but they contain less N under elevated CO_2, showing a more pronounced CO_2 effect on N decrease (Table 2.1). In fact, a conclusion can be made from those results that ammonium nitrate nutrition accentuates the biomass and N content responses to eCO_2. It also seems that the evolution of N content is less linear than biomass variation, as it appears that the biggest drop in N content occurs between 525 and 650 ppm, but do not vary as much for lower or higher concentrations. This motivated the use of 4 degrees of freedom for N content instead of 2 for biomass in our splines models.

2.4.2 Transcriptomic response to a CO_2 gradient

In order to study the transcriptomic regulations under these experimental conditions and eventually explain the observed phenotypes, we first proceeded to a global analysis of gene expression variation using RNA-Seq experiments. RNA-Seq were generated and sequenced as in the transcriptomic dataset of Publication #3.

Raw fastq files were treated for quality control with `fastp`, mapped to the TAIR10 reference genome with `STAR`, and the expression level of each gene was quantified with `htseq-count`. The raw expression matrix was normalized with the TMM method and lowly expressed genes were removed when the sum of their counts in the 40 samples did not exceed 400. The global variation of gene expression was first

	Dry biomass (mg)	N content (%)
(Intercept)	11.80^{***}	6.70^{***}
ns(CO2, df = DF)1	22.50^{***}	-0.55^{***}
ns(CO2, df = DF)2	10.78^{***}	-0.33^{***}
N_nutritionMix	4.55^{*}	0.44^{***}
ns(CO2, df = DF)1:N_nutritionMix	10.88^{*}	-0.58^{***}
ns(CO2, df = DF)2:N_nutritionMix	0.10	-1.01^{***}
ns(CO2, df = DF)3		-0.31^{**}
ns(CO2, df = DF)4		-0.57^{***}
ns(CO2, df = DF)3:N_nutritionMix		-0.83^{***}
ns(CO2, df = DF)4:N_nutritionMix		-0.66^{***}
R^2	0.74	0.88
Adj. R^2	0.73	0.87
Num. obs.	118	116

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE 2.1: Summaries of statistical models for dry biomass and N% responses to eCO_2. Biomass and N content are linearly modeled as the combination of the binary N_nutrition variable and the natural splines basis of the CO_2 gradient. Natural splines are referred to as ns. DF = 2 for biomass and DF = 4 for N content, which is why some coefficients (relative to DF 3 and 4) are not present in the biomass model.

investigated through a PCA analysis of the log-transformed normalized expression (Figure 2.3). This revealed that the major driver of gene expression is the type of N nutrition, as shown by the placement of the samples along the first principal component (57.3% of total expression variance). The second driver of gene expression is the CO_2 concentration, as samples are gradually organized by CO_2 level along the second principal component (8.3% of the total gene expression variance). The progressive divergence of the samples from different N sources as CO_2 is increased suggests that CO_2 elevation may trigger specific regulations depending on the source of N.

In order to identify differentially expressed genes (DEGs) in this experiment, we relied on the negative binomial modelling offered by `EdgeR`. The mean and dispersion of gene expression distributions were estimated under the `N_nutrition*ns(CO2, DF = 2)` design. A negative binomial generalized log-linear model was then fit to each gene, from which lists of DEGs can be established by gene-wise statistical tests for a given coefficient. At a FDR threshold of 0.5%, 2108 genes are detected as responding to CO_2 elevation in the reference N source, nitrate. 1816 genes are differentially expressed by the ammonium nitrate nutrition in the reference level of CO_2, aCO_2, and 4454 genes are differentially expressed by the interaction between ammonium nitrate nutrition and CO_2 elevation. This elevated number of 4454 DEGs responding to CO_2 more specifically under ammonium nitrate supply indicates that N source is a determining factor of the response to CO_2. Given the strong impact of the type of N supplied to the plant on gene expression, and that our previous works focused on nitrate nutrition, we decided to carry out a first detailed analysis of the 2108 CO_2 responsive genes under nitrate nutrition only.

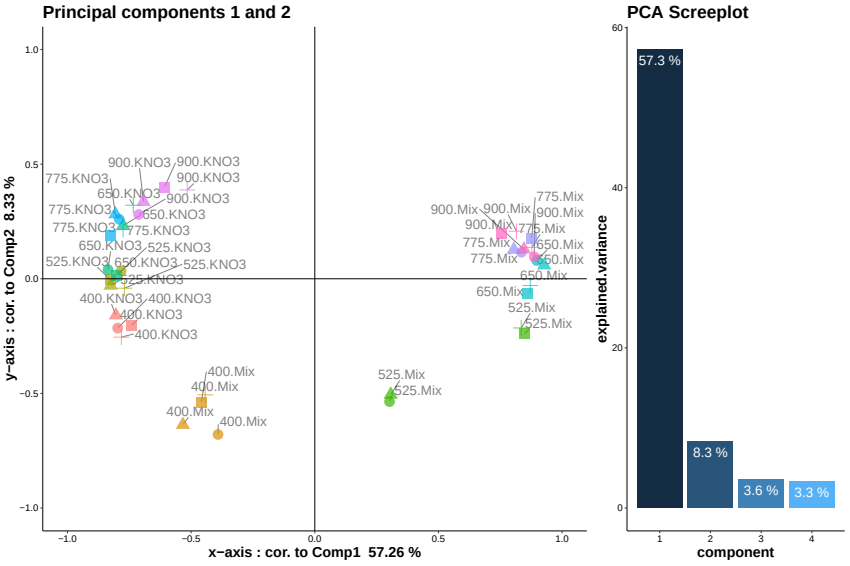


FIGURE 2.3: PCA of the log-transformed normalized expression dataset. On the left, each point represents a sample, colored by its condition, and shaped according to the replicate number. The contribution of each sample to the principal components 1 and 2 are the samples coordinates. On the right, the screeplot displays how much of the total gene expression variation is explained by each principal component.

2.4.3 GRN inference of the response to a CO_2 gradient under limiting nitrate supply

Based on the methodology developed and explored in Publication #4, we inferred a GRN of the root adaptation to rising CO_2 using expression data combined with TFBS information. As a first approach, we chose the bRF model. Indeed, it was shown to have a high precision and recall against ConnecTF when modelling the root response to N induction, and a low prediction error, but also to retrieve important interactions between genes even in the absence of knowledge regarding the TFBS of a regulator, which was not always the case for LASSO-D3S.

The input data for bRF was composed of :

1. The normalized expression matrix of the 2108 CO_2 responsive genes under nitrate supply only. Among those 2108 CO_2 responsive genes, 316 are known as transcriptional regulators in Arabidopsis.

2. The TFBS scoring matrix Π. Among the 316 CO_2 responsive regulators, 115 have a PWM available in JASPAR or the Plant Cistrome database. After a FIMO search for the 115 known PWM in the 2108 promoters of CO_2 responsive genes, Π was either filled with 1, 0, or $\frac{1}{2}$ depending on weather the PWM was found in a promoter, not found in a promoter, or is not available in the motif databases.

Preliminary explorations showed that TFBS information was efficiently integrated in this dataset in the course of inference (i.e all the edges of the inferred GRNs can rapidly be supported by TFBSs), so we set the hyper-parameter k of bRF to 1. To further investigate the effect of TFBS integration, we set the desired network density to

0.005, which corresponds to 3329 edges, and explored the PWM support, precision and recall against ConnecTF for a range of possible integration strengths α (Figure 2.4). This figure allowed us to set $\alpha = 0.8$ because it guarantees high enough PWM support and recall, while precision does not benefit from higher values of α. These parameter choices lead to a final inferred GRN containing 237 regulators, 817 target genes, with a PWM support of 79.1%, a precision of 52.9% and a recall of 1.7%.

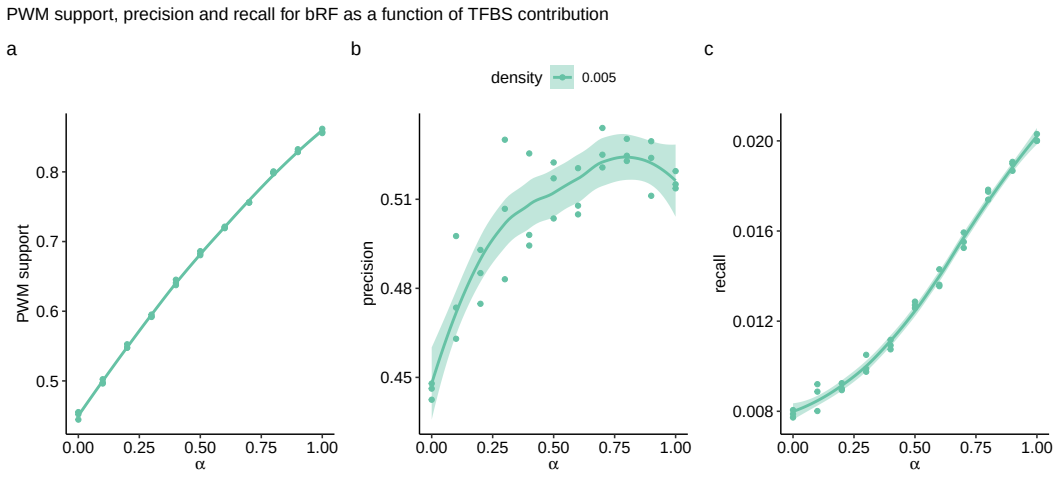


FIGURE 2.4: **Parametric exploration of bRF applied to the CO_2 gradient dataset. a.** PWM support of inferred GRNs, defined as the mean of the TFBSs scores of selected edges. A PWM support of one means that all edges of the inferred GRN are supported by a TFBS. **b.** Precision against ConnecTF. **c.** Recall against ConnecTF.

We also checked that the inferred GRN had a precision significantly higher than expected by chance, by comparing the observed precision to the precision of 50 random networks with the same gene composition and similar topologies. None of the random networks had a precision as high as our inferred GRN, the precision of which being 9.1 standard deviations away from the precision expected by chance (Figure 2.5). This supports the hypothesis that our GRN inference significantly captured relevant features of gene regulation. Furthermore, the inferred GRN exhibits standard properties of biological networks such as a degree distributions showing a small number of hub genes and a large number of lowly connected genes (Figure 2.6).

In a similar way to our previous GRN inference works, we examined the ranking of genes based on their connectivity in the predicted network. Strikingly, an elevated number of genes in the most connected regulators are regulators already known for their involvement in N nutrition pathways (Figure 2.7). CDF3 (AT3G47500) has the highest overall degree and out-degree, and has already been established to control N response and N use efficiency in Arabidopsis and tomato [Domínguez-Figueroa et al., 2020]. The second, fourth and sixth top ranked TFs belong to the NIGT1 family: HHO3, HRS1 and HHO2. The role of those TFs in the regulation of nutrient acquisition and especially of N has been clearly established, as well as their ability to bind the promoter of their targets [Kiba et al., 2018; Safi et al., 2021]. UIF1, another HRS1 homolog also known as HHO5 is the 8th most connected gene in the GRN, which reinforces the suspected role of this family in the CO_2 response. RAV1 and BZIP3 are encountered slightly lower in the nodes ranking by overall degree. Those two TFs were more recently discovered as important players in the temporal response

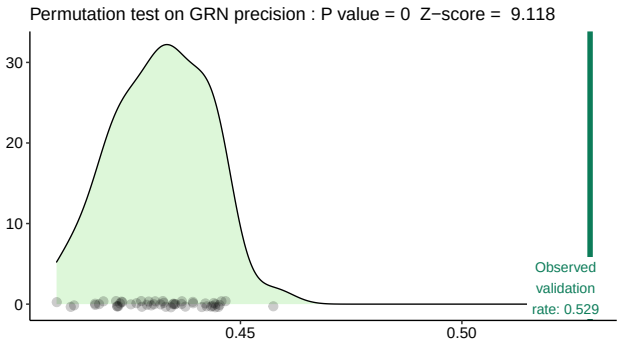


FIGURE 2.5: Permutation-based test for the inferred GRN precision. The edges of the GRN were randomly unmatched 50 times, and the precision of those 50 random networks against ConnecTF was assessed. The p-value is computed as the fraction of random networks whose precision exceeded the observed precision of our inferred GRN. The z-score is the number of standard deviations of the null distribution that separate the observed precision and the precision expected by chance.

to N induction [Brooks et al., 2019]. The 16th and 35th top ranked TFs are interestingly NLP6 and NLP7, master regulators of the nitrate networks in plants [Marchive et al., 2013]. In particular, NLP7 was proven to be post-translationnally regulated by nitrate availability, and is expected to stay strongly expressed in our experiment conducted on 0.5 mM of N. The fact that *NLP7* is differentially expressed along a CO_2 gradient hints that there may be other modes of regulation of NLP7 in the CO_2 response. Other members of the NLP family (NLP1/2/3) are also found in the GRN, but with a lower number of connections.

BT1, a negative regulator of N nutrition, was found to have 41 connections in the GRN. BT1 does not bind to DNA but has the capacity to modulate the binding to DNA of other TFs, and thus can not be supported by a PWM in the inference process. Still, its expression profile was predicted as important enough to regulate 24 targets in combination with other TFs. Finally, CCA1 is predicted as a regulator of *NRT2.1* and 23 other genes. This regulator, predominantly recognized for its control of the circadian clock in Arabidopsis, was also identified by systems biology approaches as involved in the N signalling networks [Gutiérrez et al., 2008].

Given the topological importance attributed to TFs controlling N nutrition, we looked at important target genes for nitrate transport, signalling and metabolism. We first wondered if they were present in the GRN, and if so, which genes were their regulators. The G6PD2/3, belong to the oxidative pentose phosphate pathway, a pathway that acts upon nitrate transporters in the context of light and sugar signals [Lejay et al., 2008]. They seem to be active in the response to a CO_2 gradient, as they have an important number of incoming edges in the GRN, respectively 10 and 18. In the pathway of nitrate assimilation, NIR1 is a main enzyme responsible for nitrate reduction, and receives 10 connections. Interestingly, GRXS13 was predicted as having 16 regulators. This gene comes from the glutaredoxin family, a family previously identified in the regulation of N transport through systemic signals [Ohkubo et al., 2017; Ota et al., 2020]. It was also found to be the target of many highly connected regulators in our previous GRN of the CO_2 response combined with nitrate

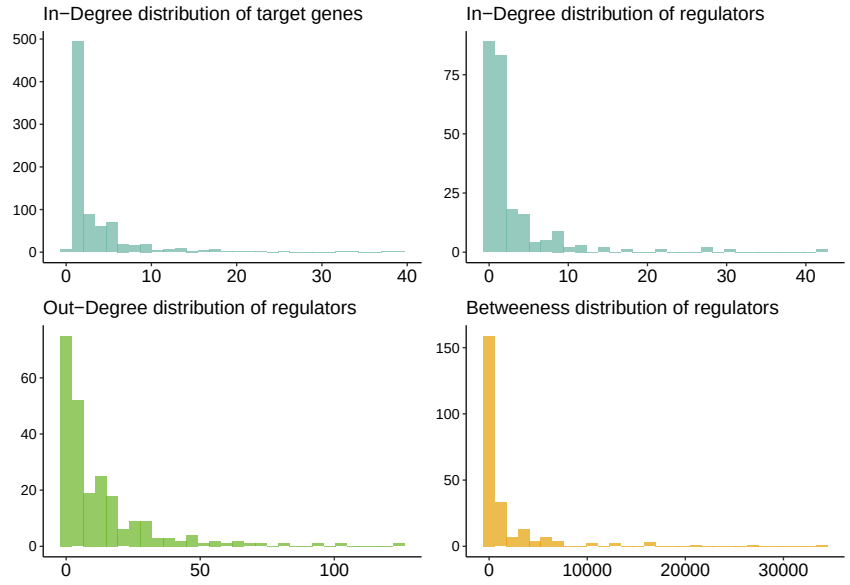


FIGURE 2.6: In-degree and out-degree distributions of the GRN regulators and target genes, and betweeness distribution of the regulators.

limitation, and consequently deserves further attention in the context of understanding the CO_2 response. Finally, nitrate transporters NRT1.12, NRT2.1 and NRT2.2 are included in the GRN. They are however less connected than the previously mentioned genes related to N metabolism or signalling, as they have respectively 1, 2 and 3 incoming regulations.

A subnetwork of N relevant genes including the aforementioned regulators and target genes shows that some important and functional links between genes involved in N nutrition are accurately modeled (Figure 2.8). For example, it was already shown that in mutant lines for NLP6, the expression of G6PD3 was significantly changed [Konishi and Yanagisawa, 2014], and that NLP6 regulates NIR1 by binding to its promoter in different studies [Gaudinier et al., 2018], which is supported by our inference as well.

Overall, these results disclose a strong implication of the actors of N signalling, metabolism and transport in the response to rising CO_2. The predominance of these genes in the structure of the inferred GRN, even though the nitrate concentration remained constant in the CO_2 gradient experiment, is striking. It puts even more in the spotlight the hypothesis that regulatory networks triggered by CO_2 could negatively act upon N acquisition and assimilation.

	label	gene_type	degree	degree_in	degree_out
AT3G47500	CDF3	Regulator	133	9	124
AT1G25550	HHO3	Regulator	116	14	102
AT2G28810	AT2G28810	Regulator	97	3	94
AT1G13300	HRS1	Regulator	85	2	83
AT5G65310	HB5	Regulator	74	2	72
AT1G68670	HHO2	Regulator	80	10	70
AT1G19000	AT1G19000	Regulator	73	9	64
AT4G37180	UIF1	Regulator	70	6	64
AT1G74840	AT1G74840	Regulator	61	1	60
AT1G13260	RAV1	Regulator	99	42	57
AT2G38090	AT2G38090	Regulator	63	6	57
AT5G15830	BZIP3	Regulator	52	2	50
AT5G17300	RVE1	Regulator	49	0	49
AT5G59990	AT5G59990	Regulator	53	7	46
AT5G47390	MYBH	Regulator	48	2	46
AT1G64530	NLP6	Regulator	48	3	45
AT2G40260	AT2G40260	Regulator	45	1	44
AT3G09600	RVE8	Regulator	52	11	41
AT5G60200	TMO6	Regulator	40	0	40
AT3G13810	ID	Regulator	41	3	38
AT5G61590	DEWAX	Regulator	68	31	37
AT2G22430	HB6	Regulator	37	1	36
AT1G31050	AT1G31050	Target Gene	43	7	36
AT4G38620	MYB4	Regulator	42	8	34
AT5G28770	BZO2H3	Regulator	62	28	34
AT1G06160	ORA59	Regulator	35	4	31
AT1G01060	LHY	Regulator	36	5	31
AT4G37260	MYB73	Regulator	33	3	30
AT2G02080	ID	Regulator	31	1	30
AT1G03040	AT1G03040	Regulator	52	22	30
AT1G07640	OBP2	Regulator	32	3	29
AT3G30260	AGL79	Regulator	35	6	29
AT5G41410	BEL1	Regulator	29	0	29
AT1G76590	AT1G76590	Regulator	35	7	28
AT4G24020	NLP7	Regulator	28	1	27
AT5G58620	TZF9	Regulator	29	2	27
AT2G25900	ATCTH	Regulator	54	28	26
AT1G21450	SCL1	Regulator	36	11	25
AT5G63160	BT1	Regulator	41	17	24
AT2G46830	CCA1	Regulator	33	9	24

FIGURE 2.7: List of the 40 genes with the highest overall degree in the inferred GRN. In-degree and out-degree are reported as well. Genes previously documented to control play a role in N nutrition are highlighted.

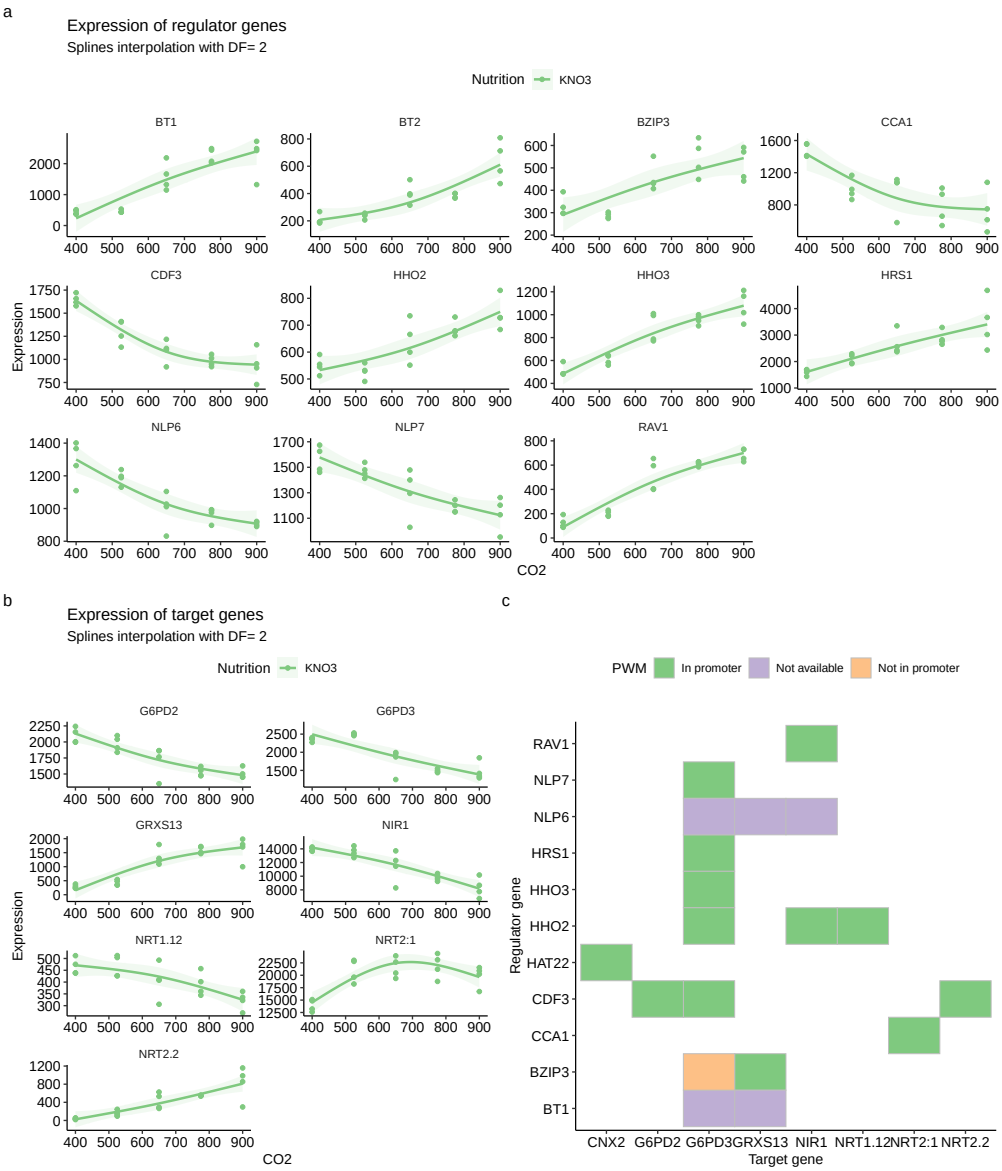


FIGURE 2.8: **Expression and predicted interactions between highly connected regulators previously described for controlling N nutrition, and key target genes involved in nitrate transport, assimilation and signalling. a.** Normalized gene expression of regulators as a function of CO_2 concentration **b.** Normalized gene expression of target genes as a function of CO_2 concentration **c.** Adjacency matrix of regulators and target genes. A cell is filled if a regulatory interaction is predicted between the regulator and the target. A cell is green if the PWM of the regulator was found in the target's promoter or yellow if it was not found. A cell is purple if the PWM of the regulator was not available in the motifs databases (like NLP6), or if the regulator is not a TF and thus does not have a PWM (like BT1).

Chapter 3

A Genome-Wide Association study identifies candidate genes in the ionome response of Arabidopsis under elevated CO_2

Existing literature discloses that there is intra-specific variability in the phenotypic response to rising CO_2 in Arabidopsis as well as in plants of agronomic interest (1.1.2.2, [Zhu et al., 2018; Myers et al., 2014a]). The variability observed in the mineral status response to eCO2 was however never explained by genomic determinants in plants, even though it has the potential to fuel the discovery of new candidate genes controlling their mineral depletion. In this chapter, we detail a GWAs project with the aim to identify genetic determinants of mineral status response to CO_2 elevation. Three populations of Arabidopsis accessions were chosen to explore different geographic scales of natural variation and maximize genetic diversity in our screening :

- A subset of the REGMAP population [Horton and Bergelson, 2012], which contains accessions originating from all around the world. It is one of the most important population to study natural variation in *Arabidopsis thaliana*.
- The LANGUEDOC population [Brachi et al., 2013] which contains accessions from the Languedoc region in France.
- The TOU-A population [Frachon et al., 2017] which contains accessions from a single meadow, in east France.

In each population, approximately 200 ecotypes were chosen to be part of our study. 5 plants from each ecotypes were cultivated in aCO_2 and eCO_2 on soil substrate, and supplied with a nutrient solution containing 10 mM of nitrate. After three weeks, plants shoots were sampled and dried. The 5 replicates of each ecotypes were then pooled together, and their mineral content was measured using acidic digestion and a microwave-plasma atomic emission spectrometer (MP-AES, Agilent). Nitrogen and carbon composition of shoots was obtained using a mass spectrometer coupled with an Elementar Pyrocube analyzer. In total, the 8 elements measured were carbon (C), nitrogen (N), sodium (Na), magnesium (Mg), manganese (Mn), iron (Fe), zinc (Zn) and copper (Cu). An overview of the study is provided in Figure 3.1. The design and generation of the biological material predates the start of this PhD and was mainly carried out by Léa-Lou Pimpare, in the growth chambers provided by the European Ecotron of Montpellier (CNRS). I was later in charge of the statistical

analysis of the phenotypic data, and of establishing the statistical associations between genetic variants and phenotypic outcomes. I also analysed and interpreted the output of the association studies and I was involved in planning experimental validations.

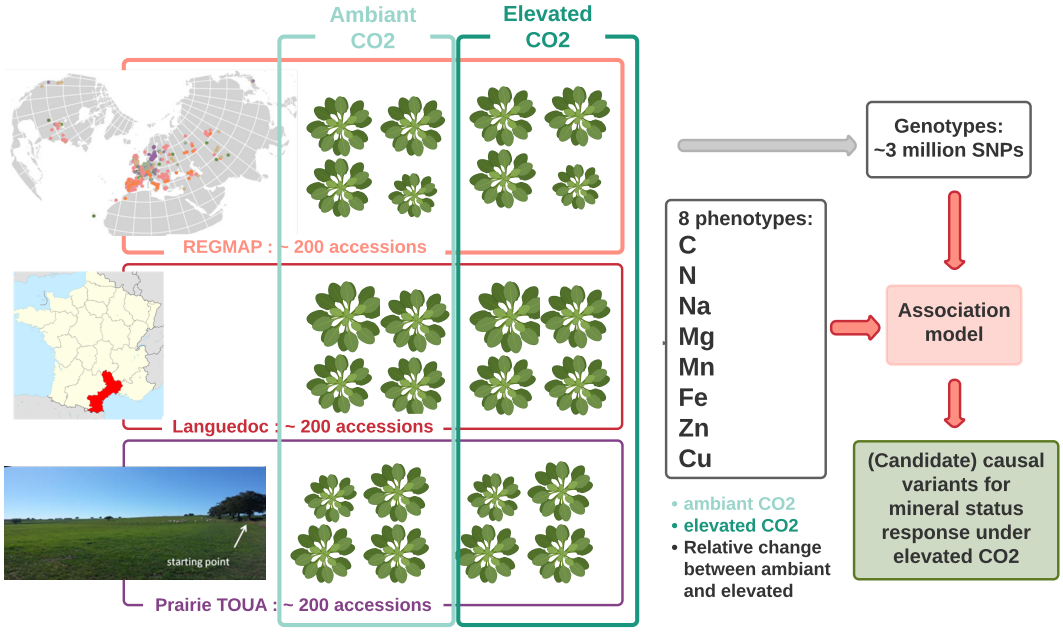


FIGURE 3.1: Experimental design of the GWAs project on the mineral status response to eCO2 in *Arabidopsis thaliana*.

Mineral composition measures were first processed to remove technical outliers. For a given element and CO_2 condition, the values more than 5 median absolute deviations (MAD) away from the median were rejected. The use of the median-based metrics was motivated by the heavy tailed distributions of the raw data with extreme values, for which median-based estimators are more robust. The value 5 was chosen as a trade-off to remove obvious clusters of outliers while preserving true biological variation. Relative changes due to eCO_2 were then derived for each element, and expressed as percentages :

$$\text{Element change} = \frac{\text{element}_{eCO_2} - \text{element}_{aCO_2}}{\text{element}_{aCO_2}} * 100$$

Those relative changes, as well as their steady state values, were the phenotypic traits of interest in the following analyses.

3.1 The ionome response to elevated CO_2 is highly variable in three natural populations of Arabidopsis

Note: Code and data to reproduce the results presented in this section are available in the github repository `https://github.com/OceaneCsn/gwas_ionome_CO2`

The overall trend in the three populations is a decrease of mineral content (Figure 3.2), as demonstrated by predominantly negative relative changes. The results of statistical tests comparing aCO_2 and eCO_2 levels of each element revealed that the

elements Fe, N and Mg significantly decrease in the three populations. Cu, Mn and Zn are significantly reduced in two populations and not significantly in the third. C is much more stable among the three populations, being significantly increased by eCO_2 in two of them. Only Na inconsistently behaved between populations, being significantly increased in the REGMAP population, reduced in the TOU-A population, and stable in the LANGUEDOC population.

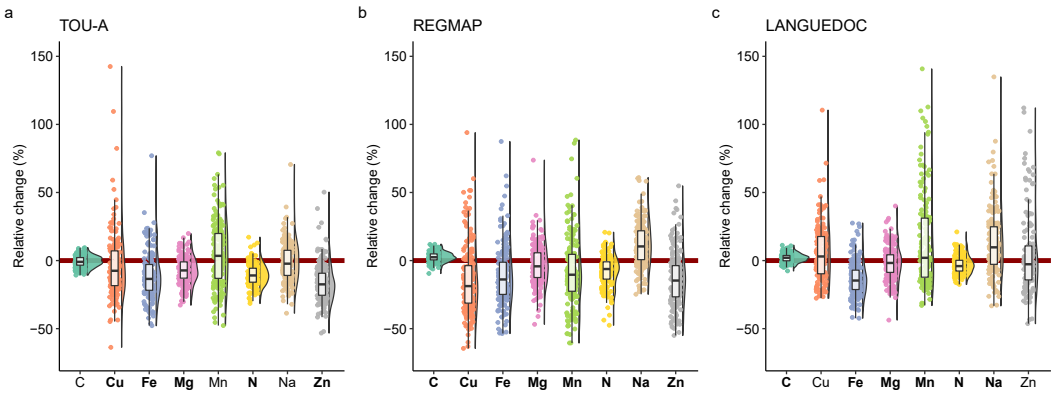


FIGURE 3.2: **Distributions of the relative changes (%) under eCO_2 of the 8 elements, in each population, and after outlier removal.** The name of the element appears in bold if the mean of the element in eCO_2 is significantly different from the mean of the element in aCO_2 (Paired wilcoxon test, significance threshold of 5%)

In order to investigate the global effect of eCO_2 on mineral composition, we performed a PCA on the relative changes in the 8 elements for the accessions of the three populations. The accessions from all populations seem to have globally similar responses to eCO_2, as suggested by the overlap of the populations in the two first principal components (Figure 3.3 a). Besides, most of the variation between accessions in term of mineral response is driven by a mechanism resulting in an inverse variation between C change and the other mineral elements (Figure 3.3 b). This captures an anticipated antagonistic trend between biomass stimulation and mineral depletion, confirming the deleterious effect of eCO_2 on Arabidopsis mineral content at a local and global scale.

Moving past this general behavior, there is a marked variability between accessions in their mineral content responses, as hinted by accessions far from central tendencies in their element changes (Figure 3.2) and accessions with extreme values in principal components 1 and 2 (Figure 3.3 a). To explore the differences between accessions in the mineral content response to eCO_2, we clustered the accessions from the REGMAP panel via a k-means approach. This multivariate clustering resulted in the partitioning of accessions in three groups, one of them showing a positive mineral content response to eCO_2 for almost all mineral elements (Figure 3.4). Indeed, plants from cluster 2 display a resilient response, with the highest relative change for almost all mineral elements, except for C content. These accessions were not found to be located in the same place among the countries of origin of the REGMAP panel. Interestingly, the fact that the accessions positively impacted by eCO_2 for one element are likely to have other mineral elements positively impacted as well promotes the view that there are general mechanisms, probably genetically driven, involved in the mineral status adaptation to eCO_2. Those general mechanisms could be additive to other mechanisms targeting specific elements like N, as the ones inferred in

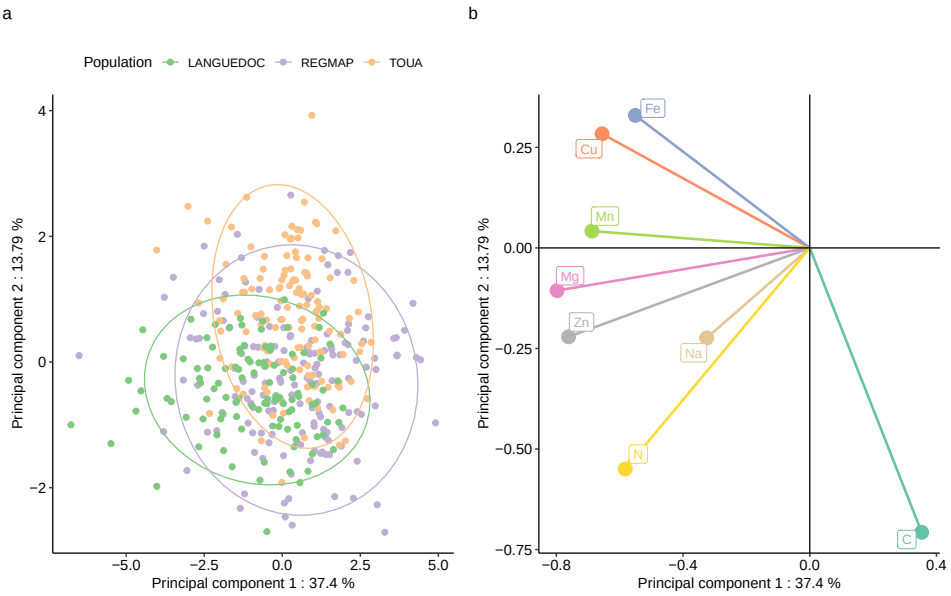


FIGURE 3.3: **PCA of the relative changes in shoot composition. a.** Accessions in the two first principal axis of the PCA. Color indicates the population of origin. **b.** Contribution (correlation) of the variables (element relative changes) to the two first principal components of the PCA.

the previous chapter.

3.2 Association models within the REGMAP population pinpoint candidate genes for the control of N, Fe and Zn accumulation in shoots

3.2.1 Association model settings and validity

In this section, we apply the state of the art association model EMMA in order to explain the variation of the mineral relative changes by means of genotype information. We focus on the REGMAP population, for which we had the earliest access in the project to genotype information at a high resolution. The 1001 Genomes project completely sequenced a large amount of Eurasian, north African and North American Arabidopsis accessions, that contained 413 accessions from the REGMAP [Alonso-Blanco et al., 2016]. However, this covers slightly less than half of our 186 REGMAP ecotypes, the rest being only genotyped with the Affymetrix 250k chip. In order to fully take advantage of the REGMAP panel, we supplemented our genotype dataset with an imputation of the complete sequence of accessions not included in the 1001 Genome project, provided by the Beagle software performing a Bayesian inference of a hidden Markov model. This tool was trained to predict the sequence of all REGMAP accessions using as ground truth the accessions from the 1001 Genomes project. This imputation was proven to be a reliable and accurate source of genotype data in existing GWAs [Arouisse et al., 2020]. During the genotype matrix preparation, duplicate SNPs and SNPs with a Minor Allele Frequency (MAF) smaller than 0.04 were removed, resulting in a total of 632694 SNPs for 186 genotypes.

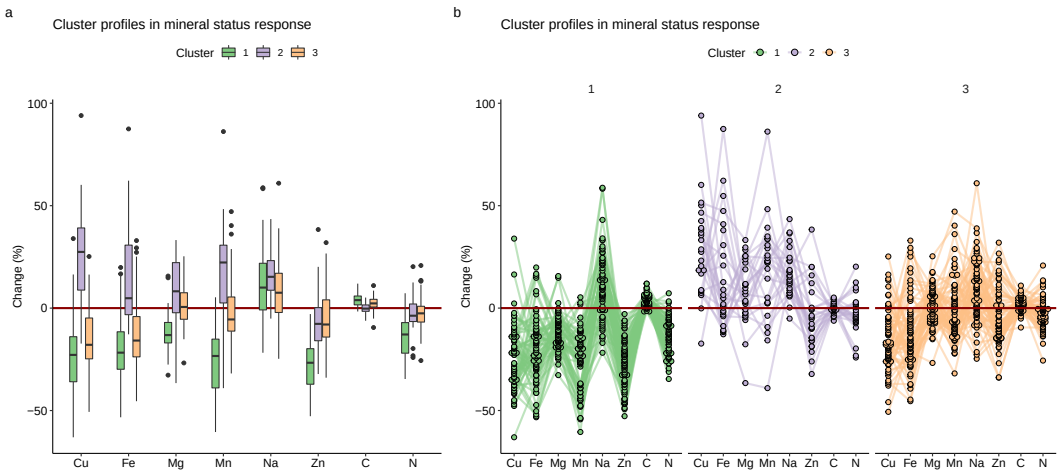


FIGURE 3.4: **Two views of the clustering of REGMAP accessions according to their relative changes in mineral content. a.** Boxplots of the relative changes in each cluster and each element. **b.** Profiles of the different clusters: each line is an accession, enabling the behavior of an accession to be followed across all elements. Cluster 1:65 accessions. Cluster 2: 25 accessions. Cluster 3: 69 accessions. The number of clusters in the k-means algorithm was chosen by the elbow method on the criteria of cluster homogeneity (within-sum of squares).

In addition to the 8 mineral elements, we added two derived phenotypes: the $\frac{N}{C}$ ratio, and the first component of the PCA on the 8 elements. We reasoned that the first principal component (PC1) was more likely to provide candidate regions involved in global mechanisms governing mineral nutrition under eCO_2. These 10 phenotypes were available in aCO_2, eCO_2, but also in their relative change version between eCO_2 and aCO_2. This resulted in 30 phenotypes on which we ran association models separately. In practice, we relied on the `statgenGWAS` R package for genotype data preparation and model estimation [van Rossum and Kruijer, 2020].

Before studying the results for the relative changes in mineral elements, we explored the associations involving mineral content under aCO_2. Using prior knowledge about N and Na nutrition in Arabidopsis allowed us to check the validity of the study. Firstly, the Na content under aCO_2 is strongly associated to SNPs falling into the gene *HKT1* (*AT4G10310*) (Figure 3.5), a well characterized sodium transporter known to control sodium accumulation in the shoot already identified through GWAs [Baxter et al., 2010]. We concurrently observe that the accessions possessing the SNP with the lowest p-value in the *HKT1* peak have a higher Na content in their shoots. Similarly, we controlled N content under aCO_2, and observed that this phenotype was in association with polymorphisms at the *NIA1* locus, an isoform of nitrate reductase. Furthermore, the accessions where the high signal variants near *NIA1* are found contain less N (Figure 3.6). The fact that some major actors of Na transport and N metabolism were found to be associated with, respectively, Na and N shoot content under reference CO_2 conditions is in accordance with existing knowledge about these genes. This demonstrates that the study design, data collection, and analysis pipelines are valid and sound enough to retrieve important genes controlling mineral nutrition.

To measure the strength of the biological signal present in the association results

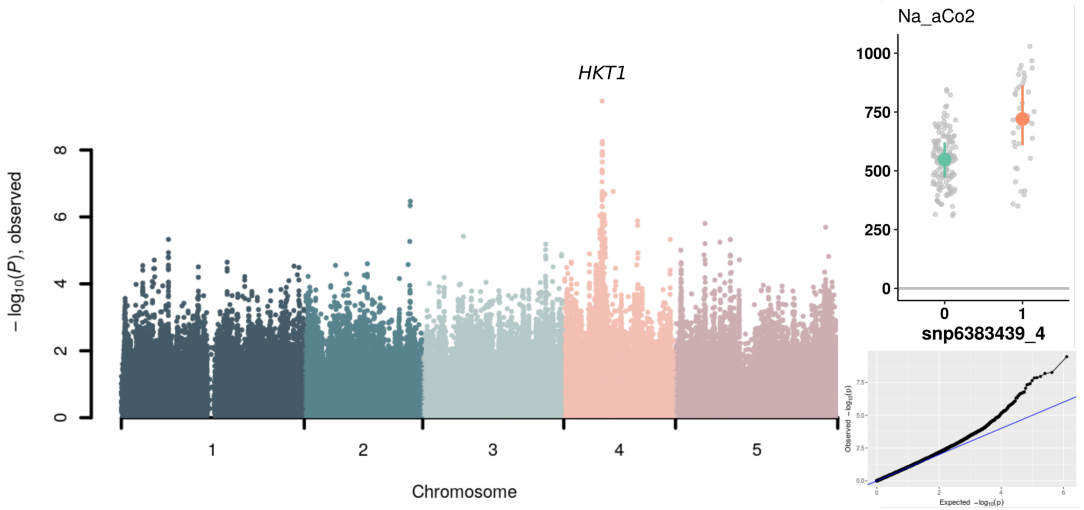


FIGURE 3.5: **Manhattan plot for Na content under aCO_2.** On the top right, the phenotype of accessions having the top SNP in the *HKT1* peak (orange) is displayed against those that do not possess this SNP (green). On the bottom right corner, the qq-plot shows observed pvalues from the LMMs on the y-axis against p-values expected without associations on the x-axis.

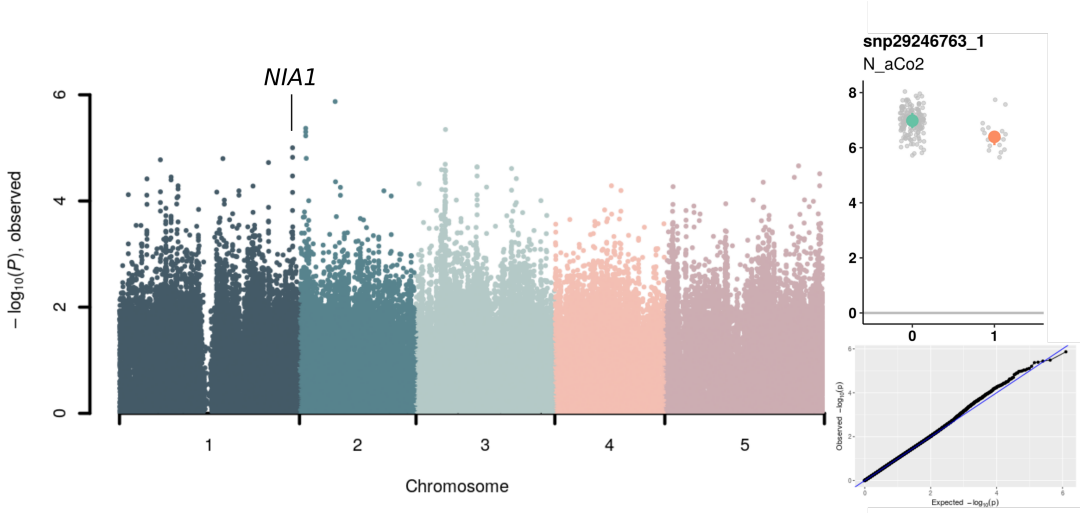


FIGURE 3.6: **Manhattan plot for N content under aCO_2.** On the top right, the phenotype of accessions having the top SNP in the *NIA1* peak (orange) is displayed against those that do not possess this SNP (green). On the bottom right corner, the qq-plot shows observed pvalues from the LMMs on the y-axis against p-values expected without associations on the x-axis.

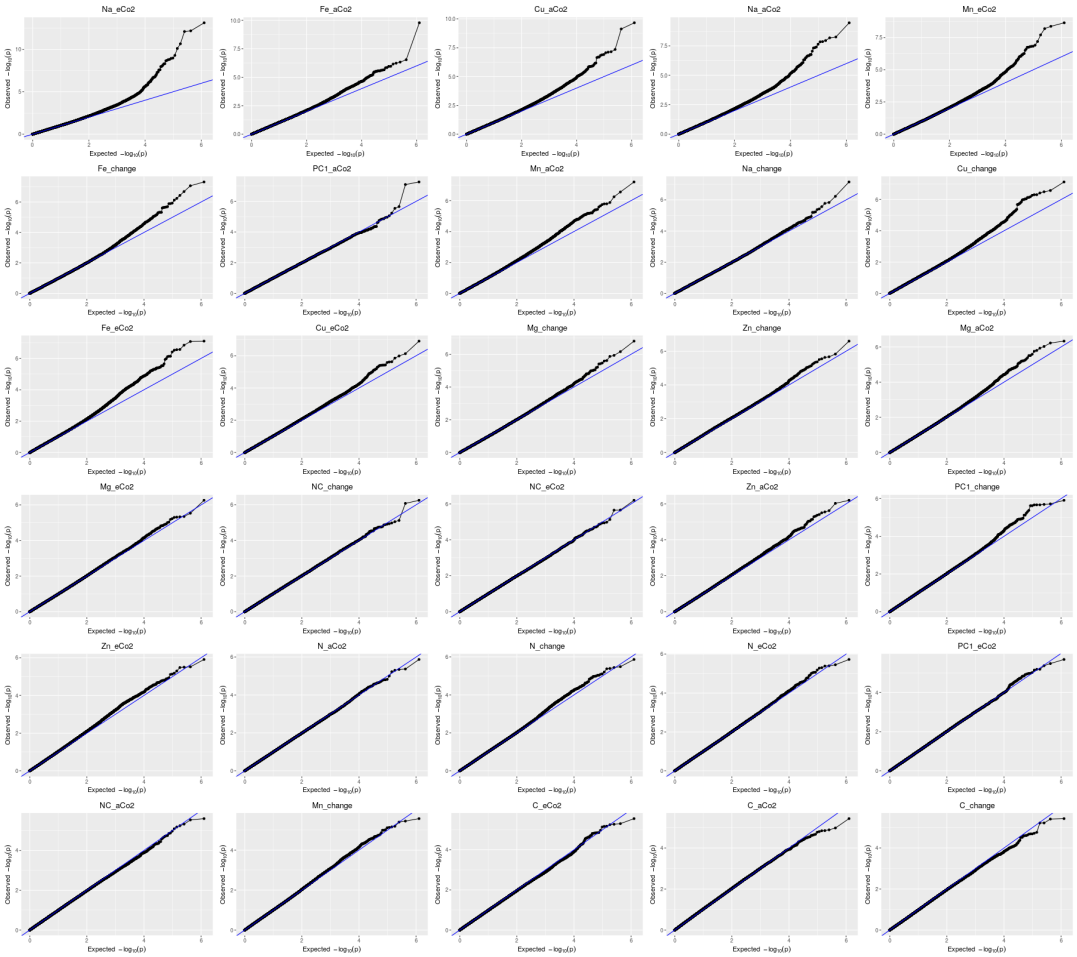


FIGURE 3.7: **Qq-plots of the 30 phenotypes**. They show the observed p-values from the LMMs on the y-axis against the p-pavlues expected without associations on the x-axis. Traits strongly associated with variants are characterized by a clear departure from the $y = x$ blue line for a few variants at the end of the curve.

of all the phenotypes, we plotted their qq-plots. They revealed that sample structure was correctly taken into account and that no covariate was neglected, but also that the degree of association varies a lot between traits (Figure 3.7). In traits of the three first lines, there are a small number of variants with very low p-values reaching a genome-wide significance level, which is expected in ideal cases of GWAs. In contrast, other traits do not display any marked departure from the null hypothesis, such as phenotypes on the three last rows. Still, such traits were not discarded from the analysis, as variants from these models can still reach a suggestive level of significance (like in the case of *NIA1* in its association to N content).

At this step, we restricted ourselves to the study of the 10 traits of relative change caused by CO_2 elevation. Indeed, *HKT1* and *NIA1* are also associated with Na and N content under eCO_2, indicating that they do not play a major role in the response of Na and N status to eCO_2. We manually explored the manhattan plots of the traits of relative changes in search for clear peaks with high p-values scores, but also peaks of moderate intensity close to genes potentially involved in mineral nutrition. By "close", we refer to a span of $\pm$ 25 kb centered on a SNP, the value of linkage desequilibrium in *Arabidopsis thaliana* being around 50 kb. We sidelined the peaks

falling into transposable elements and into pericentromeric regions. To prioritize the loci that will be further characterized functionally, we compromised between high scores of p-values and existing knowledge for genes at the vicinity of peaks. This results so far in the selection of four peaks of interest.

3.2.2 A Fe_3^+ dicitrate transport permease is a candidate gene for Fe content variation in response to eCO_2

The variation of Fe content is the trait for which we found the most intense association. Chromosome 5 harbors a strongly associated locus located near *AT5G21070*, a gene annotated as a Fe_3^+ dicitrate transport system permease. The top SNPs close to this gene are found in accessions that respond to CO_2 elevation with an increase in Fe content (Figure 3.8).

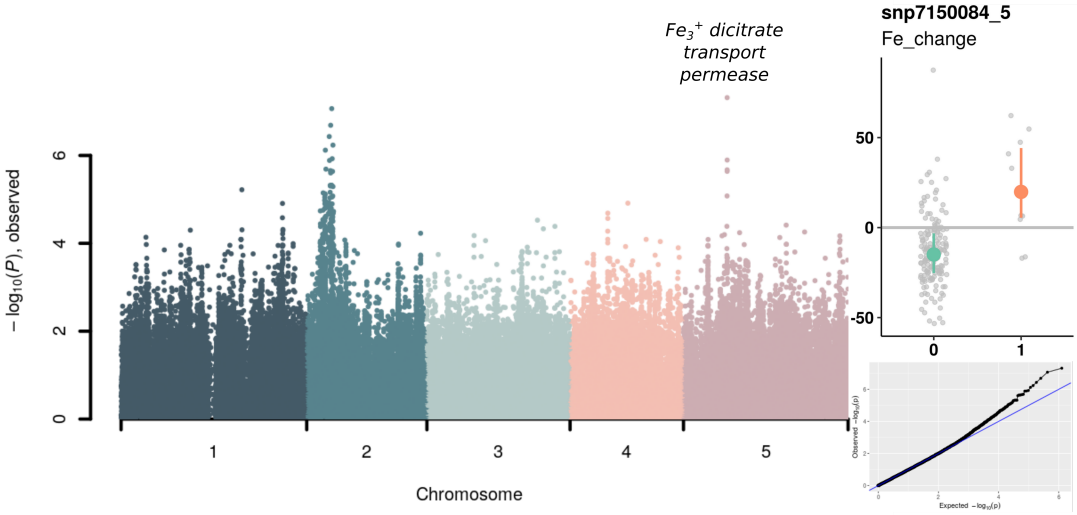


FIGURE 3.8: **Manhattan plot for Fe relative change under eCO_2.** On the top right, the phenotype of accessions having the top SNP in the Fe_3^+ dicitrate transport system permease peak (orange) is displayed against those that do not possess this SNP (green). On the bottom right corner, the qq-plot shows observed pvalues from the LMMs on the y-axis against p-values expected without associations on the x-axis. The high signal region in chromosome 2 corresponds to the pericentromeric region.

A closer look at this peak revealed that it is composed of four SNPs with high association signal. Three of them are potential expression variants: two are located in the promoter of the Fe_3^+ dicitrate transport permease and *AT5G21060*, and one is in an intronic region of *AT5G21060*. The fourth is a structural variant inside the coding region of *AT5G21080*, turning a Lysine into an Isoleucine (Figure 3.9). In addition, the intronic variant is located inside the TFBS of TFs from the MYB family (Figure 3.10).

In order to investigate the nature of the association between those variants and the change in Fe content under eCO_2, several validation experiments could be set-up, like :

- Measuring the expression of *AT5G21060* and *AT5G21070* under aCO_2 and eCO_2 in two groups of accessions: one having the haplotype of interest (composed

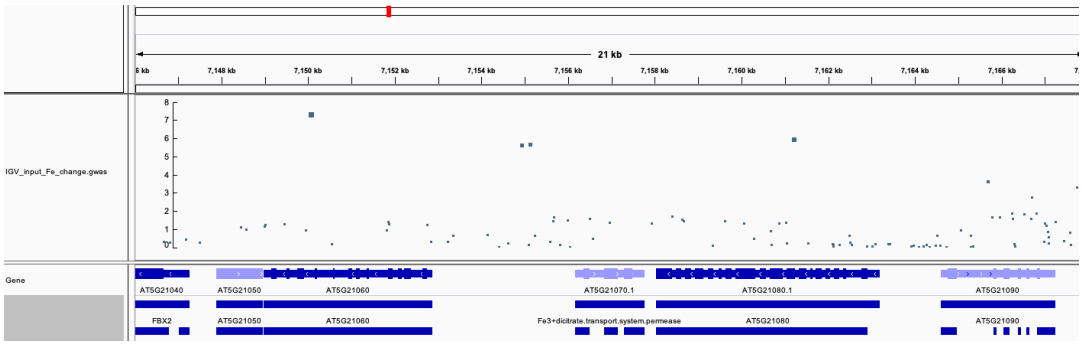


FIGURE 3.9: **Zoomed view of the SNPs strongly associated with Fe relative change.** $-log_{10}$(P) is shown as a function of chromosomic coordinates. Tracks at the bottom show gene AGIs and annotation in TAIR.

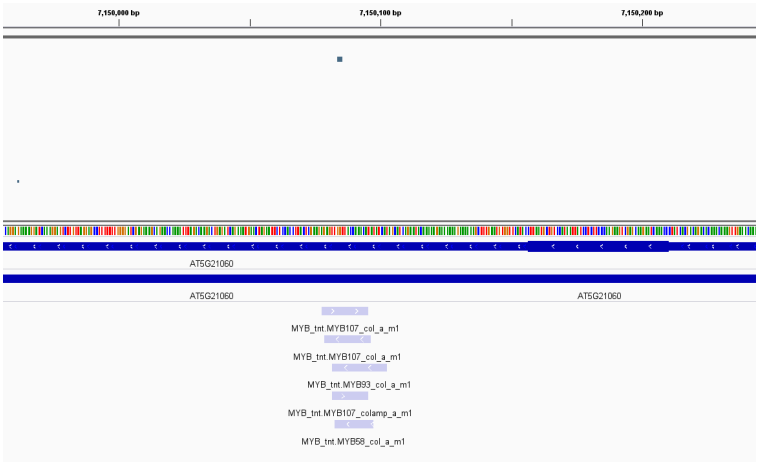


FIGURE 3.10: **Zoomed view of the SNPs strongly associated with Fe relative change, with TFBS information.** $-log_{10}$(P) is shown as a function of chromosomic coordinates. Tracks at the bottom show gene AGIs and annotation in TAIR, as well as TFBS retrieved from the Plant Cistrome database.

of the 4 top SNPs), and another group with no polymorphisms those loci. Although this would not shed light on which of the four SNPs are causal, a difference in expression change between the groups of accessions would confirm that this association is due to the regulation of the expression of *AT5G21060* or *AT5G21070*.

- Phenotyping mutant plants for *AT5G21060*, *AT5G21070* and *AT5G21080* under aCO_2 and eCO_2. If one of the mutants exhibits a different response of Fe content than the wild type, this would suggest that this candidate gene is involved in the regulation of Fe accumulation in Arabidopsis shoots under eCO_2.

The mutant lines and necessary biological material have been obtained and are now available in the team.

3.2.3 Candidate genes for the N status response to eCO_2

The relative change in N content showed less pronounced suggestive associations, but a few regions stand out for their peak shape and their closeness to genes potentially relevant to N nutrition. In particular, on chromosome three, a series of 5 variants with very low p-values are found in the promoter region of *GATA4* and its upstream neighbor *AT3G60520* (Figures 3.11 and 3.12). *GATA4* caught our attention, as it is a TF previously described to act upon N nutrition by repressing *NRT2.1*, *NRT1.1*, *GLN1.2* and *NIA1*, and modulating the phenotypic response to N limitation [Shin et al., 2017]. Accessions with polymorphisms in the *GATA4* peak have an aggravated N decline in their shoots when exposed to eCO_2 (Figure 3.11). Moreover, two of the strongly associated SNPs are inside TFBSs (Figure 3.13) belonging to TFs from the bZIP family, and to STZ, increasing the likelihood that the regulation of the expression of *GATA4* or *AT3G60520* could alter N content response to eCO_2. In particular, among the bZIP TFs with an altered TFBS we find HY5, known to be involved in C and N signalling [Bellegarde et al., 2019; Chen et al., 2016].

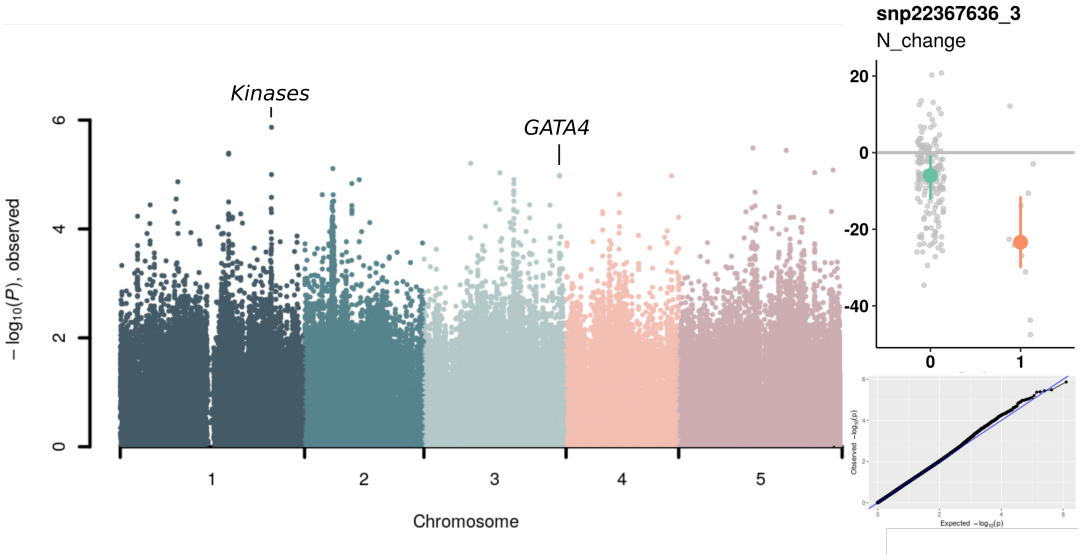


FIGURE 3.11: **Manhattan plot for N relative change under eCO_2.** On the top right, the phenotype of accessions having the top SNP in the *GATA4* peak (orange) is displayed against those that do not possess this SNP (green). On the bottom right corner, the qq-plot shows observed pvalues from the LMMs on the y-axis against p-values expected without associations on the x-axis.

In addition, the SNP with the lowest p-value genome-wide is surrounded by protein kinases (Figure 3.11) that are still uncharacterized: *AT1G66880* and *AT1G66910*. More precisely, the peak is composed of 4 high signal SNPs: two of them being non silent protein alterations in *AT1G66900* and *AT1G66910*, two of them in intronic or promoter regions. In proteomic experiments, the presence of *AT1G66880* is associated to the phosphorylation of NRT2.1 [Lejay and Schulze, unpublished]. Such markers could be experimentally tested to potentially discover new actors of N nutrition under eCO_2.

As in the previous section, the biological material to carry out validations around *GATA4* and the kinases of interest was acquired and is planned for the next phases of the project.

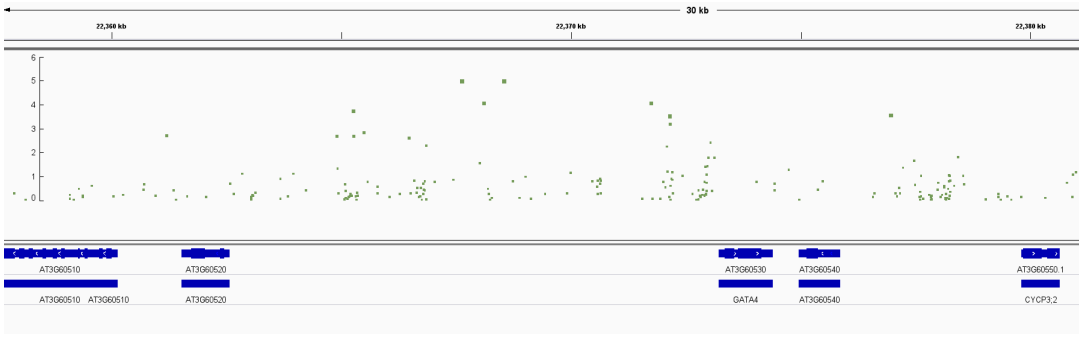


FIGURE 3.12: **Zoomed view of the SNPs strongly associated with N relative change near *GATA4*.** $-log_{10}$(P) is shown as a function of chromosomic coordinates. Tracks at the bottom show gene AGIs and annotation in TAIR.

FIGURE 3.13: **Zoomed view of the SNPs strongly associated with N relative change near *GATA4*, with TFBS information.** $-log_{10}$(P) is shown as a function of chromosomic coordinates. Tracks at the bottom show gene AGIs and annotation in TAIR, as well as TFBS retrieved from the Plant Cistrome database.

3.2.4 Strongly associated loci promote candidates genes for the control of Zn content under eCO_2

Lastly, we chose to present results regarding the associations found for Zn relative change for two reasons. First, the manhattan plot of this trait shares the same peak above the uncharacterized kinases as in the N relative change (Figure 3.14). This makes these loci even more promising in the understanding of mineral response to eCO2, as they could jointly modulate Zn and N content adaptation. Second, the strongest association is observed on the fourth chromosome, in the direct vicinity of *TIP2.2*, with a SNP in the promoter of *TIP2.2* inside the TFBS of ATHB40 and ATHB21. *TIP2.2* was recently characterized in the context of Zn distribution from roots to shoots, and in the regulation of Zn detoxification [Wang et al., 2022]. Our results strongly indicate that this gene could also be major player in the control of Zn depletion under eCO_2 in Arabidopsis. Indeed, the polymorphisms of interest close to *TIP2.2* confer to the plants a preserved Zn content under CO_2 elevation (Figure 3.14).

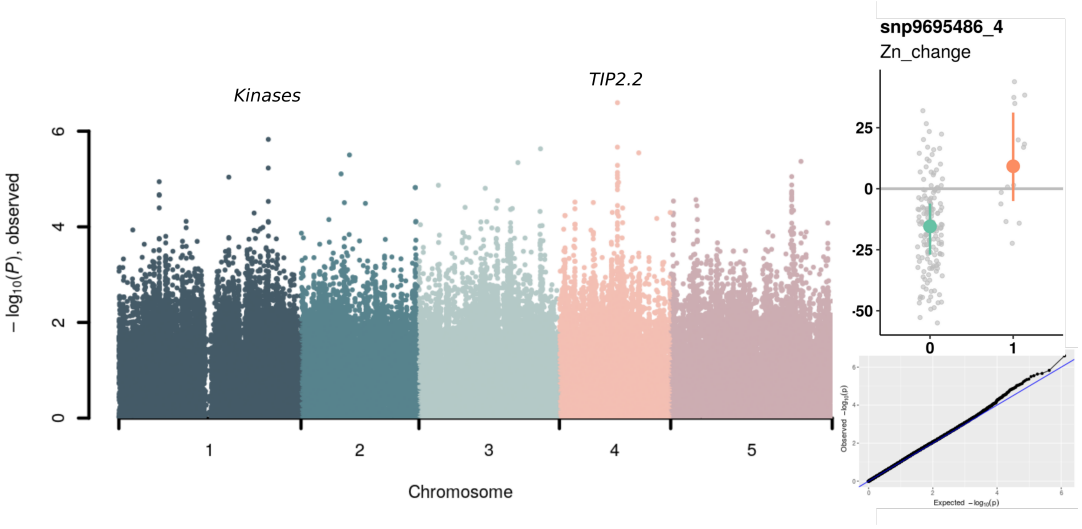


FIGURE 3.14: **Manhattan plot for Zn relative change under eCO_2.** On the top right, the phenotype of accessions having the top SNP near *TIP2.2* (orange) is displayed against those that do not possess this SNP (green). On the bottom right corner, the qq-plot shows observed pvalues from the LMMs on the y-axis against p-values expected without associations on the x-axis.

All validation experiments will be part of the project of a new PhD student commencing in october, 2022.

Chapter 4

Discussion

This manuscript summarizes our efforts to grasp genetic regulations dictating mineral nutrition under elevated CO_2. In particular, existing literature established the existence of very likely, yet unknown, regulatory mechanisms in plant roots exposed to rising CO_2 and resulting in a diminished mineral content (Publication #1). This project thus employed and developed statistical methods to model biological systems and to narrow down the search for regulatory pathways and genes involved in the mineral depletion of C3 plants under high CO_2.

4.1 Elevated CO_2 triggers gene expression reprogramming that negatively impacts mineral nutrition in Arabidopsis

4.1.1 Main physiological and molecular findings

On several levels, the joint examination of the experiments generated in this project leads to a clearer understanding of the response of Arabidopsis plants exposed to eCO_2. Firstly, **biomass was consistently increased by eCO_2 in all experiments**, whether plants were grown under abundant nitrate supply, limiting nitrate supply, limiting ammonium nitrate supply, or even iron starvation (Publication #3 and Section 2.4). This consistent increase of biomass is a strong confirmation that eCO_2 has the potential to robustly stimulate plant growth under various environmental settings, which is a promising perspective for future agricultural practices.

The fact that CO_2 increased biomass in all nutritional conditions but that N content did not decrease under abundant nitrate supply and that Fe content did not decrease at all in our combinatorial study are additional proofs that biomass stimulation and mineral content can be decoupled, and that **mineral depletion under eCO_2 is not caused by the so called "dilution effect"** alone. This is also supported by the discovery of mutant lines like *edf3* and *myb15* in which biomass stimulation is strongly altered but not N content. In addition, the dynamics of the response of biomass to a gradient of CO_2 concentrations is fairly linear while breaking points seem involved for N content (Figure 2.4.1). This decoupling between biomass stimulation and N content is already backed up by several other works [Feng et al., 2015; Wujeska-Klause et al., 2019; Myers et al., 2014b] (Publication #1). In order to further investigate those effects, it would have been interesting to also measure biomass in our panels of natural accessions and see how mineral content and biomass were linked at the population scale. This could have led to associations between biomass increase and genomic markers, similarly to a recent study that provided promising candidates in the control of growth rate under eCO_2 in Arabidopsis [Oguchi et al., 2022].

Meanwhile, the phenotypic and transcriptomic datasets we generated seem to concur on the **negative impact of eCO_2 on specific facets of plant mineral nutrition**.

The decline in N, Zn, Fe, Mg and Mn observed in the Arabidopsis populations of our GWAs is similar to what was observed in Free Air Carbon Enrichment-experiments involving diverse rice genotypes [Zhu et al., 2018] or meta-analyses encompassing several plants of agronomic interest [Myers et al., 2014a]. Moreover, the association models we fit between mineral content response and genotype point toward genes involved in different kinds of Fe or Zn transport, and a candidate regulator of N nutrition (Section 3.1). In contrast, in the plants of our combinatorial experiment, we did not see any significant decline in Fe content induced by eCO_2 (Publication #3). This could be explained by iron being more available to the roots in hydroponic culture as compared to soil substrate like in our GWAs. Still, even though the Fe shoot content did not decrease, the misregulation of key Fe nutrition genes by eCO_2 was observed in our transcriptomes. In the particular case of nitrate, **pathways associated to high affinity nitrate transport systems are especially unfavorably altered by rising CO_2**. In the transcriptomic datasets of our combinatorial study, the co-expression clustering of CO_2-responsive genes under low nitrate revealed that the responses to nitrate limitation usually observed under aCO_2 were markedly disrupted under eCO_2. Furthermore, major genes involved in nitrate transport and nitrate assimilation like *NRT2.1*, *NIR* or *GLN1.2* were repressed by eCO_2, while the expression of their negative regulators was enhanced (Publication #3). Our CO_2 gradient study confirms the over-expression of negative regulators of N uptake as CO_2 is elevated, while many of the most connected regulators of the inferred GRN of the root response to eCO_2 under low nitrate are established regulators of N acquisition and assimilation (Section 2.4) belonging for example to the NIGT or NLP families. N content in the shoots was significantly reduced by eCO_2 in our CO_2 gradient dataset for the two types of N nutrition, in the panels used to investigate natural variation, and in our combinatorial dataset under limiting nitrate supply. We also demonstrate that nitrate root uptake is diminished specifically under low nitrate, reinforcing the hypothesis that high affinity transport is negatively targeted by eCO_2 (Figure 4b in Publication #3).

4.1.2 Overview of the discovered candidate genes

GRN inference and association models predicted a list of candidate genes as important actors in the response to eCO_2, either in biomass stimulation or nutrient acquisition. Their status regarding experimental validation is also diverse, as they were obtained at different stages of the project.

1. ***MYB15*, *EDF3* and *WOX11*** were identified in an inferred GRN of the root response to eCO_2 under low nitrate in contrasted conditions of nitrate and CO_2. Those three genes control the stimulation of biomass production by eCO_2 under nitrate limitation. Their role is supported by two independent validation experiments, in which mutant plants for those genes were phenotyped, and exhibited a significantly lower biomass stimulation under eCO_2 (Publication #3).

2. The most central genes of the inferred integrative GRN in the CO_2 gradient experiment under low nitrate supply (namely ***CDF3*, *HH03*, *AT2G28810*, *HRS1*, *HB5*, *HHO2*** in Figure 2.7) are promising candidates (2.4). They were not experimentally tested yet, but their impact on biomass and N content responses could be assessed by the same experimental designs as for *MYB15*, *EDF3* and *WOX11*.

3. Strongly associated loci in the GWAs were used to identify candidate genes acting upon the relative changes in mineral content of Arabidopsis plants under eCO_2 (3.1). Those candidates are ***GATA4***, associated to N relative change, the **Fe_3^+ dicitrate transport permease**, associated to Fe relative change, ***TIP2.2***, associated to Zn relative change, two **kinases** simultaneously associated to N and Zn relative changes, as well as the direct neighbors of those genes. No wet-lab experiments to test their implication were performed yet, but they are planned and the biological material has been acquired.

4.1.3 Unifying the results between the transcriptomic datasets and the GWAs

Because our GWAs and transcriptomic datasets have the common goal to identify regulatory pathways involved in the negative regulation of mineral nutrition under eCO_2, we detail in this section the interplay between the results provided by these two branches of the project.

First, **the expression of the candidate genes from the association models** was examined in our combinatorial and CO_2 gradient transcriptomic datasets. *GATA4*, the Fe_3^+ dicitrate transport system permease are upregulated by eCO_2 in both transcriptomes, significantly in the CO_2 gradient experiment and suggestively in the combinatorial experiment. Similarly, *TIP2.2* is significantly repressed in the CO_2 gradient experiment and only suggestively in the combinatorial experiment. The neighboring genes of *GATA4*, the Fe_3^+ dicitrate transport system permease and *TIP2.2* were not differentially expressed.

Second, we find that some important **TFs identified as central nodes in inferred GRNs have their TFBS altered by strongly associated SNPs**. In the peak near the Fe_3^+ dicitrate transport system permease associated to Fe content variation, the intronic variant is located in the TFBS of several MYB TFs (Figure 3.10). One of them is *MYB93*, the 10th most connected TF of the GRN of the root response to eCO_2 under low nitrate inferred in Publication #3. Similarly, the potential expression variant in association with N change in the *GATA4* promoter falls into TFBSs of the bZIP family (Figure 3.13). *bZIP3* was among the most connected regulators in the inferred GRN of the root response to a CO_2 gradient under low nitrate (Figure 2.7), and was already documented as transcriptionally modulating the response to nitrate induction [Brooks et al., 2019]. Finally, another potential expression variant in the promoter region of *GATA4* is located in the TFBS of STZ. *STZ* is a regulator found in the inferred GRN of the root response to eCO_2 under low nitrate (Publication #3), not among the top 10 of central TFs, but it still has 11 predicted targets in this response.

However, the connections between those two parts of the project are limited by differences in experimental conditions. While the GWAs material was generated in soil substrate and under 10 mM of nitrate supply, the transcriptomic datasets and inferred GRNs were based on plants hydroponically grown and receiving 0.5 mM of nitrate. Hence, the intersection between lists of candidate genes will be restricted to pathways shared among those environmental factors.

4.1.4 Perspectives

Given the presented results, several lines of research could be followed in the future.

Firstly, **the GWAs project is still in an early stage in terms of analysis and results exploration**. Only some specific traits like Zn, N and Fe relative changes under eCO_2 could be finely investigated. To go further, the Manhattan plots of the

REGMAP panel can still be mined. For example, the exploration of the SNPs associated with the first principal component of mineral content variation could be prioritized in order to find genetic markers involved in aspecific mechanisms of mineral nutrition under eCO_2, and uncover the determinants of plants globally resilient CO_2 elevation. Moreover, the association models on the TOU-A and Languedoc populations could be performed in next phases of the project. Genotype information for the TOU-A population is available, which is however not the case for the Languedoc panel yet.

Second, **the recent application of integrative GRN inference to the root response to a gradient of CO_2 concentrations could be further developed and extended to ammonium nitrate nutrition.** GRN inference was, to date, restricted to modelling gene expression reprogramming under nitrate supply and CO_2 elevation. The same network inference could be implemented for the plants to which N was supplied in the form of ammonium nitrate. The resulting inferred GRN could be compared to the one inferred under nitrate supply in terms of network topology and central regulators, in order to pinpoint regulatory mechanisms specific to certain types of N nutrition, or those that are common. A global GRN inference including both types of N nutrition could also be performed to capture such shared regulations. Following this idea, a global GRN of the response to eCO2 could be inferred from the alliance of our combinatorial and CO2 gradient experiment, taking as input the genes robustly differentially expressed by eCO2 under low nitrate in the two experiments. In order to improve our integrative GRN inference, ATAC-Seq data could also be generated and used as additional prior information combined to TFBS to guide regression-based methods like bRF and LASSO-D3S.

Third, it would be very informative to **carry out the GRN inference of the root response to eCO_2 under low nitrate from the combinatorial dataset with TFBS integration**. This could be performed by bRF and/or LASSO-D3S and it would be interesting to see wether the candidate genes of Publication #3 are still highly connected, and if new ones emerge. More generally, in the combinatorial dataset as well as in the CO_2 gradient dataset, a more extensive study of the impact of TFBS integration though the value of α on candidate genes and network topology could offer valuable insight.

A clear direction for future research is the investigation of **the detailed physiological and molecular aspects of the CO_2 response in plants, and particularly the regulations of the high affinity transport system that we consistently report**. This can be initiated by a **more advanced functional characterization of the candidate genes** put forward in this work with the tools of molecular biology and plant physiology. These characterizations could establish their mode of operation and precise role in the regulatory pathways involved in CO_2 adaptation.

In addition, there is tangible evidence that the **root system development and plasticity** are altered by eCO_2 (Publication #1). This is also supported by the fact that three candidate genes from the GRN inferred in Publication #3 were previously associated to functions linked to the root system architecture: WOX11, AGL14 and MYB93. Furthermore, among the most connected genes of the inferred GRNs, ontologies linked to root system were enriched. This paves the way for a deeper study of the root system under eCO_2, in a reference genotype but also in mutants for our candidate genes. In addition to its interest in the context nutrient acquisition under eCO_2, the root system is a very promising field for carbon capture, as plants enhanced to sustainably store biomass in soils could be a major asset in the mitigation of atmospheric CO_2 elevation [Lynch, 2022].

Our phenotypic and transcriptomic observations leave some questions still open. We observed that phenotypic traits under eCO_2 tend to be more dispersed than under aCO_2. This could be tested by exhaustive statistical comparisons of variance between contrasted CO_2 levels in large groups of plants, such as in our GWAs dataset. Overall, this suggests that eCO_2 could cause a higher degree of variability in plant phenotypes, a phenomenon known as cryptic genetic variability. Regarding gene expression, we report a consistent over-expression of negative regulators of nitrate uptake and assimilation in all transcriptomic datasets, and a concurrent repressing of their positive regulators. However, the main nitrate transporter under low nitrate availability ***NRT2.1* is repressed by eCO_2 in our combinatorial experiment** (Figure 3a in Publication #3), **while it is over-expressed in our CO_2 gradient experiment** (Figure 2.8). This unexplained instability of the expression of *NRT2.1* under eCO_2 was confirmed in around 10 independent experiments in the team and partly hinders the construction of hypotheses for the decline of N in plants exposed to eCO_2. *NRT2.1* is a gene regulated by a multiplicity of signals like for instance oxidative stress and N status, but also light and sugars [Lejay et al., 1999]. This causes its expression to fluctuate in a 24-hours period. An explanation could be that eCO_2 modifies the oscillating pattern of *NRT2.1* in the course of the day, so that plants sampled simultaneously under aCO_2 and eCO_2 are in fact in distinct physiological phases regarding this cycle. Such mechanisms could be exposed by a time course experiment in which the expression of *NRT2.1* is measured for one or several days via non invasive means, for example by placing the luciferase gene under the control of the *NRT2.1* promoter. Such Arabidopsis lines and the necessary microscopy equipment are already available in the team.

The decorrelation between the regulation of *NRT2.1* and N content in the shoots could advocate for other mechanisms in the response to CO_2 such as an important role of efflux, or a disruption of nitrate distribution between organs. Another hypothesis is that the expression of *NRT2.1* alone does not play an essential role in the response to eCO_2, but that it is post-translationally regulated to repress nitrate uptake. Post-translational regulations have already been documented for *NRT2.1* when environmental conditions are repressive for the transport of nitrate, through the C-terminus phosphorylation of its protein [Jacquot et al., 2020]. Measuring the quantity of the NRT2.1 protein and its phosphorylation status under contrasted CO_2 levels would be an immediate way to investigate this hypothesis.

Finally, this project opens up the prospect of **translational studies**. Two research projects in the team will shift the study of the CO_2 response from Arabidopsis to plants of agronomic interest. First, GRN inference using DIANE and a LASSO-based approach applied to tomato roots under the combinatorial design was performed at the occasion of a M2 intership in biostatistics. In the GRNs inferred via both approaches, MYB15 was found among the 20 most connected regulators. This suggests that parts of the transcriptional response to eCO_2 can be conserved. Mutant lines of tomato will be acquired for the regulators *MYB15*, *EDF3* and *CDF3*, and their study under different CO_2 concentrations will provide insights on the functional role of those genes in an edible crop. Second, *Durum wheat* will be introduced in the team as a model plant to understand the response of staple crops to eCO_2. As in the first step of our GWAs project, the natural variability of the mineral content response to eCO_2 in a population of *Durum wheat* ecotypes will be studied. The examination of root system architecture of this crop exposed to eCO_2 is also planned and could be leveraged to propose solutions toward enhanced carbon capture.

4.2 Prospects for regression-based modeling and inference for systems biology

4.2.1 Summary of statistical methods used and developed

In this project, all the statistical methods chosen for candidate genes discovery have a **regression** basis. Associations between genetic markers and phenotypes are uncovered by linear regressions with variance components (LMM). The inference of GRNs, solely on the basis of expression or combined with additional prior knowledge, was composed of regression problems predicting the expression of a target gene based on the expression of regulators. In both cases however, the end goal is actually a matter of **classification**: which genetic variants are causing a phenotype and which ones are not? Among all possible transcriptional dependencies between pairs of genes, which ones are true and actually manifesting in a given response and which ones are not? The reason explaining that classification algorithms are rarely trained to directly predict causal SNPs or links between genes is that validation data is so scarce that no supervised learning framework can be properly set up. Instead, regressions are fit on available data, in our case gene expression, polymorphisms and phenotypes, and the regression models are then interpreted to perform unsupervised classification. In a GWAs, model interpretation is given by the regression coefficient p-value of the SNPs. In GRN inference, regressions are used to quantify the influence of candidate regulators on gene expression and ultimately extract relevant regulators.

While our GWAs made use of existing statistical methods, we authored two novel methodological developments for GRN inference. The first one is DIANE, a suite for transcriptomic data exploration and GRN inference in which we extended GENIE3 [Huynh-Thu et al., 2010] in order to **refine its ranking of regulatory interactions** (Publication #2). This permutation-based strategy enhances GRN precision in two organisms (Figure 6 in Publication #2), while providing **more interpretable levers** to the user through a parameter controlling network density, and another one controlling the FDR. The second development is **the implementation of bRF and LASSO-D3S, regression frameworks integrating TFBS information to guide GRN inference from expression data**. This integration is encoded into the non linear regressions of Random Forests by weighted subsampling of the regulator space during tree elongation (bRF), and by differential shrinkage into penalized generalized linear models with stability selection (LASSO-D3S). Those methods were finely studied under different strengths of data integration, and their evaluation by a series of quality criteria confirmed that TFBS integration is beneficial to GRN inference in *Arabidopsis thaliana* (Publication #4).

4.2.2 Perspectives

Direct improvements of our statistical models can be envisioned. Firstly, the LMMs in our GWAs do not **model epistasy**, a mechanism by which two or more SNPs can interact to modulate complex traits. LMMs can handle the diffuse impact of a large number of small effect loci through their additive genetic variance components, but can not model strong interaction effects between loci, which could be detected by other approaches. For example, the Multi-locus Linear Mixed Model (MLMM) strategy includes to the regression model SNPs in a forward fashion until the additional genetic variance σ_a^2 is sufficiently reduced, which is then followed by a backward

stepwise regression to eliminate the least significant SNPs [Segura et al., 2012]. Machine learning also has the potential to detect epistatic effects via the use of algorithms modelling interactions between variables, like Random Forests or Support Vector Machines [Nicholls et al., 2020; Yoosefzadeh-Najafabadi et al., 2022]. Advances in acceleration techniques for regularized least square estimation in sparse linear models can even make possible the estimation of all possible interactions between variants in so-called epistatic maps [Carré et al., 2022].

An improvement that could be brought to DIANE is a **more robust co-expression clustering**. The estimation of mixture models, as implemented by the Coseq package [Rau and Maugis-Rabusseau, 2018], can be very sensitive to parameter initialization, even with the small-EM startegy it includes. The tool DiCoExpress tackles this issue by estimating a very large number of mixture models for a narrowing range of number of clusters, which allows to choose a final model with the best Integrated Completed Likelihood [Lambert et al., 2020].

Concerning GRN inference, **the developed integrative GRN inference methods bRF and LASSO-D3S should be directly compared to their most similar existing equivalents**, namely iRafNet [Petralia et al., 2015] and the Inferelator [Skok-Gibbs et al., 2022; Miraldi et al., 2019]. bRF differs from iRafNet via its exclusive use of regulators as variables in regressions instead of all genes, and because we chose a different importance metric to score interactions. The benefit of this new importance metric, the MDA, is already demonstrated in Publication #4 (Figure S1). Like LASSO-StARS in the Inferelator, LASSO-D3S is a regularized linear regression. However, LASSO-D3S models expression counts as Poisson-distributed whereas LASSO-StARS requires normally distributed counts. Their stability selection procedure also differ, resulting in different strategies to learn the sparsity parameter λ for each gene. In the context of time-course expression data like the response to nitrate induction, the performance of bRF and LASSO-D3S could also be assessed against OutPredict [Cirrone et al., 2020]. OutPredict relies on integrative GRN inference via biased Random Forests inspired from iRafNet, but with two promising additions. First, the expression of target genes or their rate of change at a given time point is modelled as a function of the expression of the regulators at the previous time points. Second, a leave-one-out approach is used to identify the interactions between regulators and target genes that accurately predict gene expression in a time point untouched during model training. Comparing our methods to OutPredict could be a way to test the importance of modelling time in GRN inference with data integration. More generally, it would be interesting to evaluate the joint predictions of those models and combine them, as **wisdom of crowds approaches** are a promising ways of deciphering gene regulation [Schiffthaler et al., 2018; Marbach et al., 2012b].

As regression models are primarily employed in our project to extract relevant signal from large biological datasets, their **interpretability** is crucial. Many models in machine learning and even more in deep learning are considered as "black boxes", making very accurate predictions that can not be explained. In contrast, a model is said interpretable or intelligible if a human can easily comprehend its predictions and acquire valuable insights into a chosen problem via this model [Murdoch et al., 2019]. Linear models like LMMs and LASSO-D3S are easily interpretable via the coefficients associated to each input variables and their significance. Compared to other machine learning algorithms, regression trees are also interpretable because they are actually built from successive conditions on the input features that ultimately result in a prediction. However, the aggregation of trees into a forest increases the number of operations to estimate the model, adding a layer of complexity that reduces interpretability. It is thus important to **select meaningful and**

unbiased variable importance metrics for Random Forests in order to derive biological interpretations from them. We made the decision in DIANE and bRF to use the permutation based approach of the MDA instead of node purity (MDI) chosen in GENIE3 and iRafNet, because the MDA is more reliable in the case of high levels of correlation among input features [Nicodemus and Malley, 2009]. But even the MDA is not imune to dependencies between variables, and it may not theoretically converge to the desired quantity in this case. To fix this, the Sobol-MDA was proposed [Bénard, Da Veiga, and Scornet, 2021] and could replace the MDA in our GRN inference tools. In order to further improve model interpretability, alternative regression models could be used like SIRUS, a stable rule learning algorithm that exhibits many desirable properties of interpretable models: stability, simplicity and predictivity [Bénard et al., 2021].

Lastly, the ultimate form of data integration suiting this project would be **the integration of our GWAs and transcriptomic datasets together into a single model** to discover candidate genes. EGRET [Weighill et al., 2022] is a model making use of TFBS information, eQTLs, individual genotypes, protein-protein interactions and expression data. Other works have built integrative GRNs from polymorphisms and gene expression measured in the individuals of a population using regularized regressions [Kim et al., 2014]. However, eQTLs or expression data in several accessions should be available to use those methods, which is not the case in our GWAs. In contrast, the method called SIGNET does not require existing eQTL associations nor expression measured in several genotypes, but rather integrates SNPs p-values from a GWAs with a prior GRN into a global model, a Random Markov Field [Wu et al., 2017]. Bayesian inference is used to estimate this model and ultimately quantify the effect of each gene on the phenotype by testing the relevant model parameter. In the demonstration of SIGNET on human data, the prior TF-target network was constructed from TFBS information and regulatory sequences activity from the FANTOM5 project [Lizio et al., 2015], but it could be replaced by a GRN inferred by other means such as the methods developed in this project. Finally, *network-guided GWAs* are a promising area. In this category of approaches, genes are attributed an association score from a GWAs, derived from the SNP p-values from this gene or its close vicinity. Each gene is also characterized by a set of connections, given by a gene network built on prior knowledge. Several approaches can then be used to find gene modules inside this prior network enriched in association scores [Climente-González et al., 2021]. An exciting way of combining our GWAs and GRN inference work would be to apply this strategy, using as prior networks the inferred GRNs of the root response to eCO2.

4.2.3 Multidisciplinary research directions

To conclude, considerations concerning multidisciplinary aspects of systems biology can be mentioned. First, we were frequently confronted with the **gap between genome-wide models requiring a certain amount of simplifications for statistical inference purposes, and the high granularity and diversity of entities involved in the biological processes** under study. Modelling assumptions made to ensure the tractability and scalability of statistical methods can be at odds with the complexity of true biological networks, and the plurality of interactions between cellular actors. One example is the choice of the list of regulators used in GRN inference. It was in this project composed of the union between the databases PlnTFDB and AtTFDB. They contain TFs, but also other genes (like BT1/2) that can regulate gene expression even if they do not directly bind to DNA. In fact, there are many possible definitions

of what is a regulator. For example, oxidoreductase enzymes like ROXY15 are capable of interacting with the TFs TGA1/4 in order to regulate high affinity nitrate transport [Ehrary et al., 2020]. Other ROXY genes were documented as interacting with TGA1/4, suggesting that including the ROXY family to our list of regulators could be justified and envisioned for future GRN inferences. A trade-off is nonetheless necessary, as using large sets of genes as potential regulators aggravates the high dimensional setting and multi-colinearity in regressions.

An interesting perspective would be to study the **dependency between model parametrization and their application dataset**. To illustrate this, we noticed that the values of α and k in were not set to the same value when bRF was applied to the response to nitrate induction (Publication #4, $\alpha = 1$, $k = 2$) and when it was applied to the CO_2 gradient experiment (Section 2.4, $\alpha = 0.8$, $k = 1$). Indeed, TFBS-supported edges were more efficiently included to inferred GRNs in the CO_2 gradient than in the response to nitrate induction, requiring lower values of α and k to avoid the complete rejection of edges with neutral priors. A hypothesis for this variation could be that different responses are better modelled by different strengths of data integration. Another hypothesis is that the strength of data integration is influenced by the dimension of the problem, *i.e* the number of expression measures per gene n and the number of candidate regulators p, or even by the amount of noise in the data. In any case, directly providing optimal parameters to the user appears less useful that sharing important criteria and practical recipes for model parametrization. Even further, procedures to learn those parameters from the data directly could be envisioned.

Finally, data integration raises questions about **the appropriate amount of existing knowledge to use during modelling or results prioritization**. For example, our choice to manually explore Manhattan plots and prioritize genes somehow already linked to mineral nutrition is a kind of data integration based on existing gene annotation. It is however subjective and limited to current knowledge, and could be improved through the use of more automated and impartial methods processing Manhattan plot signals like the local score [Bonhomme et al., 2019] or machine learning tools performing SNP prioritization from GWAs results and other types of a priori knowledge. For example, a Random Forest-based tool extending QTG-Finder2 [Lin, Lazarus, and Rhee, 2020] found that gene structure, function and protein-protein interactions were important additional predictors for causal eQTLs in Arabidopsis. The influence of each data type could be estimated from available validated causal eQTLs and their orthologs in several plant species in plants, through feature importance metrics [Hartanto et al., 2022]. In GRN inference, data integration required explicit action levers to precisely control the contribution of each data type, like for example our parameter α (Publication #4), even though also clear indicators of model quality and biological relevance help estimating those parameters.

Data integration in our GRN inference strategies was proven useful but remains imperfect: TFBS associations can contain many false positives and the fact that not all TFBS are available introduces a form of selection bias toward previously studied TFs. This could be solved by combining gene expression to other types of data available genome-wide, like methylation or chromatin accessibility. This type of data in uncharted experimental conditions can however not be obtained from existing databases, and needs to be purposefully generated along with expression data. These are considerable motivations for building and curating **large scale matched multi-omics datasets to fuel methods development**. An example of such an initiative is The Cancer Genome Atlas (TCGA), for which data retrieval and pre-processing can even be greatly facilitated by tools like TCGAbiolinks from Bioconductor [Colaprico et al., 2015]. The development of such datasets of matched omics in plants

would offer great potential for integrative statistical inference in plant genomics.

Chapter 5

Résumé de la thèse en Français

Le manuscrit de thèse étant rédigé en anglais, il est à la demande de l'école doctorale accompagné de ce document résumant en français le travail de thèse.

Objectifs

En plus d'être à l'origine du changement climatique et de menacer les cultures par des événements climatiques extrêmes, une concentration atmosphérique élevée en CO_2 (eCO_2) peut à elle seule avoir un impact sur la physiologie des plantes. En particulier, la plupart des plantes cultivées entrent dans la catégorie des plantes C3, dans lesquelles la réaction photosynthétique est limitée par la concentration atmosphérique de CO_2. On s'attend donc à ce qu'une augmentation de la concentration atmosphérique de CO_2 entraîne une production accrue de biomasse primaire, voire une amélioration du rendement agricole en raison de cet **effet de fertilisation** [Tausz-Posch, Tausz, and Bourgault, 2019]. Néanmoins, une autre répercussion notable de l'élévation du CO_2 est source d'inquiétude: **la composition minérale des plantes C3 est appauvrie sous eCO_2** (Publication #1). Presque tous les nutriments minéraux des plantes, composant l'ionome (par exemple les éléments tels que N, P, K, S, Fe, Na, Mg, ou Zn) sont affectés, et peuvent être réduits de 5 à 25 % selon l'espèce, l'élément et les conditions. **Le nutriment dont les plantes ont besoin en plus grande quantité, l'azote (N), est souvent particulièrement affecté** [Loladze, 2014]. Cet épuisement des minéraux dans les plantes cultivées peut conduire à la consommation d'aliments contenant des quantités moindres de protéines, de vitamines ou d'oligo-éléments indispensables, et constitue une menace de malnutrition à l'échelle mondiale [Myers et al., 2017]. Bien que ces impacts négatifs aient été observés de manière consensuelle, les mécanismes expliquant cette baisse de composition minérale sous eCO_2 restent peu clairs, et la dilution des ions minéraux dans une plus grande quantité de biomasse ne peut, à elle seule, expliquer ce phénomène.

Une catégorie d'hypothèses porte sur **l'altération des mécanismes de signalisation, d'absorption et d'assimilation des nutriments, et notamment ceux de l'azote**. Même si les mécanismes de signalisation de la nutrition azotée ont été partiellement cartographiés, ils n'ont pas été étudiés sous eCO_2, et certaines données transcriptomiques montrent que les modules de signalisation pourraient être affectés. Ces changements transcriptionnels n'ont pas pu conduire à l'identification de modèles clairs, d'abord parce que trop peu de données sont actuellement disponibles dans les tissus racinaires, mais aussi parce que ces régulations peuvent être fortement influencées par des facteurs développementaux et environnementaux. Cela favorise l'hypothèse de voies de régulation complexes agissant sur l'expression des gènes clés de la nutrition azotée dans les racines sous eCO_2, avec des régulateurs inconnus orchestrant ces réseaux. A la lumière de cet état de l'art, nous avons défini plusieurs objectifs pour ce projet :

1. **Générer des données à l'échelle du génome à partir de matériel biologique, en particulier dans les racines, afin d'étudier en détail les voies de nutrition sous eCO_2.** Cela englobe des expériences transcriptomiques combinatoires en régime permanent, mais aussi des réponses adaptatives sous eCO_2, combinées à différents paramètres nutritionnels. Nous concentrons cette collecte de données sur la plante modèle *Arabidopsis thaliana*, et inférons statistiquement les réseaux de régulation de gènes (GRNs) régissant les réponses des racines au CO_2 élevé, en particulier ceux des voies nutritionnelles. Ces réseaux devraient permettre d'identifier des gènes candidats comme régulateurs clés de cette réponse.

2. **Réaliser des études d'association (GWAs) chez *Arabidopsis thaliana*.** Nous commençons par caractériser la variabilité naturelle trouvée dans l'altération du statut minéral par le CO_2 élevé de divers écotypes. Puis, sur la base des phénotypes et des informations de séquence disponibles, nous identifions les déterminants génétiques associés à la réponse du ionome sous CO_2 élevé.

Les prédictions réalisées par chacun de ces points donneront lieu à la validation expérimentale des candidats les plus prometteurs.

Pour étudier les voies nutritionnelles sous eCO_2 et réaliser l'identification des gènes candidats, nous avons eu recours à la **biologie des systèmes, et plus particulièrement à l'inférence de Réseaux de Régulation de Gènes (GRNs)**. Les GRNs sont des modèles abstraits du contrôle de la transcription et des influences mutuelles des gènes sur leurs niveaux d'expression. Dans les organismes complexes, l'inférence des GRNs est un objectif ultime mais aussi un défi non résolu car il implique de multiples couches moléculaires entrelacées et des défis statistiques importants.

Afin d'adapter et d'améliorer les solutions existantes pour l'inférence de GRN, nous avons pris la décision d'explorer les techniques basées sur la **régression**. En effet, de telles méthodes sont facilement extensibles, multivariées, et décrivent la régulation dans une formulation orientée vers la causalité. Elles ont prouvé dans des benchmarks que, même si leur performance était limitée, elle dépassait celle d'autres approches statistiques [Marbach et al., 2012b], notamment dans le cas d'ensemble d'arbres comme GENIE3 [Huynh-Thu et al., 2010].

Cependant, les méthodes basées sur la régression souffrent de certaines limites. Par exemple, l'extraction des influences régulatrices à partir des modèles de régression fournit un GRN pondéré entièrement connecté, mais la façon de seuiller ce GRN entièrement connecté n'était pas traitée dans la publication originale de GENIE3: il n'y a à ce jour aucun consensus sur la façon d'obtenir de manière optimale un GRN sparse. Compte tenu de ce potentiel d'amélioration, nous proposons dans la Publication #2 d'affiner l'inférence de GENIE3 en évaluant la significativité statistique des interactions prédites, et de comparer le gain de précision de cette procédure sur des *gold-standards* expérimentaux.

En outre, les méthodes basées sur la régression souffrent du problème de grande dimension, de forte colinéarité dans les varaibles prédictives, et leurs performances pourraient être limitées par leur utilisation exclusive de données transcriptomiques. Pour ces raisons, l'utilisation d'autres types de données omiques, non seulement pour l'évaluation des modèles, mais aussi pour leur estimation, est envisagée dans ce travail. La Publication #4 fournit un aperçu des solutions existantes pour l'intégration des données dans l'inférence basée sur la régression tout en améliorant et en explorant deux types populaires de modèles de régression pour l'inférence GRN dans

un contexte d'intégration de TFBS et d'expression. Ces approches de régression intégratives sont comparées sur la base de la précision et du rappel, de l'erreur de prédiction et de la pertinence biologique.

Principaux résultats

Développement d'un outil interfacé et sous forme de package R pour l'analyse statistique de données transcriptomiques et l'inférence de GRNs : DIANE

Avant d'interroger les données d'expression et d'inférer un GRN pour comprendre l'adaptation au CO_2 élevé chez Arabidopsis, nous avons d'abord réfléchi aux méthodes d'analyse transcriptomique qui composeraient notre pipeline. Cette réflexion a conduit au choix d'un ensemble précis d'outils.

Motivés par les analyses statistiques reproductibles, nous avons partagé notre pipeline et notre dévlopement méthodologique pour l'inférence de GRNs, via une interface utilisateur graphique déployée en ligne, qui se présente également sous la forme d'un package R : le Dashboard for the Inference and Analysis of Network from Expression data (DIANE). Cet outil est décrit en détail dans la Publication #2, résumée ci-dessous.

Les données transcriptomiques à haut débit sont souvent examinées pour découvrir de nouveaux acteurs et régulateurs d'une réponse biologique. À cette fin, des interfaces graphiques ont été développées et permettent à un large éventail d'utilisateurs de réaliser des analyses standard à partir de données RNA-seq, même avec peu d'expérience en programmation. Bien que les solutions existantes fournissent généralement des procédures adéquates pour la normalisation, l'exploration ou l'expression différentielle, des fonctionnalités plus avancées, telles que le *clustering* de gènes ou l'inférence de réseaux de régulation, manquent souvent ou ne reflètent pas l'état actuel de l'état de l'art. Nous avons développé ici une interface utilisateur appelée DIANE (Dashboard for the Inference and Analysis of Networks from Expression data) conçue pour exploiter le potentiel des jeux de données d'expression multifactorielles provenant de n'importe quel organisme grâce à un ensemble précis de méthodes. Une session de travail interactive dans DIANE permet la normalisation, la réduction de la dimensionnalité, l'expression différentielle et l'enrichissement ontologique. Le regroupement de gènes peut être effectué et exploré via des modèles de mélange configurables, et les forêts aléatoires sont utilisées pour inférer les GRNs. DIANE comprend également une nouvelle procédure pour évaluer la significativité statistique des mesures d'influence régulateur-cible basée sur des permutations pour les métriques d'importance des forêts aléatoires. Tout au long du pipeline, les rapports de session et les résultats peuvent être téléchargés pour garantir des analyses claires et reproductibles.

Nous démontrons la valeur et les avantages de DIANE en utilisant un ensemble de données récemment publiées décrivant la réponse transcriptionnelle d'*Arabidopsis thaliana* sous la combinaison de perturbations de température, de sécheresse et de salinité. Nous montrons que DIANE peut intuitivement effectuer une exploration des procédures statistiques informatives sur des données RNA-Seq, effectuer un clustering des profils d'expression génique et aller plus loin dans la reconstruction de GRNs, en fournissant des gènes candidats pertinents ou des voies de signalisation à explorer. DIANE est disponible en tant que service web, ou peut être installé et lancé localement en tant que package R.

Plus d'un an après la publication de DIANE, il est possible de faire le point sur son utilisation. DIANE a été mise à jour à plusieurs reprises afin d'ajouter quelques

améliorations fonctionnelles ou ergonomiques mineures. Elle a également été mise à jour afin de se conformer aux nouvelles versions de R et des packages dont DIANE dépend, mais aussi suite aux rapports de bugs mineurs de plusieurs utilisateurs. DIANE a reçu des contributions de code externes et a été ajouté aux portails en ligne de SouthGreen comme le Rice Genome Hub ou le Banana Genome Hub. Cet outil a également été utilisé dans l'enseignement par des étudiants en M1 de biologie végétale mais également lors de formations de recherche. Un outil interne que nous avons développé permet de visualiser les logs de l'utilisation en ligne de DIANE. Il révèle que depuis sa publication en mai 2021, il a été de plus en plus utilisé. En octobre 2022, entre 80 et 100 connexions sont reportées chaque semaine.

Un réseau de régulation des gènes révèle les caractéristiques et les régulateurs de la réponse des racines au CO_2 élevé chez Arabidopsis

Comme présenté dans l'introduction Publication #1, l'impact délétère du CO_2 élevé chez les plantes C3 est lié à la nutrition minérale de plusieurs façons. Tout d'abord, la nutrition minérale semble être négativement altérée, car la composition des feuilles de la plante diminue sous eCO_2, l'azote étant particulièrement affecté [Loladze, 2014]. Notre revue met en avant l'existence potentielle de réseaux transcriptionnels dans les racines des plantes régissant les gènes d'acquisition et de métabolisme du nitrate, mais ces réseaux n'ont pas encore été étudiés dans des transcriptomes entiers. Deuxièmement, l'effet du CO_2 élevé sur le gain de biomasse et la composition des feuilles dépend de l'état nutritionnel de la plante. Par exemple, ces réponses phénotypiques semblent être modulées par la quantité de nutriments apportés à la plante pendant sa croissance, ou par la forme de ces nutriments. Il est donc très intéressant d'étudier l'effet du CO_2 élevé en combinaison avec différentes conditions nutritionnelles pour la plante. Ceci a motivé la conception d'un jeu de données transcriptomiques combinatoires pour élucider la réponse d'eCO_2 dans des conditions d'apport suffisant et faible de deux nutriments principaux pour la plante : le nitrate et le fer (Fe).

Nous avons démontré dans la Publication #3 qu'un niveau élevé de CO_2 a un effet très modeste sur l'expression du génome des racines en cas de suffisance de nutriments, mais qu'en revanche, il entraîne des changements d'expression importants en cas de carence en nitrate et/ou en fer. De plus, nous avons démontré que le CO_2 élevé cible négativement les modules de signalisation de la carence en nitrate et en fer au niveau transcriptionnel, notamment en association avec une réduction des capacités d'absorption du nitrate à haute affinité par les racines. Nous avons enfin inféré le GRN de la réponse à un taux élevé de CO_2 sous limitation en nitrate, au moyen de la suite méthodologique DIANE (Publication #2). Cela nous a permis d'identifier et de valider expérimentalement les facteurs de transcription *MYB15*, *WOX11* et *EDF3* pour leur rôle dans la stimulation de la croissance par un taux élevé de CO_2 (Figure 5.1). Notre approche a donc bien identifié les caractéristiques et les régulateurs clés de la réponse de la plante à une concentration élevée de CO_2, dans le but de développer des cultures plus résilientes face au changement climatique.

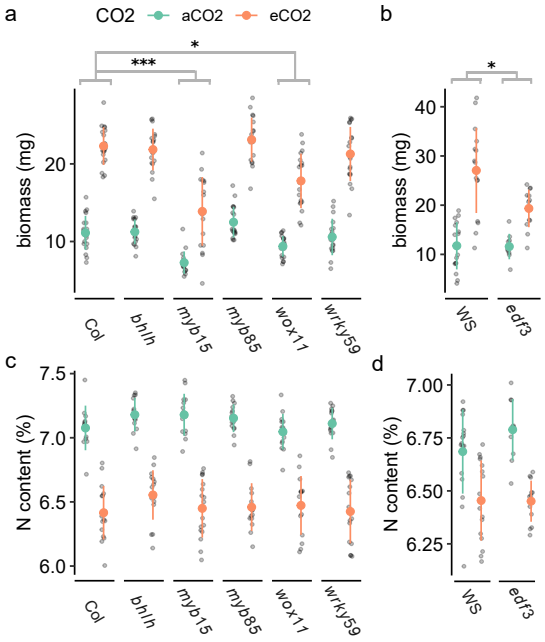


FIGURE 5.1: ***MYB15, WOX11* et *EDF3* contrôlent la stimulation de la production de biomasse par le CO_2 élevé sous limitation en nitrate**. Réponse phénotypique au CO_2 élevé des plantes présentant des mutations pour les régulateurs candidats *bhlh*, *myb15*, *myb85*, *wox11*, *wrky59* et *edf3*, par rapport à leur type sauvage relatif (Col, Columbia ; WS, Wassilewskija). Les effets d'interaction sont testés par modèles linéaires.

L'intégration des sites de fixation des facteurs de transcription (TFBSs) à des données d'expression améliore l'inférence de GRNs chez *Arabidopsis thaliana*

Les méthodes de régression sont des approches statistiques populaires et puissantes traditionnellement appliquées aux données d'expression. Plus récemment, des stratégies intégratives basées sur la régression sont apparues pour guider l'inférence des GRNs avec des données complémentaires, mais elles pourraient être adaptées pour s'adapter plus précisément aux données ou mieux modéliser la causalité. En outre, il est possible que, selon la nature du modèle de régression, les avantages de l'intégration de données varient. Pourtant, des analyses comparatives minutieuses font défaut dans la littérature.

Sur la base de la réponse temporelle à l'induction du nitrate dans les racines d'*Arabidopsis thaliana*, nous proposons d'étudier conjointement l'impact du choix du modèle pour la fonction de régression dans l'inférence GRN, et les avantages de l'intégration des TFBSs aux données d'expression. Pour ce faire, nous améliorons deux stratégies de régression prometteuses (Figure 5.2) : une version intégrative de l'algorithme Random Forest (**bRF**), et un modèle linéaire généralisé LASSO avec rétrécissement différentiel et sélection de la stabilité (**LASSO-D3S**). Nous évaluons leurs capacités de prédiction et leur précision par rapport à des gold-standards déterminés expérimentalement et répertoriés dans ConnecTF. Cette évaluation est effectuée pour une gamme de densités de réseau biologiquement pertinentes et par le biais d'un paramètre réglant finement la contribution des TFBS à l'inférence des

GRNs.
Nous concluons que l'intégration des TFBSs améliore la pertinence biologique des GRN inférés à la fois pour bRF et LASSO-D3S, et nous discutons des caractéristiques des GRNs inférés en fonction du choix du modèle. En outre, les voies de signalisation pertinentes pour la nutrition par le nitrate et attendues dans cette réponse sont modélisées de manière réaliste par bRF et LASSO-D3S, et les régulations fonctionnellement validées du transporteur de nitrate *NRT2.1* sont progressivement retrouvées à mesure que la contribution des TFBSs est renforcée. Cela souligne l'importance de nouveaux développements dans l'intégration des données multi-omiques pour l'inférence GRN basée sur la régression. Tous les scripts R pour les fonctions bRF et LASSO-D3S sont rendus disponibles.

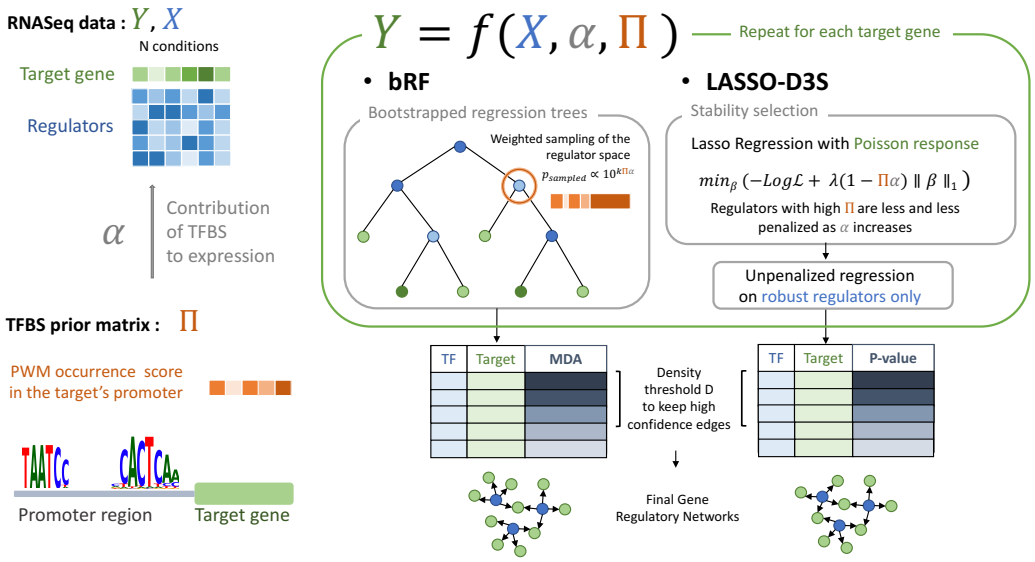


FIGURE 5.2: **Illustration des deux méthodes d'inférence de GRNs basées sur la régression et intégrant expression et TFBS : bRF et LASSO-D3S.**

La reprogrammation progressive de l'expression génétique sous gradients de CO_2 pointe vers une dérégulation vers voies de signalisation de la nutrition azotée

Les résultats de l'inférence de GRN présentés jusqu'à présent sont basés sur des transcriptomes à l'état stable dans des conditions contrastées de CO_2 et de nutrition. Il existe cependant des preuves substantielles que les mécanismes de régulation se produisent de manière adaptative, dans le temps ou le long du changement graduel d'une variable environnementale. Afin de mesurer les changements d'expression pertinents dans le contexte de l'augmentation du CO_2, nous avons généré un ensemble de données transcriptomiques de la réponse des racines à des concentrations de CO_2 progressivement croissantes. De plus, nous avons décidé d'étudier différents types de sources d'azote pour la plante : la nutrition au nitrate et au nitrate d'ammonium. Cette décision a été motivée par l'observation dans la littérature que la nutrition au nitrate provoque des réponses phénotypiques plus sévères que la nutrition au nitrate d'ammonium face à l'augmentation du CO_2. En particulier,

la nutrition à l'ammonium semble moins pénalisée en termes d'acclimatation de la photosynthèse au CO_2 élevé[Asensio, Rachmilevitch, and Bloom, 2015].

Après réalisation de l'expérience, les analyses phénotypiques ont permis de conclure que la nutrition au nitrate d'ammonium accentue l'augmentation de la stimulation biomasse, et la diminution de la teneur en N à eCO_2. Les analyses transcriptomiques des racines, et plus particulièrement une analyse de réduction de dimension ont révélé que le principal facteur influençant l'expression génétique est le type de nutrition azotée. Le deuxième facteur influençant l'expression génétique est la concentration de CO_2. De plus, la reprogrammation de l'expression semble beaucoup plus marquée en nutrition ammonium nitrate.

Sur la base de la méthodologie développée et explorée dans la Publication #4, nous avons inféré un GRN de l'adaptation des racines à l'augmentation du CO_2 sous apport en nitrate uniquement, en utilisant des données d'expression combinées aux informations TFBS. Nous avons vérifié que ce GRN avait de bonnes performances de prédiction contre des gold-standards expérimentaux, et ce dernier capture du signal biologique de manière significative comme l'ont montré des approches par permutations. Il est frappant de constater qu'un nombre élevé de gènes dans les régulateurs les plus connectés sont des régulateurs déjà connus pour leur implication dans les voies de nutrition de l'azote (Figure 5.3). Parmis ces régulateurs, CDF3 a le degré global et le degré externe les plus élevés. Il a déjà été établi qu'il contrôle la réponse à l'azote et l'efficacité d'utilisation de l'azote chez Arabidopsis et la tomate [Domínguez-Figueroa et al., 2020]. Les deuxième, quatrième et sixième TF les mieux classées appartiennent à la famille NIGT : HHO3, HRS1 et HHO2. Le rôle de ces TFs dans la régulation de l'acquisition des nutriments et notamment de l'azote a été clairement établi, ainsi que leur capacité à se lier au promoteur de leurs cibles : [Kiba et al., 2018; Safi et al., 2021]. D'autres régulateurs d'intérêts comme UIF1, RAV1 et BZIP3 [Brooks et al., 2019] sont rencontrés légèrement plus bas dans le classement des nœuds par degré global. Les 16e et 35e TF les mieux classés sont, de manière intéressante, NLP6 et NLP7, maîtres régulateurs des réseaux de nitrates chez les plantes [Marchive et al., 2013]. **Ces résultats renforcent l'hypothèse que les voies de nutrition azotée sont spécifiquement altérées par le CO_2 élevé.**

	label	gene_type	degree	degree_in	degree_out
AT3G47500	CDF3	Regulator	133	9	124
AT1G25550	HHO3	Regulator	116	14	102
AT2G28810	AT2G28810	Regulator	97	3	94
AT1G13300	HRS1	Regulator	85	2	83
AT5G65310	HB5	Regulator	74	2	72
AT1G68670	HHO2	Regulator	80	10	70
AT1G19000	AT1G19000	Regulator	73	9	64
AT4G37180	UIF1	Regulator	70	6	64
AT1G74840	AT1G74840	Regulator	61	1	60
AT1G13260	RAV1	Regulator	99	42	57
AT2G38090	AT2G38090	Regulator	63	6	57
AT5G15830	BZIP3	Regulator	52	2	50

FIGURE 5.3: **12 régulateurs les plus connectés du GRN inféré, par ordre décroissant de degré total**. Les gènes surlignés en bleu sont les régulateurs déjà caractérisés comme contrôlant les gènes de transport et d'assimilation du nitrate.

Des études d'association identifient des gènes candidats dans la réponse du ionome d'Arabidopsis sous des concentrations élevées de CO_2

Des études récentes ont examiné plusieurs lignées génétiquement diverses de riz cultivé et ont conclu que la diversité génétique pouvait être une source de variabilité dans les changements du fer, du zinc et des protéines dans des conditions de CO_2[Zhu and Ziska, 2018] élevé. L'identification des polymorphismes expliquant à cette diversité intra-espèce pourrait non seulement être un moyen de comprendre cette réponse, mais aussi de sélectionner des cultures plus résilientes. Dans ce travail, nous avons passé au crible trois populations d'écotypes de *Arabidopsis thaliana*, provenant d'échelles géographiques locales, régionales et mondiales, et caractérisé la variabilité phénotypique dans leur réponse ionomique à un CO_2 élevé. Nous avons pu confirmer un déclin global du statut minéral se produisant conjointement avec une augmentation de la teneur en carbone et identifier un sous-ensemble de lignées plus tolérantes avec une teneur en nutriments préservée. Nous avons également réalisé des GWAs sur les éléments minéraux sous CO_2 élevé pour les accessions de la population REGMAP [Horton and Bergelson, 2012], sur la base de modèles mixtes linéaires (LMM). Notre analyse a mis en évidence des haplotypes d'intérêt et des gènes candidats pour le contrôle de l'accumulation de fer, de l'azote et du zinc dans les feuilles d'Arabidopsis sous CO_2 élevé. Ces gènes candidats sont nommés en conclusion, et la Figure 5.4 illustre les résultats obtenus pour le changement relatif en fer, qui mettent en évidence une région du génome très fortement associée dans le chromosome 4.

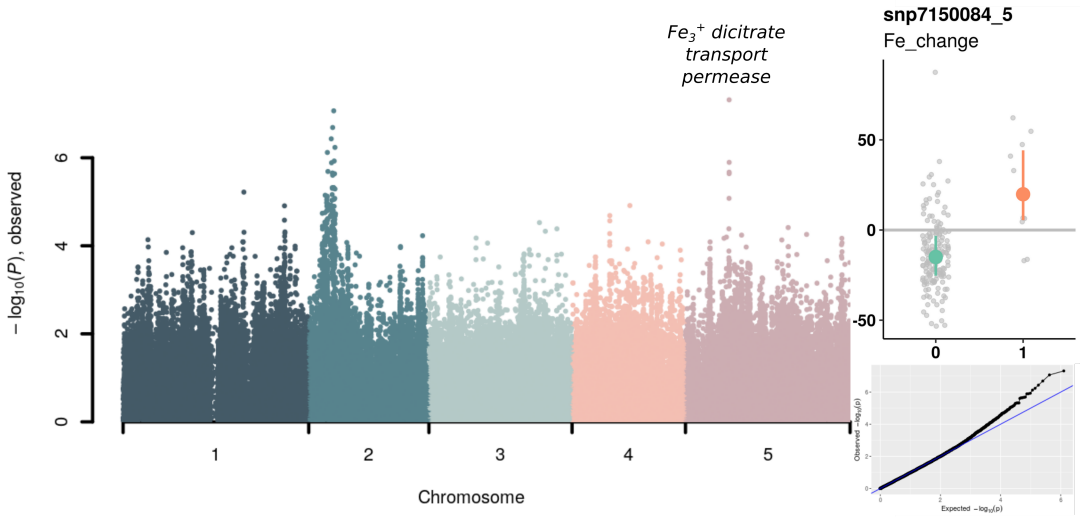


FIGURE 5.4: **Manhattan plot pour la variation relative du Fe sous eCO_2.** En haut à droite, le phénotype des accessions ayant un polymorphisme à haut signal dans le pic d'un gène candidat impliqué dans le transport du Fe_3^+ (orange) est affiché par rapport à celles qui ne possèdent pas ce variant (vert). Dans le coin inférieur droit, le diagramme montre les valeurs p observées à partir des LMM sur l'axe des ordonnées par rapport aux valeurs p attendues sans associations sur l'axe des abscisses.

Discussion et conclusion générale

Le CO_2 élevé déclenche une reprogrammation de l'expression génétique qui impacte négativement sur la nutrition minérale chez Arabidopsis

L'examen combiné des expériences réalisées dans le cadre de ce projet permet de mieux comprendre la réponse d'Arabidopsis exposée au CO_2 élevé. Tout d'abord, **la biomasse a été augmentée par le CO_2 élevé dans toutes les expériences**, que les plantes aient été cultivées sous un apport abondant de nitrate, un apport limité de nitrate, un apport limité de nitrate d'ammonium, ou même une privation de fer. Cette augmentation constante de la biomasse confirme que l'eCO_2 a le potentiel de stimuler fortement la croissance des plantes dans différents contextes environnementaux, ce qui constitue une perspective prometteuse pour les pratiques agricoles futures.

Nos mesures de quantité d'azote (N) dans diverses conditions convergent vers le fait que l'appauvrissement en minéraux sous eCO_2 n'est pas causé uniquement par le soi-disant "effet de dilution". Ceci est également confirmé par la découverte de lignées mutantes comme *edf3* et *myb15* dans lesquelles la stimulation de la biomasse est fortement altérée mais pas la teneur en azote. Ce découplage entre la stimulation de la biomasse et la teneur en N est déjà confirmé par plusieurs autres travaux [Feng et al., 2015; Wujeska-Klause et al., 2019; Myers et al., 2014b] (Publication #1).

En parallèle, les jeux de données phénotypiques et transcriptomiques générés semblent concorder sur l'impact négatif du fort CO_2 sur des facettes spécifiques de la nutrition minérale des plantes. La baisse de N, Zn, Fe, Mg et Mn observée dans les populations d'Arabidopsis de notre GWAs est similaire à ce qui a été observé dans des expériences en champs impliquant divers génotypes de riz [Zhu et al., 2018] ou des méta-analyses englobant plusieurs plantes d'intérêt agronomique [Myers et al., 2014a]. De plus, les modèles d'association que nous avons ajustés entre la réponse à la teneur en minéraux et le génotype pointent vers des gènes impliqués dans différents types de transport du Fe ou du Zn, et vers un régulateur candidat de la nutrition azotée. Ceci suggère que **l'élévation du CO_2 semble perturber les mécanismes d'absorption des minéraux**. Dans le cas particulier du nitrate, **les voies associées aux systèmes de transport du nitrate à haute affinité sont particulièrement défavorablement modifiées par l'augmentation du CO_2**. Dans les jeux de données transcriptomiques de notre étude combinatoire, le regroupement par co-expression des gènes sensibles au CO_2 sous faible teneur en nitrate a révélé que les réponses à la limitation du nitrate habituellement observées sous CO_2 ambiant étaient nettement perturbées sous eCO_2. De plus, des gènes majeurs impliqués dans le transport et l'assimilation du nitrate comme *NRT2.1*, *NIR* ou *GLN1.2* ont été réprimés par le CO_2 élevé, alors que l'expression de leurs régulateurs négatifs était augmentée (Publication #3). Notre étude du gradient de CO_2 confirme la surexpression des régulateurs négatifs de l'absorption de N lorsque le CO_2 est élevé, tandis que bon nombre des régulateurs les plus liés au GRN inféré de la réponse des racines sous CO_2 élevé et faible teneur en nitrate sont des régulateurs établis de l'acquisition et de l'assimilation de N appartenant. La teneur en N dans les feuilles a été significativement réduite par le CO_2 dans tous nos jeux de données de manière spécifique à un apport limité de nitrate. Nous démontrons également que l'absorption racinaire de nitrate est diminuée spécifiquement en cas de faible apport de nitrate, ce qui renforce l'hypothèse selon laquelle le transport à haute affinité est ciblé négativement par le CO_2 élevé.

Les modèles d'inférence de GRNs et les modèles d'associations ont prédit une liste de gènes candidats comme acteurs importants de la réponse sous CO_2 élevé, que ce soit dans la stimulation de la biomasse ou l'acquisition de nutriments. Leur statut concernant la validation expérimentale est également diversifié, car ils ont été obtenus à différentes étapes du projet.

1. ***MYB15*, *EDF3* et *WOX11*** ont été identifiés dans un GRN inféré de la réponse racinaire au CO_2 élevé sous faible nitrate dans des conditions contrastées de nitrate et de CO_2. Ces trois gènes contrôlent la stimulation de la production de biomasse par l'eCO_2 sous limitation en nitrate d'après plusieurs validations expérimentales.

2. Les gènes les plus centraux du GRN intégratif inféré dans l'expérience du gradient de CO_2 sous faible apport de nitrate (à savoir ***CDF3*, *HH03*, *AT2G28810*, *HRS1*, *HB5*, *HHO2*** sont des candidats prometteurs.

3. Les loci fortement associés dans les GWAs sont : ***GATA4***, associé au changement relatif de N, la **Fe_3^+ dicitrate transport permease**, associée au changement relatif de Fe, ***TIP2.2***, associé au changement relatif de Zn, deux **kinases** simultanément associées aux changements relatifs de N et Zn, ainsi que les voisins directs de ces gènes.

Une direction claire pour les recherches futures est l'étude des aspects physiologiques et moléculaires précis de la réponse au CO_2 chez les plantes, et particulièrement les régulations du système de transport à haute affinité que nous avons rapportées. Ceci peut être initié par une **caractérisation fonctionnelle plus avancée des gènes candidats proposés dans ce travail** avec les outils de la biologie moléculaire et de la physiologie végétale. Enfin, ce projet ouvre la perspective d'études **translationnelles**. Deux projets de recherche de l'équipe feront passer l'étude de la réponse au CO_2 d'Arabidopsis à des plantes d'intérêt agronomique : le blé dur et la tomate.

Perspectives de modélisation et d'inférence basées sur la régression pour la biologie des systèmes

Dans ce projet, toutes les méthodes statistiques choisies pour la découverte de gènes candidats ont une base de **régression**. Les associations entre les marqueurs génétiques et les phénotypes sont mises en évidence par des régressions linéaires mixtes (LMM). L'inférence de GRNs, uniquement sur la base de l'expression ou combinée à des connaissances préalables, est composée de problèmes de régression prédisant l'expression d'un gène cible sur la base de l'expression des régulateurs. Dans les deux cas cependant, l'objectif final est en fait une question de **classification** : quels sont les variants génétiques qui causent un phénotype? Parmi toutes les dépendances transcriptionnelles possibles entre paires de gènes, lesquelles sont se manifestent réellement dans une réponse donnée ? La raison pour laquelle les algorithmes de classification sont rarement entraînés à prédire directement les variants causaux ou les liens entre les gènes est que les données de validation sont si rares qu'aucun cadre d'apprentissage supervisé ne peut être correctement mis en place. Au lieu de cela, des régressions sont ajustées sur les seules données disponibles en grande quantités, dans notre cas l'expression des gènes, les polymorphismes et les phénotypes, et les modèles de régression sont ensuite interprétés pour effectuer une classification non supervisée. Dans un GWAs, l'interprétation du modèle est donnée par la p-valeur du coefficient de régression de chaque variant. Dans l'inférence

GRN, les régressions sont utilisées pour quantifier l'influence des régulateurs candidats sur l'expression des gènes et finalement extraire les régulateurs pertinents.

Alors que nos GWAs ont utilisé des méthodes statistiques existantes, nous sommes les auteurs de deux nouveaux développements méthodologiques pour l'inférence GRN. Le premier est DIANE, une suite pour l'exploration de données transcriptomiques et l'inférence de GRN dans laquelle nous avons étendu GENIE3 [Huynh-Thu et al., 2010] afin de **raffiner son classement des interactions réglementaires** (Publication #2). Cette stratégie basée sur la permutation améliore la précision du GRN dans deux organismes, tout en fournissant **des leviers plus interprétables** à l'utilisateur par le biais d'un paramètre contrôlant la densité du réseau, et d'un autre contrôlant le FDR. Le deuxième développement est **l'implémentation de bRF et LASSO-D3S, des cadres de régression intégrant les informations TFBS pour guider l'inférence GRN à partir de données d'expression**. Cette intégration est encodée dans les régressions non linéaires des forêts aléatoires par sous-échantillonnage pondéré de l'espace des régulateurs pendant l'élongation des arbres (bRF), et par rétrécissement différentiel dans des modèles linéaires généralisés pénalisés avec sélection de stabilité (LASSO-D3S). Ces méthodes ont été finement étudiées sous différentes forces d'intégration de données, et leur évaluation par une série de critères de qualité a confirmé que l'intégration des TFBS est bénéfique à l'inférence GRN chez *Arabidopsis thaliana* (Publication #4).

En termes de perspectives, **bRF et LASSO-D3S devraient être directement comparées à leurs équivalents existants les plus similaires**, à savoir iRafNet [Petralia et al., 2015] et l'Inferelator [Skok-Gibbs et al., 2022; Miraldi et al., 2019], et ce, sur les mêmes jeux de données. De plus, est important de **sélectionner des métriques d'importance des variables intelligibles et non biaisées pour les forêts aléatoires afin d'en déduire des interprétations biologiques**. Nous avons pris la décision dans DIANE et bRF d'utiliser l'approche basée sur des permutations (le MDA) au lieu de la pureté des nœuds (MDI) choisie dans GENIE3 et iRafNet, car le MDA est plus fiable dans le cas de niveaux élevés de corrélation entre les caractéristiques d'entrée [Nicodemus and Malley, 2009]. Mais même le MDA n'est pas insensible aux dépendances entre les variables, et il peut théoriquement s'écarter de la quantité désirée. Pour résoudre ce problème, le Sobol-MDA a été proposé [Bénard, Da Veiga, and Scornet, 2021] et pourrait remplacer le MDA dans nos outils d'inférence. Afin d'améliorer encore l'interprétabilité des modèles, des modèles de régression alternatifs pourraient être utilisés comme SIRUS, un algorithme d'apprentissage de règles stables qui présente de nombreuses propriétés souhaitables des modèles interprétables : stabilité, simplicité et prédictivité [Bénard et al., 2021].

Enfin, la forme ultime d'intégration de données convenant à ce projet serait **l'intégration de nos jeux de données GWAs et transcriptomiques dans un modèle unique pour découvrir des gènes candidats**. Dans cet optique, les *network-guided GWAs* constituent un domaine prometteur. Dans ces approches, les gènes se voient attribuer un score d'association à partir d'un GWAs, dérivé des p-values des variants dans ce gène ou son proche voisinage. Chaque gène est également caractérisé par un ensemble de connexions, donné par un réseau de gènes construit à partir de connaissances préalables. Plusieurs approches peuvent alors être utilisées pour trouver des modules de gènes enrichis en scores d'association à l'intérieur de ce réseau préalable. Nous pourrions combiner nos GWAs à l'inférence de GRNs en appliquant cette stratégie, en utilisant comme réseaux d'entrée les GRN inférés de la réponse des racines au CO2 élevé.

Bibliography

Ainsworth, Elizabeth A. and Stephen P. Long (2020). "30 years of free-air carbon dioxide enrichment (FACE): What have we learned about future crop productivity and its potential for adaptation?" In: *Global Change Biology* 27.1, pp. 27–49. DOI: 10.1111/gcb.15375. URL: https://doi.org/10.1111%2Fgcb.15375.

Alonso-Blanco, Carlos et al. (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana". In: *Cell* 166.2, pp. 481–491. DOI: 10.1016/j.cell.2016.05.063. URL: https://doi.org/10.1016%2Fj.cell.2016.05.063.

Alvarez, José M. et al. (2014). "Systems approach identifies TGA1 and TGA4 transcription factors as important regulatory components of the nitrate response ofiArabidopsis thaliana/iroots". In: *The Plant Journal* 80.1, pp. 1–13. DOI: 10.1111/tpj.12618. URL: https://doi.org/10.1111%2Ftpj.12618.

Alvarez, José M. et al. (2020). "Transient genome-wide interactions of the master transcription factor NLP7 initiate a rapid nitrogen-response cascade". In: *Nature Communications* 11.1. DOI: 10.1038/s41467-020-14979-6. URL: https://doi.org/10.1038%2Fs41467-020-14979-6.

Anders, S., P. T. Pyl, and W. Huber (2014). "HTSeq–a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2, pp. 166–169. DOI: 10.1093/bioinformatics/btu638. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtu638.

Araus, Viviana et al. (Apr. 2016). "Members of BTB gene family regulate negatively nitrate uptake and nitrogen use efficiency in Arabidopsis thaliana and Oryza sativa". In: *Plant Physiology*, pp.01731.2015. DOI: 10.1104/pp.15.01731. URL: https://doi.org/10.1104/pp.15.01731.

Arouisse, Bader et al. (2020). "Imputation of 3 million SNPs in the Arabidopsis regional mapping population". In: *The Plant Journal* 102.4, pp. 872–882. DOI: 10.1111/tpj.14659. URL: https://doi.org/10.1111%2Ftpj.14659.

Asensio, Jose Salvador Rubio, Shimon Rachmilevitch, and Arnold J. Bloom (2015). "Responses of Arabidopsis and Wheat to Rising COsub2/sub Depend on Nitrogen Source and Nighttime COsub2/sub Levels". In: *Plant Physiology* 168.1, pp. 156–163. DOI: 10.1104/pp.15.00110. URL: https://doi.org/10.1104%2Fpp.15.00110.

Astle, William and David J. Balding (2009). "Population Structure and Cryptic Relatedness in Genetic Association Studies". In: *Statistical Science* 24.4. DOI: 10.1214/09-sts307. URL: https://doi.org/10.1214%2F09-sts307.

Banf, Michael and Seung Y. Rhee (2017). "Computational inference of gene regulatory networks: Approaches, limitations and opportunities". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1860.1, pp. 41–52. DOI: 10.1016/j.bbagrm.2016.09.003. URL: https://doi.org/10.1016%2Fj.bbagrm.2016.09.003.

Barbosa, Sara et al. (2018). "A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints". In: *Biosystems* 174, pp. 37–48.

Bargmann, Bastiaan O.R. et al. (2013). "TARGET: A Transient Transformation System for Genome-Wide Transcription Factor Target Discovery". In: *Molecular Plant* 6.3, pp. 978–980. DOI: 10.1093/mp/sst010. URL: https://doi.org/10.1093%2Fmp%2Fsst010.

Baxter, Ivan et al. (2010). "A Coastal Cline in Sodium Accumulation in Arabidopsis thaliana Is Driven by Natural Variation of the Sodium Transporter AtHKT11". In: *PLoS Genetics* 6.11. Ed. by Gregory P. Copenhaver, e1001193. DOI: 10.1371/journal.pgen.1001193. URL: https://doi.org/10.1371%2Fjournal.pgen.1001193.

Bellegarde, Fanny, Alain Gojon, and Antoine Martin (2017). "Signals and players in the transcriptional regulation of root responses by local and systemic N signaling in Arabidopsis thaliana". In: *Journal of Experimental Botany* 68.10, pp. 2553–2565. DOI: 10.1093/jxb/erx062. URL: https://doi.org/10.1093%2Fjxb%2Ferx062.

Bellegarde, Fanny et al. (2019). "The chromatin factor HNI9 and ELONGATED HYPOCOTYL5 maintain ROS homeostasis under high nitrogen provision". In: *Plant physiology* 180.1, pp. 582–592.

Bellot, Pau et al. (2015). "NetBenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference". In: *BMC Bioinformatics* 16.1. DOI: 10.1186/s12859-015-0728-4. URL: https://doi.org/10.1186%2Fs12859-015-0728-4.

Bénard, Clément, Sébastien Da Veiga, and Erwan Scornet (2021). "MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA". In: *arXiv preprint arXiv:2102.13347*.

Bénard, Clément et al. (2021). "Interpretable random forests via rule extraction". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 937–945.

Bloom, Arnold J et al. (2010). "Carbon dioxide enrichment inhibits nitrate assimilation in wheat and Arabidopsis". In: *Science* 328.5980, pp. 899–903.

Bonhomme, Maxime et al. (2019). "A local score approach improves GWAS resolution and detects minor QTL: application to Medicago truncatula quantitative disease resistance to multiple Aphanomyces euteiches isolates". In: *Heredity* 123.4, pp. 517–531. DOI: 10.1038/s41437-019-0235-x. URL: https://doi.org/10.1038%2Fs41437-019-0235-x.

Brachi, Benjamin et al. (2013). "Investigation of the geographical scale of adaptive phenological variation and its underlying genetics in iArabidopsis thaliana/i". In: *Molecular Ecology* 22.16, pp. 4222–4240. DOI: 10.1111/mec.12396. URL: https://doi.org/10.1111%2Fmec.12396.

Breiman, Leo et al. (2017). *Classification and regression trees*. Routledge.

Broido, Anna D. and Aaron Clauset (2019). "Scale-free networks are rare". In: *Nature Communications* 10.1. DOI: 10.1038/s41467-019-08746-5. URL: https://doi.org/10.1038%2Fs41467-019-08746-5.

Brooks, Matthew D. et al. (2019). "Network Walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions". In: *Nature Communications* 10.1. DOI: 10.1038/s41467-019-09522-1. URL: https://doi.org/10.1038%2Fs41467-019-09522-1.

Brooks, Matthew D et al. (Nov. 2020). "ConnecTF: A platform to integrate transcription factor–gene interactions and validate regulatory networks". In: *Plant Physiology* 185.1, pp. 49–66. ISSN: 0032-0889. DOI: 10.1093/plphys/kiaa012. eprint: https://academic.oup.com/plphys/article-pdf/185/1/49/36389080/kiaa012.pdf. URL: https://doi.org/10.1093/plphys/kiaa012.

Campos, Adrian I. and Julio A. Freyre-González (2019). "Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total

number of genetic interactions". In: *Scientific Reports* 9.1. DOI: 10.1038/s41598-019-39866-z. URL: https://doi.org/10.1038%2Fs41598-019-39866-z.

Carré, Clément et al. (2022). "Full epistatic interaction maps retrieve part of missing heritability and improve phenotypic prediction". In: DOI: 10.1101/2022.07.20.500572. URL: https://doi.org/10.1101%2F2022.07.20.500572.

Cassan, Océane, Sophie Lèbre, and Antoine Martin (May 2021). "Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite". In: *BMC Genomics* 22.1. DOI: 10.1186/s12864-021-07659-2. URL: https://doi.org/10.1186/s12864-021-07659-2.

Chen, Shifu et al. (2018). "fastp: an ultra-fast all-in-one FASTQ preprocessor". In: *Bioinformatics* 34.17, pp. i884–i890. DOI: 10.1093/bioinformatics/bty560. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbty560.

Chen, Xiangbin et al. (2016). "Shoot-to-root mobile transcription factor HY5 coordinates plant carbon and nitrogen acquisition". In: *Current Biology* 26.5, pp. 640–646.

Chèneby, Jeanne et al. (2019). "ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments". In: *Nucleic Acids Research*. DOI: 10.1093/nar/gkz945. URL: https://doi.org/10.1093%2Fnar%2Fgkz945.

Cheng, Chia-Yi et al. (2021). "Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships". In: *Nature Communications* 12.1. DOI: 10.1038/s41467-021-25893-w. URL: https://doi.org/10.1038%2Fs41467-021-25893-w.

Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin (Dec. 2018). "Variational inference for probabilistic Poisson PCA". In: *The Annals of Applied Statistics* 12.4. DOI: 10.1214/18-aoas1177. URL: https://doi.org/10.1214/18-aoas1177.

Chiquet, Julien, Guillem Rigaill, and Martina Sundqvist (2019). "A multiattribute Gaussian graphical model for inferring multiscale regulatory networks: an application in breast cancer". In: *Gene Regulatory Networks*. Springer, pp. 143–160.

Choi, Yoonha et al. (2017). "A Poisson Log-Normal Model for Constructing Gene Covariation Network Using RNA-seq Data". In: *Journal of Computational Biology* 24.7, pp. 721–731. DOI: 10.1089/cmb.2017.0053. URL: https://doi.org/10.1089%2Fcmb.2017.0053.

Cirrone, Jacopo et al. (Apr. 2020). "OutPredict: multiple datasets can improve prediction of expression and inference of causality". In: *Scientific Reports* 10.1. DOI: 10.1038/s41598-020-63347-3. URL: https://doi.org/10.1038/s41598-020-63347-3.

Clercq, Inge De et al. (2021). "Integrative inference of transcriptional networks in Arabidopsis yields novel ROS signalling regulators". In: *Nature Plants* 7.4, pp. 500–513. DOI: 10.1038/s41477-021-00894-1. URL: https://doi.org/10.1038%2Fs41477-021-00894-1.

Climente-González, Héctor et al. (2021). "Boosting GWAS using biological networks: A study on susceptibility to familial breast cancer". In: *PLOS Computational Biology* 17.3. Ed. by Sushmita Roy, e1008819. DOI: 10.1371/journal.pcbi.1008819. URL: https://doi.org/10.1371%2Fjournal.pcbi.1008819.

Colaprico, Antonio et al. (2015). "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data". In: *Nucleic Acids Research* 44.8, e71–e71. DOI: 10.1093/nar/gkv1507. URL: https://doi.org/10.1093%2Fnar%2Fgkv1507.

Cortijo, Sandra et al. (2019). "Widespread inter-individual gene expression variability in iArabidopsis thaliana/i". In: *Molecular Systems Biology* 15.1. DOI: 10.15252/msb.20188591. URL: https://doi.org/10.15252%2Fmsb.20188591.

Crawford, Nigel M and Anthony DM Glass (1998). "Molecular and physiological aspects of nitrate uptake in plants". In: *Trends in plant science* 3.10, pp. 389–395.

De Smet, Riet and Kathleen Marchal (2010). "Advantages and limitations of current network inference methods". In: *Nature Reviews Microbiology* 8.10, pp. 717–729.

Devlin, B. and Kathryn Roeder (1999). "Genomic Control for Association Studies". In: *Biometrics* 55.4, pp. 997–1004. DOI: 10.1111/j.0006-341x.1999.00997.x. URL: https://doi.org/10.1111%2Fj.0006-341x.1999.00997.x.

Dobin, Alexander et al. (2012). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1, pp. 15–21. DOI: 10.1093/bioinformatics/bts635. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbts635.

Domínguez-Figueroa, José et al. (2020). "The Arabidopsis Transcription Factor CDF3 Is Involved in Nitrogen Responses and Improves Nitrogen Use Efficiency in Tomato". In: *Frontiers in Plant Science* 11. DOI: 10.3389/fpls.2020.601558. URL: https://doi.org/10.3389%2Ffpls.2020.601558.

Dondelinger, Frank, Sophie Lèbre, and Dirk Husmeier (2013). "Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure". In: *Machine Learning* 90.2, pp. 191–230.

Efron, Bradley et al. (2004). "Least angle regression". In: *The Annals of Statistics* 32.2. DOI: 10.1214/009053604000000067. URL: https://doi.org/10.1214%2F009053604000000067.

Ehrary, Ahmad et al. (2020). "Glutaredoxin iAtGRXS8/i represses transcriptional and developmental responses to nitrate in iArabidopsis thaliana/i roots". In: *Plant Direct* 4.6. Ed. by Ivan Baxter. DOI: 10.1002/pld3.227. URL: https://doi.org/10.1002%2Fpld3.227.

Faith, Jeremiah J et al. (2007). "Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles". In: *PLoS Biology* 5.1. Ed. by Andre Levchenko, e8. DOI: 10.1371/journal.pbio.0050008. URL: https://doi.org/10.1371%2Fjournal.pbio.0050008.

Feng, Zhaozhong et al. (2015). "Constraints to nitrogen acquisition of terrestrial plants under elevated CO 2". In: *Global change biology* 21.8, pp. 3152–3168.

Frachon, Léa et al. (2017). "Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time". In: *Nature Ecology &amp Evolution* 1.10, pp. 1551–1561. DOI: 10.1038/s41559-017-0297-1. URL: https://doi.org/10.1038%2Fs41559-017-0297-1.

Friedman, N and K Murphy (1998). "S. Russell,"Learning the Structure of Dynamic Probabilistic Networks"". In: *Proc. of the Con onUncertaintyin Artijlcial Intelligence*.

Fujita, André et al. (2007). "Modeling gene expression regulatory networks with the sparse vector autoregressive model". In: *BMC systems biology* 1.1, pp. 1–11.

Gaudinier, Allison et al. (2018). "Transcriptional regulation of nitrogen-associated metabolism and growth". In: *Nature* 563.7730, pp. 259–264. DOI: 10.1038/s41586-018-0656-3. URL: https://doi.org/10.1038%2Fs41586-018-0656-3.

Geurts, Pierre et al. (2018). "dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data". In: *Scientific reports* 8.1, pp. 1–12.

Gutiérrez, Rodrigo A. et al. (2008). "Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene iCCA1/i". In: *Proceedings of the National Academy of Sciences* 105.12, pp. 4939–4944. DOI: 10.1073/pnas.0800211105. URL: https://doi.org/10.1073%2Fpnas.0800211105.

Hartanto, Margi et al. (2022). "Prioritizing Candidate eQTL Causal Genes in Arabidopsis using Random Forests". In: *G3 Genes|Genomes|Genetics*. DOI: 10.1093/g3journal/jkac255. URL: https://doi.org/10.1093%2Fg3journal%2Fjkac255.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Haury, Anne-Claire et al. (2012). "TIGRESS: Trustful Inference of Gene REgulation using Stability Selection". In: *BMC Systems Biology* 6.1. DOI: 10.1186/1752-0509-6-145. URL: https://doi.org/10.1186%2F1752-0509-6-145.

Hayes, Wayne, Kai Sun, and Nataša Pržulj (2013). "Graphlet-based measures are suitable for biological network comparison". In: *Bioinformatics* 29.4, pp. 483–491. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts729.

Horton, Matthew W and Joy Bergelson (Jan. 2012). "Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel". In: *Nature Genetics* 44.2, pp. 212–216. DOI: 10.1038/ng.1042. URL: https://doi.org/10.1038/ng.1042.

Hu, Taishan et al. (2021). "Next-generation sequencing technologies: An overview". In: *Human Immunology* 82.11, pp. 801–811. DOI: 10.1016/j.humimm.2021.02.012. URL: https://doi.org/10.1016%2Fj.humimm.2021.02.012.

Huynh-Thu, Vân Anh et al. (2010). "Inferring regulatory networks from expression data using tree-based methods". In: *PloS one* 5.9, pp. 1–10.

IPCC (2013). "Index". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T.F. Stocker et al. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Chap. Index, 1523–1535. ISBN: ISBN 978-1-107-66182-0. DOI: 10.1017/CBO9781107415324. URL: www.climatechange2013.org.

Jacquot, Aurore et al. (2020). "NRT2. 1 C-terminus phosphorylation prevents root high affinity nitrate uptake activity in Arabidopsis thaliana". In: *New Phytologist* 228.3, pp. 1038–1054.

Kang, Hyun Min et al. (2008). "Efficient control of population structure in model organism association mapping". In: *Genetics* 178.3, pp. 1709–1723.

Kang, Hyun Min et al. (2010). "Variance component model to account for sample structure in genome-wide association studies". In: *Nature Genetics* 42.4, pp. 348–354. DOI: 10.1038/ng.548. URL: https://doi.org/10.1038%2Fng.548.

Kiba, Takatoshi et al. (2018). "Repression of nitrogen starvation responses by members of the Arabidopsis GARP-type transcription factor NIGT1/HRS1 subfamily". In: *The Plant Cell* 30.4, pp. 925–945.

Kim, Dong-Chul et al. (2014). "Inference of SNP-Gene Regulatory Networks by Integrating Gene Expressions and Genetic Perturbations". In: *BioMed Research International* 2014, pp. 1–9. DOI: 10.1155/2014/629697. URL: https://doi.org/10.1155%2F2014%2F629697.

Kim, S. Y. (2003). "Inferring gene networks from time series microarray data using dynamic Bayesian networks". In: *Briefings in Bioinformatics* 4.3, pp. 228–235. DOI: 10.1093/bib/4.3.228. URL: https://doi.org/10.1093%2Fbib%2F4.3.228.

Konishi, Mineko and Shuichi Yanagisawa (2014). "Emergence of a new step towards understanding the molecular mechanisms underlying nitrate-regulated gene expression". In: *Journal of Experimental Botany* 65.19, pp. 5589–5600. DOI: 10.1093/jxb/eru267. URL: https://doi.org/10.1093%2Fjxb%2Feru267.

Koutrouli, Mikaela et al. (2020). *A Guide to Conquer the Biological Network Era Using Graph Theory*. DOI: 10.3389/fbioe.2020.00034.

Lambert, Ilana et al. (2020). "DiCoExpress: a tool to process multifactorial RNAseq experiments from quality controls to co-expression analysis through differential analysis based on contrasts inside GLM models". In: *Plant Methods* 16.1. DOI:

10.1186/s13007-020-00611-7. URL: https://doi.org/10.1186%2Fs13007-020-00611-7.

Langfelder, Peter and Steve Horvath (2008). "WGCNA: an R package for weighted correlation network analysis". In: *BMC bioinformatics* 9.1, p. 559.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Lebre, Sophie et al. (2010). "Statistical inference of the time-varying structure of gene-regulation networks". In: *BMC systems biology* 4.1, pp. 1–16.

Lecca, Paola (2021). "Machine learning for causal inference in biological networks: Perspectives of this challenge". In: *Frontiers in Bioinformatics* 1, p. 746712.

Leclerc, Robert D. (2008). "Survival of the sparsest: Robust gene networks are parsimonious". In: *Molecular Systems Biology* 4. ISSN: 17444292. DOI: 10.1038/msb.2008.52.

Lejay, Laurence et al. (1999). "Molecular and functional regulation of two NO3–uptake systems by N-and C-status of Arabidopsis plants". In: *The Plant Journal* 18.5, pp. 509–519.

Lejay, Laurence et al. (2008). "Oxidative pentose phosphate pathway-dependent sugar sensing as a mechanism for regulation of root ion transporters by photosynthesis". In: *Plant physiology* 146.4, pp. 2036–2053.

Li, Ying, Kranthi Varala, and Gloria M. Coruzzi (2015). "From milliseconds to lifetimes: tracking the dynamic behavior of transcription factors in gene networks". In: *Trends in Genetics* 31.9, pp. 509–515. DOI: 10.1016/j.tig.2015.05.005. URL: https://doi.org/10.1016%2Fj.tig.2015.05.005.

Liang, Shoudan, Stefanie Fuhrman, and Roland Somogyi (1998). "Reveal, a general reverse engineering algorithm for inference of genetic network architectures". In: *Biocomputing*. Vol. 3.

Lin, Fan, Elena Z Lazarus, and Seung Y Rhee (2020). "QTG-Finder2: A Generalized Machine-Learning Algorithm for Prioritizing QTL Causal Genes in Plants". In: *G3 Genes|Genomes|Genetics* 10.7, pp. 2411–2421. DOI: 10.1534/g3.120.401122. URL: https://doi.org/10.1534%2Fg3.120.401122.

Liu, Han, Kathryn Roeder, and Larry Wasserman (2010). "Stability approach to regularization selection (stars) for high dimensional graphical models". In: *Advances in neural information processing systems* 23.

Lizio, Marina et al. (2015). "Gateways to the FANTOM5 promoter level mammalian expression atlas". In: *Genome biology* 16.1, pp. 1–14.

Loladze, Irakli (2014). "Hidden shift of the ionome of plants exposed to elevated CO2 depletes minerals at the base of human nutrition". In: *elife* 3, e02245.

López-Maury, Luis, Samuel Marguerat, and Jürg Bähler (2008). "Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation". In: *Nature Reviews Genetics* 9.8, pp. 583–593. DOI: 10.1038/nrg2398. URL: https://doi.org/10.1038%2Fnrg2398.

Love, Michael I., Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550. ISSN: 1474760X. DOI: 10.1186/s13059-014-0550-8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

Lüthi, Dieter et al. (2008). "High-resolution carbon dioxide concentration record 650,000–800,000 years before present". In: *nature* 453.7193, pp. 379–382.

Lynch, Jonathan P (2022). "Harnessing root architecture to address global challenges". In: *The Plant Journal* 109.2, pp. 415–431.

Maathuis, Marloes et al., eds. (2018). *Handbook of Graphical Models*. CRC Press. DOI: 10.1201/9780429463976. URL: https://doi.org/10.1201%2F9780429463976.

Maghiaoui, Amel, Alain Gojon, and Liên Bach (2020). "NRT1. 1-centered nitrate signaling in plants". In: *Journal of Experimental Botany* 71.20, pp. 6226–6237.

Marbach, Daniel et al. (2012a). "Predictive regulatory models in iDrosophila melanogaster/i by integrative inference of transcriptional networks". In: *Genome Research* 22.7, pp. 1334–1349. DOI: 10.1101/gr.127191.111. URL: https://doi.org/10.1101%2Fgr.127191.111.

Marbach, Daniel et al. (2012b). "Wisdom of crowds for robust gene network inference". In: *Nature Methods* 9.8, pp. 796–804. DOI: 10.1038/nmeth.2016. URL: https://doi.org/10.1038%2Fnmeth.2016.

Marchive, Chloé et al. (2013). "Nuclear retention of the transcription factor NLP7 orchestrates the early response to nitrate in plants". In: *Nature communications* 4.1, pp. 1–9.

Margolin, Adam A et al. (2006). "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context". In: *BMC Bioinformatics* 7.S1. DOI: 10.1186/1471-2105-7-s1-s7. URL: https://doi.org/10.1186%2F1471-2105-7-s1-s7.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: *Nucleic Acids Research* 40.10, pp. 4288–4297. DOI: 10.1093/nar/gks042.

Meinshausen, Nicolai and Peter Bühlmann (2010). "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, pp. 417–473.

Meinshausen, Nicolai and Peter Bühlmann (2006). "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34.3. DOI: 10.1214/009053606000000281. URL: https://doi.org/10.1214%2F009053606000000281.

Mercatelli, Daniele et al. (2020). "Gene regulatory network inference resources: A practical overview". In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1863.6, p. 194430.

Michailidis, George and Florence d'Alché Buc (2013). "Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues". In: *Mathematical Biosciences* 246.2, pp. 326–334. DOI: 10.1016/j.mbs.2013.10.003. URL: https://doi.org/10.1016%2Fj.mbs.2013.10.003.

Miraldi, Emily R et al. (Mar. 2019). "Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells". en. In: *Genome Res.* 29.3, pp. 449–463.

Mochida, Keiichi et al. (2018). "Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets". In: *Frontiers in Plant Science* 9, p. 1770.

Murdoch, W. James et al. (2019). "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44, pp. 22071–22080. DOI: 10.1073/pnas.1900654116. URL: https://doi.org/10.1073%2Fpnas.1900654116.

Murphy, Kevin, Saira Mian, et al. (1999). *Modelling gene expression data using dynamic Bayesian networks*. Tech. rep. Technical report, Computer Science Division, University of California ...

Myers, Samuel S. et al. (2014a). "Increasing CO2 threatens human nutrition". In: *Nature* 510.7503, pp. 139–142. DOI: 10.1038/nature13179. URL: https://doi.org/10.1038%2Fnature13179.

Myers, Samuel S et al. (2014b). "Increasing CO2 threatens human nutrition". In: *Nature* 510.7503, pp. 139–142.

Myers, Samuel S et al. (2017). "Climate change and global food systems: potential impacts on food security and undernutrition". In: *Annual review of public health* 38.

Nicholls, Hannah L. et al. (2020). "Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci". In: *Frontiers in Genetics* 11. DOI: 10.3389/fgene.2020.00350. URL: https://doi.org/10.3389%2Ffgene.2020.00350.

Nicodemus, K. K. and J. D. Malley (2009). "Predictor correlation impacts machine learning algorithms: implications for genomic studies". In: *Bioinformatics* 25.15, pp. 1884–1890. DOI: 10.1093/bioinformatics/btp331. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtp331.

O'Brien, José A. et al. (2016). "Nitrate Transport, Sensing, and Responses in Plants". In: *Molecular Plant* 9.6, pp. 837–856. DOI: 10.1016/j.molp.2016.05.004. URL: https://doi.org/10.1016%2Fj.molp.2016.05.004.

Oguchi, Riichi et al. (2022). "Enhanced growth rate under elevated CO2 conditions was observed for transgenic lines of genes identified by intraspecific variation analyses in Arabidopsis thaliana". In: *Plant Molecular Biology*, pp. 1–13.

Ohkubo, Yuri et al. (2017). "Shoot-to-root mobile polypeptides involved in systemic regulation of nitrogen acquisition". In: *Nature Plants* 3.4. DOI: 10.1038/nplants.2017.29. URL: https://doi.org/10.1038%2Fnplants.2017.29.

O'Malley, Ronan C. et al. (2016). "Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape". In: *Cell* 165.5, pp. 1280–1292. DOI: 10.1016/j.cell.2016.04.038. URL: https://doi.org/10.1016%2Fj.cell.2016.04.038.

Ota, Ryosuke et al. (2020). "Shoot-to-root mobile CEPD-like 2 integrates shoot nitrogen status to systemically regulate nitrate uptake in Arabidopsis". In: *Nature communications* 11.1, pp. 1–9.

Park, Sungjoon et al. (2018). "BTNET : boosted tree based gene regulatory network inference algorithm using time-course measurement data". In: *BMC Systems Biology* 12.S2. DOI: 10.1186/s12918-018-0547-0. URL: https://doi.org/10.1186%2Fs12918-018-0547-0.

Perković, Emilija, Markus Kalisch, and Maloes H. Maathuis (2017). *Interpreting and using CPDAGs with background knowledge*. DOI: 10.48550/ARXIV.1707.02171. URL: https://arxiv.org/abs/1707.02171.

Petralia, Francesca et al. (2015). "Integrative random forest for gene regulatory network inference". In: *Bioinformatics* 31.12, pp. i197–i205. DOI: 10.1093/bioinformatics/btv268. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtv268.

Puelma, Tomas, Rodrigo A. Gutiérrez, and Alvaro Soto (July 2012). "Discriminative local subspaces in gene expression data for effective gene function prediction". In: *Bioinformatics* 28.17, pp. 2256–2264. DOI: 10.1093/bioinformatics/bts455. URL: https://doi.org/10.1093/bioinformatics/bts455.

Pušnik, Žiga et al. (2022). "Review and assessment of Boolean approaches for inference of gene regulatory networks". In: *Heliyon*, e10222. DOI: 10.1016/j.heliyon.2022.e10222. URL: https://doi.org/10.1016%2Fj.heliyon.2022.e10222.

Qin, Jing et al. (2014). "Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods". In: *Methods* 67.3, pp. 294–303. DOI: 10.1016/j.ymeth.2014.03.006. URL: https://doi.org/10.1016%2Fj.ymeth.2014.03.006.

Rau, Andrea and Cathy Maugis-Rabusseau (2018). "Transformation and model choice for RNA-seq co-expression analysis". In: *Briefings in Bioinformatics* 19.3, pp. 425–436. ISSN: 14774054. DOI: 10.1093/bib/bbw128.

Rau, Andrea et al. (Nov. 2019). "Individualized multi-omic pathway deviation scores using multiple factor analysis". In: DOI: 10.1101/827022. URL: https://doi.org/10.1101/827022.

Robinson, Mark D and Alicia Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biology* 11.3, R25. DOI: 10.1186/gb-2010-11-3-r25. URL: https://doi.org/10.1186%2Fgb-2010-11-3-r25.

Rohart, F et al. (2017). "mixOmics: an R package for 'omics feature selection and multiple data integration". In: *biorxiv.org*. URL: http://biorxiv.org/content/early/2017/05/05/108597.

Rubin, Grit et al. (2009). "Members of the LBD family of transcription factors repress anthocyanin synthesis and affect additional nitrogen responses in Arabidopsis". In: *The plant cell* 21.11, pp. 3567–3584.

Rubio-Asensio, José S and Arnold J Bloom (2017). "Inorganic nitrogen form: a major player in wheat and Arabidopsis responses to elevated CO2". In: *Journal of Experimental Botany* 68.10, pp. 2611–2625.

Safi, Alaeddine et al. (2021). "GARP transcription factors repress Arabidopsis nitrogen starvation response via ROS-dependent and -independent pathways". In: *Journal of Experimental Botany* 72.10. Ed. by Miriam Gifford, pp. 3881–3901. DOI: 10.1093/jxb/erab114. URL: https://doi.org/10.1093%2Fjxb%2Ferab114.

Sanguinetti, Guido and Vân Anh Huynh-Thu (2019). "Gene Regulatory Networks". In.

Schiffthaler, Bastian et al. (2018). "Seiðr: Efficient Calculation of Robust Ensemble Gene Networks". In: DOI: 10.1101/250696. URL: https://doi.org/10.1101%2F250696.

Schrynemackers, Marie, Robert Küffner, and Pierre Geurts (2013). "On protocols and measures for the validation of supervised methods for the inference of biological networks". In: *Frontiers in Genetics* 4. DOI: 10.3389/fgene.2013.00262. URL: https://doi.org/10.3389%2Ffgene.2013.00262.

Seçilmiş, Deniz, Thomas Hillerton, and Erik L L Sonnhammer (2022). "GRNbenchmark - a web server for benchmarking directed gene regulatory network inference methods". In: *Nucleic Acids Research* 50.W1, W398–W404. DOI: 10.1093/nar/gkac377. URL: https://doi.org/10.1093%2Fnar%2Fgkac377.

Segura, Vincent et al. (2012). "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations". In: *Nature Genetics* 44.7, pp. 825–830. DOI: 10.1038/ng.2314. URL: https://doi.org/10.1038%2Fng.2314.

Shin, Ji Min et al. (2017). "The chimeric repressor for the GATA4 transcription factor improves tolerance to nitrogen deficiency in &lti&gtArabidopsis&lt/i&gt". In: *Plant Biotechnology* 34.3, pp. 151–158. DOI: 10.5511/plantbiotechnology.17.0727a. URL: https://doi.org/10.5511%2Fplantbiotechnology.17.0727a.

Skok-Gibbs, Claudia et al. (2022). "High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0". In: *Bioinformatics* 38.9. Ed. by Anthony Mathelier, pp. 2519–2528. DOI: 10.1093/bioinformatics/btac117. URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtac117.

Smith, Matthew R and Samuel S Myers (2018). "Impact of anthropogenic CO2 emissions on global human nutrition". In: *Nature Climate Change* 8.9, pp. 834–839.

Sun, Peng et al. (2022). "Countering elevated CO2 induced Fe and Zn reduction in Arabidopsis seeds". In: *New Phytologist*.

Tabata, Ryo et al. (2014). "Perception of root-derived peptides by shoot LRR-RKs mediates systemic N-demand signaling". In: *Science* 346.6207, pp. 343–346.

Tausz-Posch, S., M. Tausz, and M. Bourgault (2019). "Elevated [scpCO/scp sub2/sub] effects on crops: Advances in understanding acclimation, nitrogen dynamics and interactions with drought and other organisms". In: *Plant Biology* 22.S1. Ed. by L. J. De Kok, pp. 38–51. DOI: 10.1111/plb.12994. URL: https://doi.org/10.1111%2Fplb.12994.

Thorne, Thomas (2018). "Approximate inference of gene regulatory network models from RNA-Seq time series data". In: *BMC bioinformatics* 19.1, pp. 1–12.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society (Series B)* 58, pp. 267–288.

van Rossum, Bart-Jan and Willem Kruijer (2020). *statgenGWAS: Genome Wide Association Studies*. R package version 1.0.5. URL: https://CRAN.R-project.org/package=statgenGWAS.

VanRaden, P.M. (2008). "Efficient Methods to Compute Genomic Predictions". In: *Journal of Dairy Science* 91.11, pp. 4414–4423. DOI: 10.3168/jds.2007-0980. URL: https://doi.org/10.3168%2Fjds.2007-0980.

Varala, Kranthi et al. (2018). "Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants". In: *Proceedings of the National Academy of Sciences* 115.25, pp. 6494–6499. DOI: 10.1073/pnas.1721487115. URL: https://doi.org/10.1073%2Fpnas.1721487115.

Vidal, Elena A. et al. (2020). "Nitrate in 2020: Thirty Years from Transport to Signaling Networks". In: *The Plant Cell* 32.7, pp. 2094–2119. DOI: 10.1105/tpc.19.00748. URL: https://doi.org/10.1105%2Ftpc.19.00748.

Wang, Yuqi et al. (2022). "AtTIP22 facilitates resistance to zinc toxicity via promoting zinc immobilization in the root and limiting root-to-shoot zinc translocation in Arabidopsis thaliana". In: *Ecotoxicology and Environmental Safety* 233, p. 113333. DOI: 10.1016/j.ecoenv.2022.113333. URL: https://doi.org/10.1016%2Fj.ecoenv.2022.113333.

Weighill, Deborah et al. (2022). "Predicting genotype-specific gene regulatory networks". In: *Genome Research* 32.3, pp. 524–533. DOI: 10.1101/gr.275107.120. URL: https://doi.org/10.1101%2Fgr.275107.120.

Wu, Mengmeng et al. (2017). "Simultaneous inference of phenotype-associated genes and relevant tissues from GWAS data via Bayesian integration of multiple tissue-specific gene networks". In: *Journal of Molecular Cell Biology* 9.6, pp. 436–452. DOI: 10.1093/jmcb/mjx059. URL: https://doi.org/10.1093%2Fjmcb%2Fmjx059.

Wujeska-Klause, Agnieszka et al. (2019). "Lower photorespiration in elevated CO2 reduces leaf N concentrations in mature Eucalyptus trees in the field". In: *Global Change Biology* 25.4, pp. 1282–1295.

Yoosefzadeh-Najafabadi, Mohsen et al. (2022). "Machine-Learning-Based Genome-Wide Association Studies for Uncovering QTL Underlying Soybean Yield and Its Components". In: *International Journal of Molecular Sciences* 23.10, p. 5538. DOI: 10.3390/ijms23105538. URL: https://doi.org/10.3390%2Fijms23105538.

Yu, Guangchuang et al. (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters". In: *Omics: a journal of integrative biology* 16.5, pp. 284–287.

Zhang, W and F Janssen (2006). "The relationship between PR and ROC curves". In: *Darmstadt: Technische Universität Darmstadt*.

Zhu, Chunwu and Lewis H. Ziska (May 2018). "Carbon dioxide (CO 2) levels this century will alter the protein, micronutrients, and vitamin content of rice grains with potential health consequences for the poorest rice-dependent countries". In: *Science Advances* 4.5. DOI: 10.1126/sciadv.aaq1012. URL: https://doi.org/10.1126/sciadv.aaq1012.

Zhu, Chunwu et al. (2018). "Carbon dioxide (CO sub2/sub) levels this century will alter the protein, micronutrients, and vitamin content of rice grains with potential health consequences for the poorest rice-dependent countries". In: *Science Advances* 4.5. DOI: 10.1126/sciadv.aaq1012. URL: https://doi.org/10.1126%2Fsciadv.aaq1012.