Title: Cracking the code of signal interactions in *Arabidopsis.*

Authors

Gabriel Krouk[1], Daniel Tranchina[1,3], Laurence Lejay[4,1], Alexis A. Cruikshank[1], Dennis Shasha[3], Gloria M. Coruzzi[1], and Rodrigo A. Gutierrez[1,2,§].

Affiliations

1 Center for Genomics & Systems Biology, New York University, Department of Biology, New York, NY, 10003, USA

2 Departamento de Genética Molecular y Microbiología, Pontificia Universidad Católica de Chile. Alameda 340. 8331010. Santiago, Chile

3 Courant Institute of Mathematical Sciences, New York University, New York, NY, 10003, USA

Contact Information

§ To whom correspondence should be addressed. E-mail : rgutierrez@uc.cl

Additional title page Footnotes

4 Current Address: Institut de Biologie Intégrative des Plantes, UMR 5004, Biochimie et Physiologie Moléculaire des Plantes, Agro-M/CNRS/INRA/SupAgro/UM2, F-34060 Montpellier, France.

SUMMARY

Sessile organisms, such as plants, need to cope with dramatic variations in their environment. Genomic reprogramming induced by different internal or external signals has been extensively studied during the past decade. Nevertheless, very few reports have systematically studied genome-wide effects of multiple signals and their combinations on gene expression. Here, we evaluate an unprecedented level of signal integration. We determined genome-wide expression patterns under a factorial combination of Carbon (C), Light (L) Nitrogen (N) as binary factors in two Organs (O), roots and leaves, for a total of 16 different experimental conditions. In an effort to derive models that explain gene expression as a function of the explanatory factors C, N, L, O and their interactions (e.g. CO, NL, CNO) we performed analysis of variance. The statistical analysis revealed an unexpected high level of coordination between the applied factors on the global control of gene expression. Although 3^{15} different classes of models are theoretically possible for the four factors and all their interactions, all the gene expression patterns observed can be explained by a small number of models. For instance, 63% of regulated genes belong to groups of 45 or more genes with the same expression pattern (or model). The probability of this occurring by chance if genes were randomly assigned to models is below the smallest computable number ($<10^{-308}$). A second major finding is that the extent of factor interaction for C, N and L differs between organs, being of greater importance in roots as compared to shoots. In leaves, L is the major factor controlling gene expression and is largely insensitive to the influence of other factors. In contrast, in roots L is predominantly perceived as C. Surprisingly, the N effect is highly dependent on the context and is primarily detected as an interaction with C and/or L. These major factor interaction events, detected on a genome-wide scale in *Arabidopsis* are summarized in a comprehensive Boolean-model. Our combined experimental and bioinformatics approach unravelled a basic rule of signal propagation, linking the number of genes controlled by an experimental factor to the magnitude of its control on individual gene expression. The strong entanglement of signal interactions, prompted us to propose the existence of a 'Code' of signal integration.

INTRODUCTION

Living organisms need to integrate both internal and external information in order to program the appropriate responses for survival. Signalling pathways that respond to single nutrient or hormonal signals are on the way to be resolved (Baena-Gonzalez et al., 2007; Camargo et al., 2007; Castillon et al., 2007; Krouk et al., 2006; Maruyama-Nakashita et al., 2006; Muños et al., 2004; Rolland et al., 2006). However, little is known about how multiple signals are integrated on a genome-wide scale to change gene expression, physiological adjustments and/or direct new programs of development. In plants, some early clues to these molecular mechanisms come from the study of hormonal crosstalks (Achard et al., 2006; Nemhauser et al., 2004). The prevalence of multiple hormone-resistant mutants suggest that such crosstalk is very frequent (Gazzarrini and McCourt, 2003). In plant nutrition, it has been clearly established that proteins involved in glucose sensing (HXK1), nitrate sensing (NRT1.1, NRT2.1) and light signalling (HY5), are respectively involved in the crosstalk with auxin/cytokinin (Moore et al., 2003), auxin (Guo et al., 2002; Little et al., 2005; Malamy, 2005) and abscisic acid signalling (Chen et al., 2008). These crosstalks are proposed to allow regulation of growth to be tuned to nutrient availability. However, very few of the molecular elements generating crosstalk between nutritional signalling pathways themselves are known. For instance, Carbon (C), Light (L) and Nitrogen (N) signals are well known to be finely coordinated to ensure the appropriate C/N ratio needed for amino acid synthesis under a specific light regime. In particular, N transport and assimilation genes are known to be under the control of L/C/N signals (Coruzzi and Zhou, 2001). For genes encoding nitrate transporters, this C/L control can involve different C-related signalling pathways (Lejay et al., 2008). It has also been demonstrated that photosynthetic genes are under regulation by N and C (Moore et al., 2003; Rolland et al., 2002). Previous genome-wide studies have shown that C, N and CN control major cellular functions such as *energy*, *metabolism*, *C-metabolism*, and fundamental processes such as *ribosome biogenesis* (Gutierrez et al., 2007b; Palenchar et al., 2004; Price et al., 2004). Together, the evidence indicates a strong coordination between the C/N/L signals. However, the underlying mechanism(s) and models of signal integration involved in this crosstalk is/are yet to be proposed.

Recently, a bioinformatics approach was undertaken to characterize the crosstalk between seven different hormones (Nemhauser et al., 2006). By analyzing lists of hormone-responsive genes, the authors concluded that a very low level of interaction between hormone signalling pathways exists because of the small overlap between these lists. However, they do predict that the biosynthesis of each hormone is susceptible to control by others, which has been recently proven for ethylene controlled auxin synthesis (Stepanova et al., 2008; Tao et al., 2008). In our study, we used a related experimental and bioinformatics approach to evaluate interactions of nutrient and light signals using gene expression as a reporter of signal effects. Indeed, we analyzed the *Arabidopsis* transcriptome (using Affymetrix ATH1 GeneChips) under a complete factorial combination of C (Carbon), N (Nitrogen), L (Light) on two different Organs (O), roots and shoots. The response of

each gene was modelled as a function of each factor (C, N, L, O) and all possible interactions using analysis of variance (ANOVA). Thus, if a gene is controlled for instance by N and C, it constitutes a marker of convergence for signals from these two factors. By considering the whole set of regulated genes (a third of the genome), this logic allowed us to follow signal interaction on a genome-wide scale. This unique and quantitative vision of factor interactions allowed us: i) to discover an unexpected strong level of signal integration that we consider as a 'Code' of gene expression control, ii) to decipher major relationships between factors (C, N, L, O) at a genomic scale, iii) to uncover a basic rule of signal propagation, linking the number of genes controlled by a signal to the magnitude of its control on individual gene expression.

RESULTS

Genome-wide investigation of gene expression response to Carbon (C), Nitrogen (N), Light (L) and Organ (O).

We analyzed global gene expression patterns in all possible combinations of C, L and N as binary factors (presence or absence) on two different organs (leaves and roots). Plants were grown hydroponically in L/D cycles (8/16h) for six weeks, with 1mM nitrate as the N-source and without exogenous C. They were then treated for 8h with combinations of 30mM sucrose, 5mM nitrate in the light or in darkness. Roots and leaves were harvested separately and used for total RNA isolation. This corresponds to 16 different experimental conditions, including organ as a factor (Figure 1). RNA samples were used to hybridize the Arabidopsis ATH1 genome array from Affymetrix to evaluate global gene expression. All experiments were performed in duplicates. All hybridizations were normalized using the MAS*v*5.0 package and analyzed with custom made R functions. To our knowledge this constitutes the largest complete factorial design ever studied on a genome-wide scale, in plants or any other organism. To evaluate the effect of the experimental treatments on gene expression, we used ANOVA on the expression of each gene represented on the microarray. We used two different models for ANOVA analysis. The first model considers the organs as a factor, such that the expression Y_i of a $gene_i$ is given by: $Y_i = \alpha_0 + \alpha_1 C + \alpha_2 L + \alpha_3 N + \alpha_4 O + \alpha_5 CL + \alpha_6 CN + \alpha_7 CO + \alpha_8 LN + \alpha_9 NO + \alpha_{10} LO + \alpha_{11} CNL + \alpha_{12} LNO + \alpha_{13} CNO + \alpha_{14} CLO + \alpha_{15} CLNO + Z$. In this model, α_0 represents the expression under a "control" condition (without C, without N , without L, in roots); *Z* represents the noise; and α_1 to α_{15} represent the coefficients quantifying the effect of each factor (C, N, L, O) or combination of factors. For example, the coefficient of CNL represents the effect of C, N and L in combination, over and above the main effects of C, N, L and O, and all 2-way interactions among these factors. The second model is just a simplified version of the first model in which gene expression in roots and leaves datasets were analyzed separately: $Y_i = \alpha_0 + \alpha_1 C + \alpha_2 L + \alpha_3 N + \alpha_4 CL + \alpha_5 CN + \alpha_6 LN + \alpha_7 CNL + Z$. The two modeling approaches were used as they highlight three different aspects of the data (1- Whole data set, 2- Leaves only, 3-Roots only). Indeed, we found that the O (organ) effect is a predominant factor that controls gene expression (see below), and that its dramatic effect on gene expression can mask the weaker effects of other factors. On the other hand, analysis of the whole dataset provides an insight into how the O factor is integrated and how it influences the other factors. Results of modeling are provided as Table S1 for the whole dataset, Table S2 for leaves and Table S3 for roots. These tables summarize the significant coefficients for each factor or combination of factors in the model for each gene and constitute the basis for further analyses. From the modeling using the entire dataset, 8,036 genes (35% of the genome) were found to be significantly controlled by at least one factor or combination of the four factors (C, L, N, O). We found 3,279 (14.3%) and 1,002 (4.4%) genes regulated by at least one factor (C, N, L) or combination of factors in leaves and roots respectively.

A 'Code' of signal interaction.

To understand the global patterns of response to the experimental factors, we simplified the matrixes with the gene expression models described in the previous section using a binary code. We replaced coefficients that were negative, not significant or positive with a -1, 0 or 1, respectively. Thus, genes harbouring the same expression pattern (successions of 0, 1 or -1) could be grouped in the same *model of regulation*. Considering the whole data set, a gene can be either induced, repressed or not affected by 15 terms (C, L, N, O, CL, CN, CO, LN, NO, LO, CNL, LNO, CNO, CLO, CLNO) derived from the combinations of the 4 factors (C, L, N, O) and their 1^{st} ,2^{nd} , and/or 3^{rd} order interactions. Thus, a gene can respond in any one of 3^{15} = 14,348,907 possible ways. Our global analysis led to the surprising result that a very large number of genes are controlled by a very small number of *models of regulation* (Figure 2, Table 1 as truncated version; Table S4 as full version). For instance, we found 6,422 out of the 8,036 regulated genes (79.9%) explained by only 87 of the 3^{15} possible models of gene regulation. This result suggests there is a major constraining structure in plant cell signalling pathways. We thus hypothesize the existence of a 'Code' governing signal integration at the organism level responsible for the observed global gene expression reprogramming in response to C, N and L in two different organs.

Cracking the code.

To elucidate the code controlling regulation of gene expression by the experimental factors and their interactions, we developed two approaches. The first one is based on clustering across the three matrixes described above (all data, root, shoot). This method adapted from Speed (2003), enables qualitative analysis of the co-occurrence of each term in the models of gene expression (Figure 3 A to E). The second method uses the Sungear software (Poultney et al., 2007) to quantitatively evaluate the importance of each term, as assessed by the number of genes, in the models of gene expression (Figure 3 F to H). We used average linkage hierarchical cluster analysis with euclidean distance on the simplified matrix of regulatory models (Table S4). To do this, we multiplied each column in Table 1 by the number of genes with the corresponding *model* (last column in Table 1) to weight each row proportionally to the number of genes. The dendrograms generated by the clustering algorithm allowed us to infer the relationship between the signals (Speed, 2003) in the control of gene expression (Figure 3A to 3E). To evaluate the signal (as defined in Figure 2A) strength as determined by the number of genes controlled by each signal we used Sungear, a software tool designed for dynamic analysis of multiple lists of genes (Gutierrez et al., 2007a; Poultney et al., 2007). The combined hierarchical clustering and Sungear analysis revealed that O is the predominant factor controlling gene expression (Figure 3A and 3F). In leaves, the main signal is L (Figure 3B, 3D and 3G) while in roots the L effect manifests as an interaction with C (Figure 3C, 3E, 3H). That is, genes controlled by L in

leaves do not typically respond to other signals, but in roots genes controlled by L are also largely controlled by C. This logic can be used to decipher the relationships and strengths of any of the signals (Figure 3).

In a second analysis, we used hierarchical clustering analysis on the model coefficients (Tables S1; S2; S3). In this case, we grouped signals based both on their relationship and magnitude of their effect on gene expression. Interestingly, the hierarchy of signals in this analysis seems to be comparable to our first analysis based on model size (compare dendrograms in Figure 3 and Figure S1). This finding suggested that for a given signal, its strength on individual gene regulation and the number of genes in the genome that are controlled by this signal are correlated. To test this hypothesis, we plotted the absolute values of the model coefficient (an indicator of the strength of regulation irrespective of induction or repression) against the number of genes controlled by each individual signal (Figure 4). We observed a logarithmic relationship linking these two parameters at the whole dataset level (R^2= 0.50) and at the organ specific level (R^2=0.82) (Figure 4). Note here, that logarithmic regression excluding the L signal in leaves is still very significant (R^2=0.74). The two terms with the largest coefficient and number of genes (top right sector of the graph), C and L, seem to be the ones which behave the most differently in the roots and leaves datasets. Treating data from root and leaves separately allowed us to reduce this constraint and improved the regression. Thus, if we sort signals by their availability to control gene expression, two components can be identified. The first component encompasses weaker interactions, controlling few genes (< 500 genes). In this component, the strength of the signal increases without a concomitant increase in the number of genes regulated. In the second component (>500 genes), we observe the inverse relationship. The strength of the regulation reaches a 'plateau' (at a value of approximately 450 in the coefficients), but there is a large increase in the number of regulated genes (Figure 4).

The rules of signal integration

To reach a better understanding of how plants respond and integrate multiple experimental factors, we looked at the number of genes controlled by *x* number of signals (as defined in Figure 1B). This revealed that signal integration is stronger in roots than in leaves (Figure 5A). In leaves the large majority (89.8%) of genes are controlled by only one factor whereas in roots, 86.2% of genes are controlled by 2 or more factors (Figure 5A). To decipher the relationships underlying the dichotomy between leaves and roots, gene lists corresponding to each group (a to h) were subjected to hierarchical clustering (Figure 5 B-C). This approach demonstrated for instance that in leaves, 99.6% of the 89% of the genes controlled by only one signal (Fig 5B-a), are controlled by L, and thus L signalling in leaves is mainly independent of the other considered signals. In roots, genes with simple models with one significant term, also show a dominant of L (78% are induced by light only) (Figure 5C-e). However, as the models are more complex (Figure 5B e towards h), L and C appear related (compare Figures 5C e to h). Furthermore, the effect of N is mainly observed as an interaction with C and L, indicating that the effect of N is largely dependent on the context of the other signals.

To confirm this hypothesis and decipher signal cross-talks, we analyzed the number of genes controlled by a given factor (C, N, L or O) alone, and the subsequent effect of adding *x* signals (Figure 6). This approach provides information of how signals superimpose to control gene expression at the whole plant level (Figure 6A), or at the organ-specific level (Figure 6B). We found that 78% of genes with organ-specific expression (Organ effect) are also controlled by at least 1 other signal (Figure 6A), leading to the conclusion that signals such as C, L or N are differently transduced in roots and in leaves. This analysis is in agreement with our previous observation that L has a strong and independent effect in leaves (Figure 5A), but also reveals that C and N-controlled genes in leaves are also controlled by 1 or 2 other signals (Figure 6A). To date, no gene was found to be controlled in leaves by N or C alone. This led to the striking conclusion that C or N are mainly sensed in leaves as an interaction, and not as single signals. In roots, N is highly interconnected since genes under N regulation are also under the control of three other signals (Figure 6B), and N-controlled genes are on average controlled by one more signal than genes controlled by C or L (Compare Figure 6B, panel-roots, C-L-N-slopes). Observation of the integration of C or L signal in roots again showed that the L signal is integrated in roots like the C signal. These results are consistent with the idea that the light effect in roots is sensed as the downstream consequence of sugar production through photosynthesis (Figure 6B). To discover the signal(s) that interact with N, we analyzed centroid-plots of the coefficient of genes controlled by N and other signals (Figure 6C). No particular pattern was found for genes controlled by N plus 1 or 2 signals (Figure 6C a-b). Nevertheless, a strong and unexpected pattern appeared for N plus 3 other signals (Figure6 C-c), corresponding to the largest gene list and the most dominant level of interaction (50% of the N controlled genes). Here, we can strongly conclude that N-controlled genes, at this degree of interaction, are also negatively controlled by C and L signals and positively by CL signal (100% of the 22 genes in this gene list follow this pattern).

To conclude this analysis, we wanted to define if this mode of gene regulation (N, CL induced, C and L repressed, Figure 6C-c) was unique to this cluster of 22 N-controlled genes and specific to this level of integration (N plus 3 added signals), or if we were capturing a more general pattern of signal integration valid on a genome-wide scale. To answer this question, co-effects between each pair of signals were evaluated (Figure 7). This approach again demonstrated very highly coordinated and integrated gene regulation in roots, even when considering the sign of the regulation (induction or repression). For instance, the influence of light is highly positively correlated to the influence of carbon in the control of gene expression, suggesting that in roots, the L signal is mainly sensed as sugars, the products of photosynthesis (Figure 7 a). In addition, in roots, C signals are anti-correlated with CL and CN signals. This demonstrates that the magnitude of the C effect is lowered in the presence of L and N (Figure 7 e-b). In a similar way, L effect is attenuated by C or N presence (Figure 7 f-c). The relationships found by the analysis of these plots, corresponding to the major signalling pathways in roots, have been summarized in the proposed Boolean model (Figure 7). This model is discussed later for its predicted physiological

consequences. This Boolean model constitutes a first attempt in plant biology to decipher genome-wide signal interaction in the *Arabidopsis* root. It is noteworthy that in leaves light is so predominant and insensitive to any other signal tested here (Figure 6B), that the effects of other signals (particularly when 2 signals interact) are much less meaningful (data not shown).

DISCUSSION

Four factor factorial design: The key of 'Code' discovery.

For the past decade, transcriptome studies have been used to understand molecular events involved in response to biotic, abiotic, hormonal treatments or development series (for an overview see https://www.genevestigator.ethz.ch/ or http://bbc.botany.utoronto.ca/efp/cgi-bin/efpWeb.cgi). Nevertheless, only three reports have systematically addressed the interaction between experimental factors genome-wide (C *vs* N, C *vs* L) (Gutierrez et al., 2007b; Palenchar et al., 2004; Thum et al., 2004). These approaches revealed gene networks involved in plant adaptation to a fluctuating N, C, L environment. Here, increasing the number of factors to 4 (C, N, L, O) allowed us to reach a new level of complexity. Indeed, when analyzing single factors, there are 3^1 different models possible (induced, repressed or not regulated). This same logic (depicted Figure1 B) applies to 2 factors (3^3= 27 different models), 3 factors (3^7=2,187), 4 factors (3^{15}=14,348,907) and so on. But it is only by performing the experiments to address the existence or not of all these possible models that we uncovered the tremendous constrain in signalling pathways in living organisms. In this study we found that the distribution of gene expression patterns fell within very few *models of expression*, revealing a strong coordination between signals. The probability of finding the observed associations by chance is below the smallest computable number ($<10^{-323}$). This is what motivates our concept of a 'code of signal interaction'. It is clear that our modeling approach can explain only part of the gene expression variability. However, our results already suggest that the plant cell signalling pathways are constrained such as the possible outputs in response to changes in multiple external factors belong to a very small portion of the universe of possibilities. Since our model might also miss non-linear relationships, we hypothesize that the universe of constraining structure in plant cell signalling pathways can even be more dense than proposed here. This is yet another dimension to the robustness characteristic of living organisms.

A basic rule of signal transduction

Our current analysis uncovered a relationship between the strength of signals (absolute value of model coefficient) and the number of genes controlled by these signals (Figure 4). A recurrent logarithmic law in biology is known to link the perceived sensation/response of biological systems to true stimulus intensity. The Weber–Fechner equation (Fechner, 1860) can be applied to many different biological systems: from human odor perception (Omur-Ozbek and Dietrich, 2008), time perception (Takahashi, 2007), to prefrontal cortex neuron activity of monkeys under visual stimulation (Nieder and Merten, 2007), or cockroach neuron response to light intensity (Mizunami et al., 1986). It is thus tempting to hypothesize that the

plant transcriptome response could be under the same kind of mechanistic stimulus/perception relationship. However, our study does not directly link the strength of the applied signal, but more two components of the sensed signals (1- number of regulated genes and 2- gene regulation magnitude). This will deserve further investigation to (i) validate this link between gene response and applied signal intensity in *Arabidopsis* and (ii) demonstrate that this strong logarithmic relationship can be found in other living organisms transcriptomes.

Working model validation and finding of Boolean-like signal integration.

In the proposed models to explain gene expression in response to multiple experimental factors (Figure 1), we hypothesised that plants can sense combinations of signals (Figure 1B). This assumption is first supported by experimental data. For instance, it as been demonstrated that *NRT2.1*/*NRT3.1* repression (coding a major component of NO_3^- High Affinity Transport System) is effective only when high NO_3^- and high NH_4^+ are both present in the media (Krouk et al., 2006). Our present study also supports this point of view. Indeed, the ANOVA model that we used was able to uncover genes behaving as proposed in Figure 2B. For instance, modelling of leaf data detects 3 genes controlled as a single independent signal by the presence of CL, 2 by CN (as defined in Figure 2B gene #2), or 4 by LN. In roots, 2 genes are found controlled by CL, 9 by CN, 6 by LN as a single and independent signal (Figure S2). This *post hoc* analysis provides support for the modelling approach and suggests that plants can sense combinations of factors as single signals. From another stand point, this analysis suggests that genes are under the control of AND-type logic-gate-like signallings, as we previously showed for C/L, and for NH_4/NO_3 (Krouk et al., 2006; Thum et al., 2003). Our present study suggests that this kind of Boolean-regulation can affect genome-wide expression in plants (Figure 7).

Signal integration overview in *Arabidopsis*.

The role of autotrophic leaves as an energy converter has been known since the 18th century. Shoots of plants capture photon solar energy and capture it into sugars through photosynthesis, constituting the major entry of energy in food chains. Our current findings have shown that the management of signal integration and their consequences on a genome-wide scale follows this centuries old paradigm. Our study shows that signal integration, for the considered signals, is more important in roots than in leaves. In photosynthetic leaves, the main signal in the control of gene expression is L. We also show that the L signal in leaves is insensitive to C, N, or combinations of these signals (Figure 6B). Corresponding L-controlled genes in leaves have significantly over-represented functions including metabolism, and photosynthesis (data not shown). In contrast, in the heterotrophic roots, Light is very poorly sensed on its own (Figure 5A, C-e), and L and C act on genes in an unexpected highly coordinated fashion (Figure 7 panel-a and Boolean-model). Our genome-wide study also suggests that sensing systems in roots are very sensitive to the presence of sugar, no matter if this resource comes from an

external source or from leaves as photosynthates. Recent findings on root ion transporters support this hypothesis, by showing that 16 out of 19 light or carbon regulated transporters were directly controlled by a carbon signalling pathway (Lejay et al., 2008). Moreover, we showed that CL signal exerts a negative feed-back loop on the action of C and L. This means that gene regulation by C or L reaches a plateau, and that the CL signal does not have any synergistic effect on gene expression control. This observation reinforces the notion that roots primarily sense L as C. More interestingly, we found a pronounced effect of CN as repressor of C or L signals (Figure 7 e-f). This corresponds to genes controlled by C or L, for which control is disrupted (the level of the CN coefficient is equal to C effect) by the CN presence. In other words, these 136 genes (union of gene-list corresponding to Figure7 e and f) are under the control of yet to be identified C/N balance sensing system and are up or down regulated only when C and not N is applied to plants. This type of genomic regulation could correspond to signalling evoked by Moore et al. (2003) for photosynthetic genes. Indeed sugar repression of *CAB1* and *RBCS* are antagonized by nitrate. These newly discovered candidate genes as a group will deserve further analysis to identify molecular mechanisms involved in their control, and consequently elements of the C/N balance sensing system.

In conclusion, this analysis provides mathematical models that explain global gene expression as a function of C, N, L in roots and leaves. Analysis of the models provided new insights into nutrient signal transduction pathways in a sessile organism, *Arabidopsis*. Our findings provide a new model of C, N, L signal management and suggest that many of the effects seen for single genes (Lejay et al., 2003; Lejay et al., 1999; Lejay et al., 2008; Moore et al., 2003; Rolland et al., 2002; Thum et al., 2003), are in fact managed by the plant at a systemic level (hundreds of genes: Figure 7). We believe that our findings have broad relevance since not only are plants primary providers of C and N through sugars and amino acid biosynthesis, and are thus critical for crop productivity, but carbon fixation through photosynthesis is also a major factor affecting global warming. In this context, understanding C/N/L signal interaction at a genomic scale in plants could provide new ways to tackle agricultural productivity and other socio-economical and environmental problems.

EXPERIMENTAL PROCEDURES

Plant culture and transcriptome analysis

Arabidopsis thaliana Col-0 were grown hydroponically in nutrient solution as described previously (Gutierrez et al., 2007b). Briefly, plants were grown on sand, placed in custom-designed styrofoam rafts in a growth chamber (EGC, Chagrin Falls, OH, USA) at 22°C with 60 µE light intensity and 8 h/16 h light/dark cycles. The seeds were initially germinated in tap water (1 week), then transferred to a complete nutrient solution which was renewed weekly (Krouk et al., 2006). After six weeks plant were transferred towards fresh media on the day before the experiments. For treatments, individual rafts were transferred to containers with 300 ml of nutrient solution supplemented with various concentrations of nitrate (as a mix of 2/1 KNO_3/$Ca(NO_3)_2$) and/or sucrose. The N-free nutrient solutions contained 0.25 mM K_2SO_4 and 0.25 mM $CaCl_2$ instead of KNO_3/$Ca(NO_3)_2$. Plants were transferred to treatment media at the beginning of the light period and were harvested 8 h afterwards. Roots and leaves were collected separately and quickly frozen in liquid nitrogen.

Microarray hybridization

Total RNA extraction was performed as described previously (Gutierrez et al., 2007b). Briefly, cDNA were synthesized from 8 µg total RNA using T7- Oligo(dT) promoter primer and reagents recommended by Affymetrix (Santa Clara, CA, USA). Biotin-labeled cRNA was synthesized using the Enzo BioArray HighYield RNA Transcript Labeling Kit (Enzo, New York, NY, USA). The concentration and quality of the cRNA was evaluated by A260/280 nm reading and 1% agarose gel electrophoresis. We used 15 µg of labeled cRNA to hybridize the *Arabidopsis* ATH1 Affymetrix gene chip for 16 h at 42°C. Washing, staining and scanning were performed as recommended by Affymetrix. Image analysis and normalization to a target median intensity of 150 was performed with the Affymetrix MAS v5.0 set at default values. We analyzed the reproducibility of replicates using the correlation coefficient and visual inspection of scatter plots of pairs of replicates. One pair of duplicates failed this quality control. Thus, to improve reliability of the measure we performed 2 more Affymetrix chips from independent samples corresponding to the condition : Roots, Light, No nitrogen, No carbon.

Modelling of gene expression patterns

All data manipulations were performed on R (http://www.r-project.org/). The ANOVA analysis was carried out using the R *lm()* function with three models. The first model considers the organs as a factor, such that the expression Y_i of a $gene_i$ is given by: $Y_i = \alpha_0 + \alpha_1 C + \alpha_2 L + \alpha_3 N + \alpha_4 O + \alpha_5 CL + \alpha_6 CN + \alpha_7 CO + \alpha_8 LN + \alpha_9 NO + \alpha_{10} LO + \alpha_{11} CNL + \alpha_{12} LNO + \alpha_{13} CNO + \alpha_{14} CLO + \alpha_{15} CLNO + Z$. In this model, α_0 represents the expression under a “control” condition (without C, without N ,

without L, in roots); *Z* represents the noise; and α_1 to α_{15} represent the coefficients quantifying the effect of each factor (C, N, L, O) or combination of factors. The second model is just a simplified version of the first model in which gene expression in roots and leaves datasets were analyzed separately: $Y_i = \alpha_0 + \alpha_1 C + \alpha_2 L + \alpha_3 N + \alpha_4 CL + \alpha_5 CN + \alpha_6 LN + \alpha_7 CNL + Z$. Each gene was analyzed separately. We addressed multiple testing by controlling the false discovery rate (FDR) at 1% at each stage of the evaluation procedure as described previously (Gutierrez et al., 2007b). A rigorous statistical procedure was implemented to avoid over-fitting. The complete models were used to assess whether gene expression could be explained at all by any combination of the coefficients. If the model was significant at 1% FDR, then each significant term in the model was evaluated to determine if its presence contributed to the final model. Terms with higher p-values were tested first. We used the anova() function to compare models at each iteration of the procedure. Significant coefficients were organized as presented in supplemental Tables S1, S2, S3.

Clustering

Hierarchy between signals were evaluated by average linkage hierarchical clustering. First, euclidian distances were calculated using the *dist()* function in the R software. Second, clusters were generated by the *hclust()* function. Third, plot were generated using the *plot()* (default values) function. Dendrogram interpretations were carried out as previously described (Speed, 2003).

ACKNOWLEDGMENTS

We thank Sandrine Ruffel, Francisco Melo and Miriam Gifford for helpful discussion and critical reading of the manuscript. This work was funded by NIH NIGMS GRANT GM032877 to GC, and NSF Arabidopsis 2010 Genome Grant (IOB 0519985) to GC, DS, MILLENIUM NUCLEUS (P06-009F) and FONDECYT (1060457) to RG and by an European-FP7-International Outgoing Fellowships (Marie Curie) (AtSYSTM-BIOL) to GK.

REFERENCES

Achard, P., Cheng, H., De Grauwe, L., Decat, J., Schoutteten, H., Moritz, T., Van Der Straeten, D., Peng, J., and Harberd, N.P. (2006). Integration of plant responses to environmentally activated phytohormonal signals. Science *311*, 91-94.

Baena-Gonzalez, E., Rolland, F., Thevelein, J.M., and Sheen, J. (2007). A central integrator of transcription networks in plant stress and energy signalling. Nature *448*, 938-942.

Camargo, A., Llamas, A., Schnell, R.A., Higuera, J.J., Gonzalez-Ballester, D., Lefebvre, P.A., Fernandez, E., and Galvan, A. (2007). Nitrate signaling by the regulatory gene NIT2 in Chlamydomonas. Plant Cell *19*, 3491-3503.

Castillon, A., Shen, H., and Huq, E. (2007). Phytochrome Interacting Factors: central players in phytochrome-mediated light signaling networks. Trends Plant Sci *12*, 514-521.

Chen, H., Zhang, J., Neff, M.M., Hong, S.W., Zhang, H., Deng, X.W., and Xiong, L. (2008). Integration of light and abscisic acid signaling during seed germination and early seedling development. Proc Natl Acad Sci U S A *105*, 4495-4500.

Coruzzi, G., and Zhou, L. (2001). Carbon and nitrogen sensing and signaling in plants: emerging 'matrix effects'. Curr Opin Plant Biol *4*, 247-253.

Fechner, G. (1860). Elemente der psychophysik. *Vol II*, Leipzig: Breitkopf and Hartel.

Gazzarrini, S., and McCourt, P. (2003). Cross-talk in plant hormone signalling: what Arabidopsis mutants are telling us. Ann Bot (Lond) *91*, 605-612.

Guo, F.Q., Wang, R., and Crawford, N.M. (2002). The Arabidopsis dual-affinity nitrate transporter gene AtNRT1.1 (CHL1) is regulated by auxin in both shoots and roots. J Exp Bot *53*, 835-844.

Gutierrez, R.A., Gifford, M.L., Poultney, C., Wang, R., Shasha, D.E., Coruzzi, G.M., and Crawford, N.M. (2007a). Insights into the genomic nitrate response using genetics and the Sungear Software System. J Exp Bot *58*, 2359-2367.

Gutierrez, R.A., Lejay, L.V., Dean, A., Chiaromonte, F., Shasha, D.E., and Coruzzi, G.M. (2007b). Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. Genome Biol *8*, R7.

Krouk, G., Tillard, P., and Gojon, A. (2006). Regulation of the high-affinity NO_3^- uptake system by a NRT1.1-mediated "NO_3^--demand" signalling in *Arabidopsis*. Plant Physiol *142*, 1075-1086.

Lejay, L., Gansel, X., Cerezo, M., Tillard, P., Muller, C., Krapp, A., von Wiren, N., Daniel-Vedele, F., and Gojon, A. (2003). Regulation of root ion transporters by photosynthesis: functional importance and relation with hexokinase. Plant Cell *15*, 2218-2232.

Lejay, L., Tillard, P., Lepetit, M., Olive, F., Filleur, S., Daniel-Vedele, F., and Gojon, A. (1999). Molecular and functional regulation of two NO_3^- uptake systems by N- and C-status of Arabidopsis plants. Plant J *18*, 509-519.

Lejay, L., Wirth, J., Pervent, M., Cross, J.M., Tillard, P., and Gojon, A. (2008). Oxidative pentose phosphate pathway-dependent sugar sensing as a mechanism for regulation of root ion transporters by photosynthesis. Plant Physiol *146*, 2036-2053.

Little, D.Y., Rao, H., Oliva, S., Daniel-Vedele, F., Krapp, A., and Malamy, J.E. (2005). The putative high-affinity nitrate transporter NRT2.1 represses lateral root initiation in response to nutritional cues. Proc Natl Acad Sci USA *102*, 13693-13698.

Malamy, J.E. (2005). Intrinsic and environmental response pathways that regulate root system architecture. Plant Cell Environ *28*, 67-77.

Maruyama-Nakashita, A., Nakamura, Y., Tohge, T., Saito, K., and Takahashi, H. (2006). Arabidopsis SLIM1 is a central transcriptional regulator of plant sulfur response and metabolism. Plant Cell *18*, 3235-3251.

Mizunami, M., Tateda, H., and Naka, K. (1986). Dynamics of cockroach ocellar neurons. The Journal of general physiology *88*, 275-292.
Moore, B., Zhou, L., Rolland, F., Hall, Q., Cheng, W.H., Liu, Y.X., Hwang, I., Jones, T., and Sheen, J. (2003). Role of the Arabidopsis glucose sensor HXK1 in nutrient, light, and hormonal signaling. Science *300*, 332-336.
Muños, S., Cazettes, C., Fizames, C., Gaymard, F., Tillard, P., Lepetit, M., Lejay, L., and Gojon, A. (2004). Transcript profiling in the chl1-5 mutant of Arabidopsis reveals a role of the nitrate transporter NRT1.1 in the regulation of another nitrate transporter, NRT2.1. Plant Cell *16*, 2433-2447.
Nemhauser, J.L., Hong, F., and Chory, J. (2006). Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. Cell *126*, 467-475.
Nemhauser, J.L., Mockler, T.C., and Chory, J. (2004). Interdependency of brassinosteroid and auxin signaling in Arabidopsis. PLoS biology *2*, E258.
Nieder, A., and Merten, K. (2007). A labeled-line code for small and large numerosities in the monkey prefrontal cortex. J Neurosci *27*, 5986-5993.
Omur-Ozbek, P., and Dietrich, A.M. (2008). Developing hexanal as an odor reference standard for sensory analysis of drinking water. Water research.
Palenchar, P.M., Kouranov, A., Lejay, L.V., and Coruzzi, G.M. (2004). Genome-wide patterns of carbon and nitrogen regulation of gene expression validate the combined carbon and nitrogen (CN)-signaling hypothesis in plants. Genome Biol *5*, R91.
Poultney, C.S., Gutierrez, R.A., Katari, M.S., Gifford, M.L., Paley, W.B., Coruzzi, G.M., and Shasha, D.E. (2007). Sungear: interactive visualization and functional analysis of genomic datasets. Bioinformatics *23*, 259-261.
Price, J., Laxmi, A., St Martin, S.K., and Jang, J.C. (2004). Global transcription profiling reveals multiple sugar signal transduction mechanisms in Arabidopsis. Plant Cell *16*, 2128-2150.
Rolland, F., Baena-Gonzalez, E., and Sheen, J. (2006). Sugar sensing and signaling in plants: conserved and novel mechanisms. Annual review of plant biology *57*, 675-709.
Rolland, F., Moore, B., and Sheen, J. (2002). Sugar sensing and signaling in plants. Plant Cell *14 Suppl*, S185-205.
Speed, T. (2003). Statistical analysis of gene expression in microarray data. Book CRC press, 240p.
Stepanova, A.N., Robertson-Hoyt, J., Yun, J., Benavente, L.M., Xie, D.Y., Dolezal, K., Schlereth, A., Jurgens, G., and Alonso, J.M. (2008). TAA1-mediated auxin biosynthesis is essential for hormone crosstalk and plant development. Cell *133*, 177-191.
Takahashi, T. (2007). Hyperbolic discounting may be reduced to electrical coupling in dopaminergic neural circuits. Medical hypotheses *69*, 195-198.
Tao, Y., Ferrer, J.L., Ljung, K., Pojer, F., Hong, F., Long, J.A., Li, L., Moreno, J.E., Bowman, M.E., Ivans, L.J.*, et al.* (2008). Rapid synthesis of auxin via a new tryptophan-dependent pathway is required for shade avoidance in plants. Cell *133*, 164-176.
Thum, K.E., Shasha, D.E., Lejay, L.V., and Coruzzi, G.M. (2003). Light- and carbon-signaling pathways. Modeling circuits of interactions. Plant Physiol *132*, 440-452.
Thum, K.E., Shin, M.J., Palenchar, P.M., Kouranov, A., and Coruzzi, G.M. (2004). Genome-wide investigation of light and carbon signaling interactions in Arabidopsis. Genome Biol *5*, R10.

Figure and Table Legends

Figure 1. Scheme of experimental design and working model of gene control by multiple signals at the organ-specific level.

A) 6 week old plants were treated for 8h with all combinations of 4 (C,L,N,O) binary (0/1) factors for a total of 16 experimental conditions. Treatments were N (5 mM NO_3^-), C (30mM sucrose), Light (50 μE). RNAs were extracted from roots and shoots separately and hybridized to ATH1 Affymetrix chips. Microarray data analysis was performed as described in the Experimental Procedures. B) Scheme presenting the concept used to decipher signal interactions in the control of gene expression. We propose signals can be produced from a factor (C, N, L represented as blue squares) or combination of factors (white squares). These signals can then affect the expression of a particular gene. The expression of a gene (e.g. black circles labeled 1 and 2) can be affected by (red arrow) one factor (e.g. C alone for number 1) or combination of factors (e.g. when C and N are together for number 2). C) Idealized gene expression patterns produced by the signal effects shown in (B) for the two genes 1 and 2**.**

Figure 2. A small number of models explain most gene expression patterns in response to 16 different experimental conditions.

The gene expression patterns obtained in the 16 different experimental conditions were modeled as a function of the four experimental factors (C, N, L, O) and their interactions using a rigorous statistical procedure (see Experimental Procedures). Genes with the same model of expression were grouped. The graph shows the number of genes (Y-axis) explained by the different models of gene expression (X-axis).

Figure 3. Signal strength and relationship for the control of gene expression.

A,B) Analysis using the entire data set; C,D) Analysis using data from Leaves; E,F) Analysis using data from Roots; A-C) Dendrograms produced by average linkage hierarchical clustering analysis with euclidean distance carried out on the simplified model matrixes as described in the text. B-F) Analysis of signal strength using the Sungear software. The Sungear polygon shows the signals at the vertices (anchors). The circles inside the polygon (vessels) represent the genes controlled by different signals as indicated by the arrows around the vessels. The area of each vessel (size) is proportional to the number of genes associated with that vessel. Thus, visually and quantitatively it is possible to identify the main signal at the whole dataset level as O. In leaves, L predominates, and in roots C and L are similar with regard to the number of genes affected by the signals.

Figure 4. Relationship between the number of regulated genes and the magnitude of gene regulation (coefficients of the model).

The graphs show the relationship between the average coefficient and the number of genes that showed the coefficient as significant in the regulation model. Circles are labeled with the corresponding signal. The coefficient of determinations (R^2) for each logarithmic regression analysis is indicated in the graphs. (A) Analysis for the complete data set. (B) Analysis for Roots and Leaves data set separately.

Figure 5. Signal integration at the organ specific level.

A) Percentage of regulated genes as a function of the number of signals. In leaves most genes are regulated by only 1 signal, labeled with the letter "a". B) Genes belonging to the groups labeled with letters in panel A, were subjected to average linkage hierarchical clustering with euclidean distance to analyze the signal relationship across increasingly complex models of gene expression. Dendrograms show hierarchy of signals in the control of gene expression (a to d for leaves, e to h for roots).

Figure 6. Signal integration of C, L, N, O factors (A-B): Case of study for Nitrogen (C)

Effect of added signal on the percentage of controlled genes considering each factor at A) the whole dataset level, B) the organ specific level. C) Centroid-plots of LM coefficients for the gene lists (a to d) considered in B-roots-Nitrogen. Note that in C-b, N effect is significant (by definition), but no trend between genes gathered in the list can be visualized. Some of them the are positively controlled by N some of them negatively.

Figure 7.. Signal integration in roots.

Genes controlled by at least each pair of considered signals were sorted then plotted based on their gene expression model coefficients. The strong relationships discovered (a-e) were summarized by a Boolean model. Pearson correlation coefficients and number of genes are provided in each panel, and in the general Boolean model respectively.

Table 1. Predominant model of expression at the whole data set level.

Expression of each gene has been modeled as a function of C, L, N, O factors and their interaction. Each gene model was recoded replacing by -1, 0 or 1, coefficients respectively for negatively, not significant or positively regulated genes.

Number of genes in each class is indicated in the last column. Red lines underline complex models gathering a large number of genes. The full version of this table is provided in Table S4.

TABLES

Table 1

C	N	L	O	CN	CL	CO	LN	NO	LO	CNL	LNO	CNO	CLO	CNLO	#genes
0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	**1009**
0	0	0	1	0	0	0	0	0	-1	0	0	0	0	0	**676**
0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	**502**
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	**485**
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**337**
0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	**229**
0	0	-1	-1	0	0	0	0	0	0	0	0	0	0	0	**226**
-1	0	-1	-1	0	1	1	0	0	1	0	0	0	-1	0	**154**
0	0	1	0	0	0	0	0	0	-1	0	0	0	0	0	**153**
0	0	-1	-1	0	0	0	0	0	1	0	0	0	0	0	**142**
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	**139**
1	0	1	0	0	0	-1	0	0	0	0	0	0	0	0	**135**
1	0	1	0	0	0	-1	0	0	-1	0	0	0	0	0	**112**
															...