Observability of Plant Metabolic Networks Is Reflected in the Correlation of Metabolic Profiles[1]

Kevin Schwahn, Anika Küken, Daniel J. Kliebenstein, Alisdair R. Fernie, and Zoran Nikoloski*

Systems Biology and Mathematical Modeling Group (K.S., A.K., Z.N.) and Central Metabolism Group (K.S., A.R.F.), Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany; Department of Plant Sciences, University of California, Davis, Davis, California 95616 (D.J.K.); and DynaMo Center of Excellence, University of Copenhagen, DK-1871 Frederiksberg C, Denmark (D.J.K.)

ORCID IDs: 0000-0003-1367-0719 (A.K.); 0000-0001-5759-3175 (D.J.K.); 0000-0003-2671-6763 (Z.N.).

Understanding whether the functionality of a biological system can be characterized by measuring few selected components is key to targeted phenotyping techniques in systems biology. Methods from observability theory have proven useful in identifying sensor components that have to be measured to obtain information about the entire system. Yet, the extent to which the data profiles reflect the role of components in the observability of the system remains unexplored. Here we first identify the sensor metabolites in the model plant Arabidopsis (*Arabidopsis thaliana*) by employing state-of-the-art genome-scale metabolic networks. By using metabolic data profiles from a set of seven environmental perturbations as well as from natural variability, we demonstrate that the data profiles of sensor metabolites are more correlated than those of nonsensor metabolites. This pattern was confirmed with in silico generated metabolic profiles from a medium-size kinetic model of plant central carbon metabolism. Altogether, due to the small number of identified sensors, our study implies that targeted metabolite analyses may provide the vast majority of relevant information about plant metabolic systems.

Systems biology aims at developing models that allow for a complete characterization of how the inputs and outputs of a biological system are interconnected and jointly relate to the molecular phenotypes. The experimental systems biology studies attempt to obtain a substantial coverage of the (molecular) components of a biological system using various technological platforms, such as transcriptomics (Weber et al., 2007) and metabolomics (Fiehn, 2002), and, more recently, phenomics (Araus and Cairns, 2014). The aim of these research efforts is to utilize the read-outs about the components for estimating how the biological system functions.

However, while these efforts are rapidly becoming faster and cheaper, they still encounter both financial and logistical problems when attempting to scale up to measure large populations or the vast space of conceivable physiological environments. These problems quickly become irresolvable for any studies attempting to combine genetic and environmental variation in the same system. Thus, until the technical problems are removed, alternative solutions are in demand that can allow as much of the system to be measured (i.e. observed) as possible. Therefore, we are faced with the question: Is it possible to identify a subset of transcripts or metabolites that can provide complete information about an investigated system?

One way to identify these subsets is based on networks structures generated by systems biology approaches. This line of research aims at finding a small number of molecular components (with respect to what can be measured) whose measurement can characterize the internal state of a biological system. Given the myriad of output components from any biological system (e.g. generated within a plant leaf cell and exported to any other tissue type), it is of great interest to determine the number and the identity of these output components that may provide insights into the state of the system. However, components deemed as outputs of a modeled biological system are usually not external to the system, but rather, actively participate in shaping the levels of its underlining components. For instance, amino acids are used to build proteins that, in turn, drive the entirety of metabolism, including amino acid and sugar metabolism that provide the energy and building blocks to create the plant cell wall (Cosgrove, 2005; Singh and Ghosh, 2006). Therefore, the connectivity of components due to

[1] K.S. was funded by the International Max Planck Research Schools "Primary Metabolism and Plant Growth" program of the Max Planck Institute of Molecular Plant Physiology. This effort was also funded by the National Science Foundation under grant numbers DBI 0820580 (D.J.K.) and MCB 330337 (D.J.K.); the USDA National Institute of Food and Agriculture under Hatch project number CA-D-PLS-7033-H (D.J.K.); and the Danish National Research Foundation under grant number DNRF99 (D.J.K.).

* Address correspondence to nikoloski@mpimp-golm.mpg.de.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Zoran Nikoloski (nikoloski@mpimp-golm.mpg.de).

Z.N. conceived the project and wrote the article with contribution of all authors; K.S. performed most the analysis and analyzed the data; A.K. contributed the kinetic model and related functions; A.R.F. and D.J.K. provided biological interpretation.

www.plantphysiol.org/cgi/doi/10.1104/pp.16.00900

regulatory, signaling, and metabolic interactions must be considered when determining the sensor components.

Metabolic networks are among the best described networks in systems biology to test our ability to identify metabolites that can serve as sensors to describe metabolism. We would like to emphasize that the concept of sensor metabolites does not correspond to the in vivo notion of sensing and signaling metabolites. Sensing and signaling metabolites are involved in coregulating and integrating the metabolic status with other cellular events (Templeton and Moorhead, 2004). Our concept of sensor metabolites is that the metabolites would need to be measured by the researcher to acquire the majority of information present in the sample.

A metabolic network of a given cellular system consists of the entirety of biochemical reactions interconverting nutrients obtained from the environment, into basic and more complex building blocks used to create the cell and allow it to defend itself. The components of a metabolic network are, therefore, the metabolites and the accompanying conversion reactions. These components are fully specified by the levels of all the metabolites and the rates/fluxes of all reactions. While the levels of many metabolites can be determined with modern metabolomics technologies (Goodacre et al., 2004), the reaction rates cannot be measured but are estimated from the combination of labeling and modeling (Kauffman et al., 2003; Nöh et al., 2007). Recent advances in modeling of plants have resulted in genome-scale metabolic networks for a variety of species, from Arabidopsis (*Arabidopsis thaliana*), as a plant model, to maize and rice, as important agronomic crops (de Oliveira Dal'Molin et al., 2010a, 2010b; Saha et al., 2011; Seaver et al., 2014).

Well-established methods from control theory utilize network structure to determine the sensors that must be measured to observe the internal state of a system, biological or otherwise (Liu et al., 2011, 2013; Jha and van Schuppen, 2001; Rios et al., 2013). These methods are not concerned with calculating the internal states from the sensors, but determining if the system is observable with particular components. For nonlinear biological systems, such as metabolism, obtaining the internal state from the sensors is still a challenging problem (Chaves and Sontag, 2002). The issue of determining the set of metabolites that needs to be measured in a labeling experiment to characterize a unique flux distribution of a given system has been tackled in the framework of constrained-based modeling (Chang et al., 2008).

Here, we address the observability problem from a data-driven perspective: To begin, we apply the graphical approach of Liu et al. (2013) to large-scale plant metabolic networks. We then investigate if, and to what extent, the data profiles about metabolites predicted as sensors relate to the rest of the metabolites in the network. In this way, we aim to bridge the gap between the existing powerful control-theoretic methods and the plethora of accumulated data from metabolomics studies. To inspect the model-based effects in the identification of sensor metabolites, we tested the robustness of the findings with two different models that guarantee good coverage with the metabolomics data. In addition, we used a medium-scale kinetic model for central carbon metabolism to further strengthen our findings from the large-scale models. The findings are further discussed with respect to the role of sensor metabolites as dead-end metabolites in the respective metabolic networks (with and without consideration of biomass reactions, used in the simulating growth). The small number of identified sensor metabolites in relation to the size of the entire metabolic network suggests that targeted metabolite analyses could provide the vast majority of relevant information about plant metabolic systems and could prove effective in strategies for crop improvement (Gu et al., 2010, 2012; Lu et al., 2011; Fernie and Schauer, 2009).

RESULTS AND DISCUSSION

Number and Position of Sensor Metabolites in Models of Plant Primary Metabolism

By applying the graphical approach to identify root strongly connected components (SCCs) in the Arabidopsis core model (AraCORE; Arnold and Nikoloski, 2014), we found 23 sensor metabolites listed in Supplemental Table S1. Aside from two sugars, trehalose and cellulose, and nucleoside triphosphates, the remaining sensor metabolites were amino acids. The metabolomics data set of Caldana et al. (2011) contained the metabolic profiles of 11 of the identified sensor metabolites. Overall, 30 of the 91 measured metabolites could be mapped to AraCORE, as indicated in Supplemental Table S5 that includes the metabolites used as sensors and nonsensors for the investigation of this model. Consideration of biomass and sink reactions in the model led to the identification of only 15 sensor metabolites, consisting of the amino acids and cellulose (see Supplemental Table S2—"AraCORE Sensors with Biomass Function"). The finding that the sensors identified upon consideration of biomass also act as sensors when biomass is excluded was in line with the observation that the biomass reaction includes all amino acids, alongside cellulose and nucleotides.

Additionally, we also considered the Arabidopsis model downloaded from PlantSEED (AraSEED; Seaver et al., 2014). We identified 198 sensor metabolites, given in Supplemental Table S3, of which 10 could be mapped to the metabolomics data. Overall, we were able to map 47 metabolites to the measured data. The list of all metabolites identified as sensor and nonsensor in the investigation of the AraSEED model is provided in Supplemental Table S6. In agreement with the AraCORE model, the sensors again included amino acids and sugars, in addition to a variety of complexes with Coenzyme A and Plastoquinone. In brief, the findings from the two models were similar in that all models have root SCCs that largely overlap sugar and amino acid metabolism. However, the small number of mapped metabolites in comparison to the size of the models employed is a challenge, largely due to the limitations of current metabolomics technologies. For instance, a quarter of the detected analytes could not be annotated to

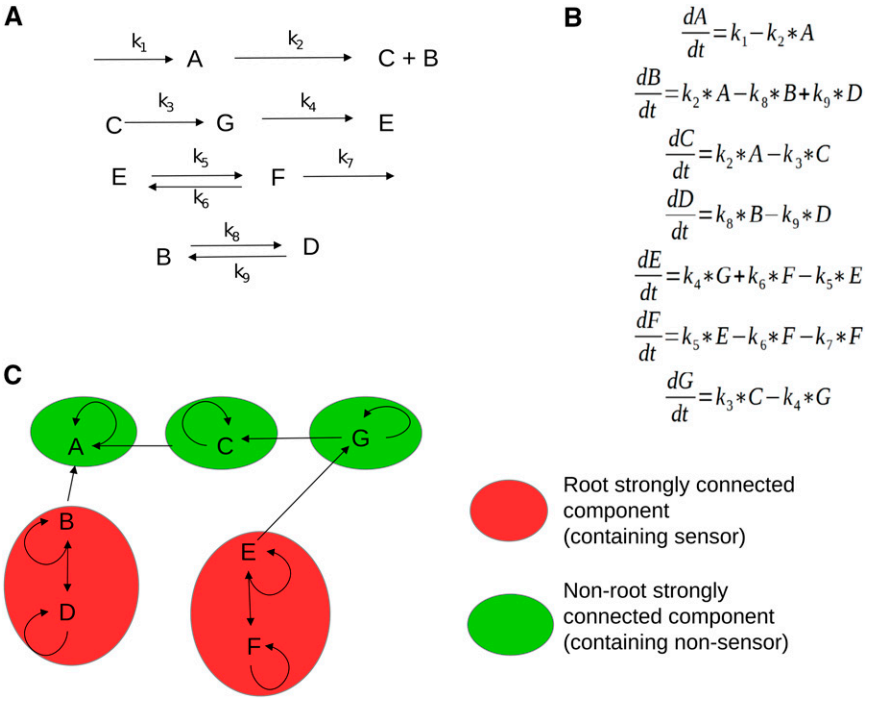

Figure 1. Schematic overview of the implemented algorithm, adapted from Liu et al. (2013). A, Example of a metabolic network with nine irreversible reactions. B, System of differential equations for the change in concentration for each metabolite (A–D) in the network shown in A assuming mass action kinetic. C, Inference graph for the metabolic network and system of differential equations in A and B. Node *u* is connected by directed edge to node *v* if metabolite *v* occurs in the differential equation for metabolite *u* from B. The green circles represent nonroot SCC, whereas the red circles indicate root SCC. Each node in a root SCC can act as a sensor node.

known metabolites; moreover, secondary metabolites could not be mapped in all models, because some of the models used in this study include pathways of central metabolism.

Data Profiles of Sensor Metabolites Show Stronger Correlations Than Nonsensor Metabolites

The underlying approach states that information from all root-SCC allows the reconstruction of the state of the system. A minimum set of sensor metabolites can then be used to specify the metabolic profiles of the sensors and the rest of the network. It is important to emphasize that the metabolites from the root SCCs, containing the sensors, can be connected to different nonroot SCCs. Therefore, one may expect that there is a relation within sensors based on whether they are connected to the same nonroot SCCs (see Supplemental Fig. S1 for illustration). If all nodes in a nonroot SCC have directed path to sensors in two SCCs, a single

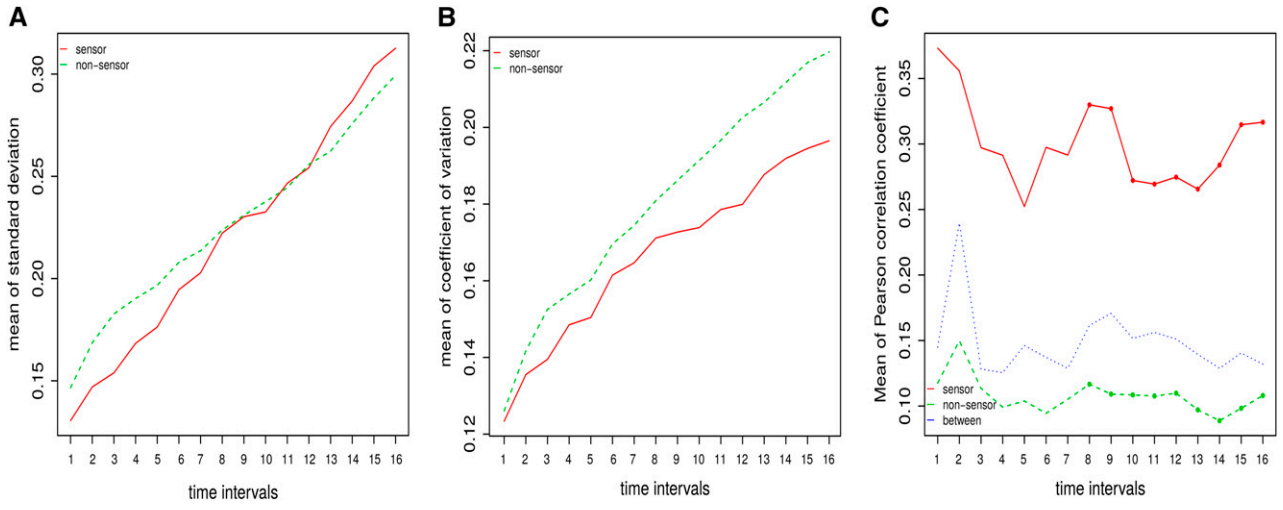

Figure 2. Statistical comparison of sensors and nonsensors in the AraCORE model. The *x* axis represents the investigated time interval, from 1 to 16. The *y* axis represents values for the three statistics, respectively: A, SD; B, CV; and C, Pearson correlation of sensors and nonsensors. The red line corresponds to the values for the statistics between sensor metabolites, while the green line corresponds to values between nonsensor metabolites. The blue line in C is used for the correlation between sensors and nonsensors. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and nonsensors.

sensor may suffice to reconstruct the state of the nonroot SCCs; in this case, the other sensor will be needed to describe its own profile. For the investigated networks, the majority of the identified sensor metabolites were connected, via a directed path, to the same nonroot SCC. Therefore, sensors in different root SCCs detect the same network and each could be employed to reconstruct the state of the nonroot SCC. Therefore, if the data profiles of a sensor metabolite can be used to reconstruct the profiles of the nonsensor metabolites, it may be expected that sensor metabolites are more correlated to each other than to the rest of the metabolites; by corollary, for nonsensor metabolites, it may be expected that they are less correlated to each other than to the sensor metabolites. Within the AraCORE model all sensors are connected to the same nonroot SCC, whereas in the AraSEED model 35 of the 198 sensors were not connected to the largest nonroot SCC. Out of the 35 sensors, only Glc and Fru were mapped to the data. We did not observe a different behavior with respect to the findings from the previous analysis for two types of sensor groups (see Supplemental Fig. S4).

To empirically test these sensor hypotheses and their biological utility, we used time-series metabolomics data from Arabidopsis Col-0 exposed to seven different environments. To this end, we determined the correlation for each pair of measured metabolites over all conditions; we then divided the resulting correlation values in three categories: (1) between two sensors, (2) two nonsensors, and (3) between a sensor and a nonsensor metabolite. Because the available time-series data captured the response to the applied perturbations caused by the different light and temperature conditions across different time scales, we determined the correlation between the time series with consideration of different time points (i.e. intervals). More specifically, we determined the correlations by using k (with $5 \leq k \leq 20$) consecutive time points from the experimental measurements, starting with the first time point (Fig. 2). This results in 16 time intervals, so that the first interval consists of the first five measured time points and the last of all 20 time points. In addition, we investigated the correlation obtained by jointly considering the data from all time points and conditions.

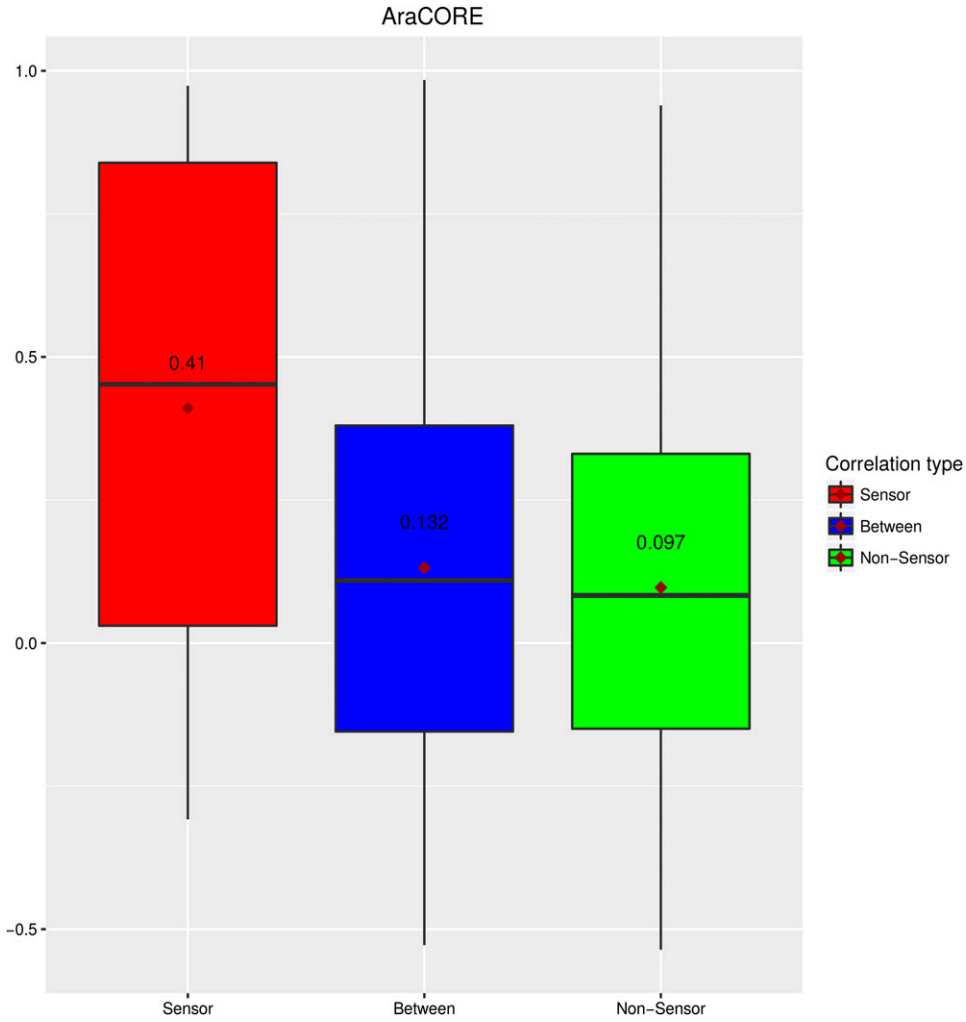

Figure 3. Distribution of correlation values of the AraCORE model. Box plots of the distribution of correlation values between sensors, between sensors and nonsensors, and between nonsensors are colored in red, blue, and green, respectively. The mean value is given above the square symbol, while the median is given by the solid line.

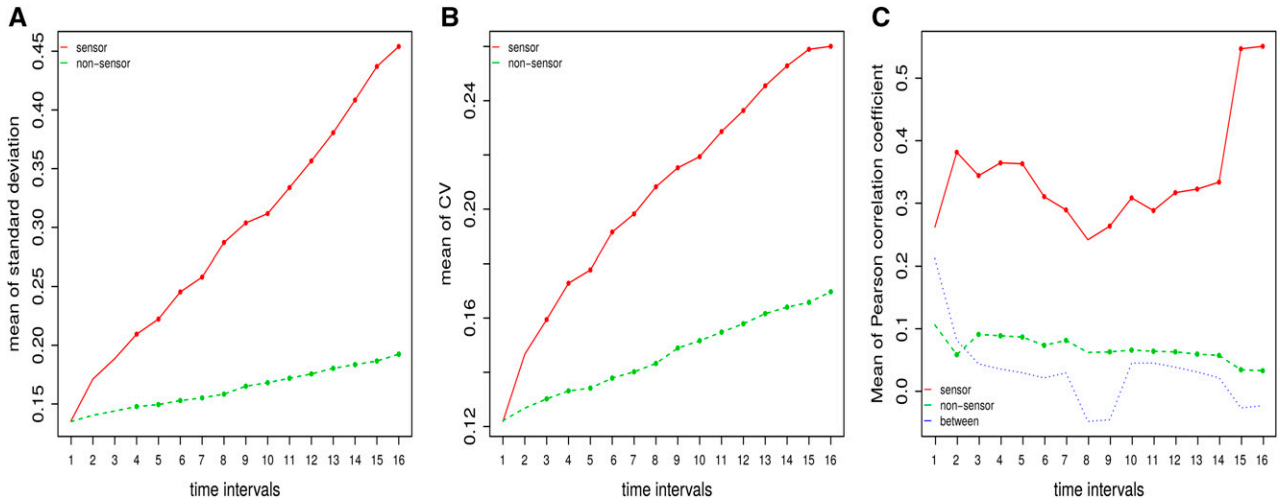

Figure 4. Statistical comparison of sensors and nonsensors in the AraSEED model. The *x* axis represents the investigated time interval, from 1 to 16. The *y* axis represents values for the three statistics, respectively: A, SD; B, CV; and C, Pearson correlation of sensors and nonsensors. The red line corresponds to the values for the statistics between sensor metabolites, while the green line corresponds to values between nonsensor metabolites. The blue line in C is used for the correlation between sensors and nonsensors. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and nonsensors.

This analysis provided the distributions of correlation values across the three classes of metabolite pairs in a given time interval over all considered conditions. We then tested the null hypothesis that the means of the distributions do not statistically differ between the classes of metabolite pairs, by applying a two-sided *t* test. In accordance with the observation that the majority of the identified sensor metabolites were connected to the same (nonroot) SCCs, for the AraCORE model, we found that the mean of correlations between sensor metabolites was greater than the mean of correlations between nonsensor metabolites in 9 of the 16 investigated intervals. The statistical significance in the later time points is due to the larger power of the test due to the larger number of data points available (Schönbrodt and Perugini, 2013). We also observed that the mean of the correlations between sensor and nonsensor metabolites was greater than the correlations between nonsensor metabolites, but smaller than the correlations between sensor metabolites only. These results were reproducible if all time points and conditions were jointly used (Fig. 3).

However, another possible source of this result is that the metabolic profiles of the sensor metabolites have lower variability than nonsensors, and, thus, show higher correlations. To test this hypothesis, we first determined the distributions of SD and coefficient of variation (CV) or the sensor and nonsensor metabolites equivalent to the setup for investigating the correlation values. We then tested if the means of each measure of variability differed between the classes of metabolites. The means of the CV and the SD were not statistically different between the nonsensors and sensors; thus, differences in the variation of sensors and nonsensors were likely not causing the difference in the correlation structure. Therefore, we concluded that the observed difference in correlation was a result of the position of the sensor nodes in the network and not due to smaller variability.

For a comparison of genome-scale models, we investigated the relationship of sensors and nonsensors in the AraSEED model. The mean correlations of the sensors were significantly different and larger than those of the nonsensors in 13 of 16 time intervals (see Fig. 4), thus conforming to our previous findings. This was additionally confirmed through the investigation of all time points and conditions (Fig. 5). The correlations in sensors were significantly higher than in nonsensors. However, the results of the SD and the CV were in contrast to our previous results: In the majority of time intervals, we found significantly higher values in the sensors than for the nonsensors. This is likely due to the difference in the number of sensors and nonsensors mapped from the metabolomics data in the two models.

Altogether, we demonstrated that, with the used data set, sensors show larger correlation than between nonsensors and sensors, and that that the latter is greater than the correlation within nonsensors. We also showed that these findings remained largely unaltered when models of different size and structure are explored. The evidence indicates that in the case of AraCORE, these findings are likely not related to the variability in the metabolic profiles. In addition, we investigated correlation of the metabolic traits gathered in a study by Sulpice et al. (2013). The data were obtained under three different growth conditions with respect to nitrogen (N) and carbon availability, and included the levels of 45 metabolites from 97 Arabidopsis accessions. Because the models largely encompass the reactions from central carbon metabolism, we expect that the structure of the metabolic network remains unaltered between accessions;

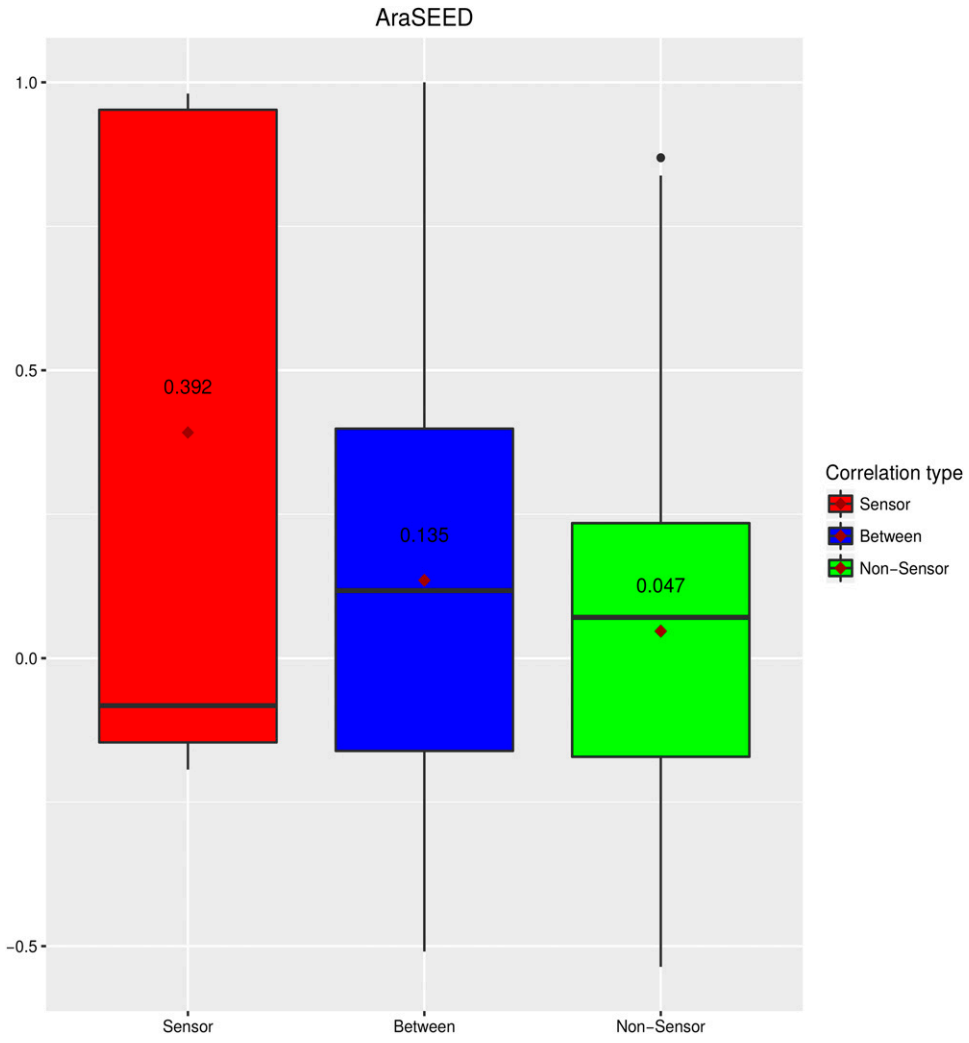

Figure 5. Distribution of correlation values of the AraSEED model. Box plots of the distribution of correlation values between sensors, between sensors and nonsensors, and between nonsensors are colored in red, blue, and green, respectively. The mean value is given above the square symbol, while the median is given by the solid line.

under this assumption, the data profiles can be regarded as realizations of the same network. Therefore, we selected the sensors and nonsensors from the two models and repeated the correlation analysis.

In the AraCORE model, we could map 12 sensors and 20 nonsensors, while data were available for 10 sensors and 25 nonsensors in the AraSEED model. The correlation between sensors was significantly higher compared to nonsensors in AraCORE and the AraSEED model (Supplemental Fig. S2). This analysis demonstrated that, under simplifying assumptions about robustness of central carbon metabolism in plants, similar patterns between sensors and nonsensors as in the analysis of single genotypes can also be found by using data from genetically variable populations.

Analysis of Robustness for the Observed Sensor/Nonsensor Patterns

To determine if observed pattern of correlations within and between sensors and nonsensors were not artifacts of the used network and could not have resulted by arbitrary grouping of metabolites, we conducted two types of robustness analyses. In the first, we randomized the partition of metabolites into the two classes, while in the second we inspected the effect of the reversibility of reactions considered in the metabolic network.

In the first analysis of robustness, we determined the probability that a random partition of metabolites into same number of sensor and nonsensor metabolites (as in the findings) results in the observed pattern of correlations. To this end, we shuffled the assignment of sensor and nonsensor metabolites 500 times, while keeping their respective total numbers fixed, and determined the data properties, namely, SD and CV, as well as correlation for the classes of metabolites and metabolite pairs. This robustness analyses demonstrated that the observed larger correlation of sensors in comparison to nonsensors was statistically significant. In addition, the correlation between sensor and nonsensors was similar to the other two estimated correlations of sensors to sensors and nonsensors to nonsensors. We further

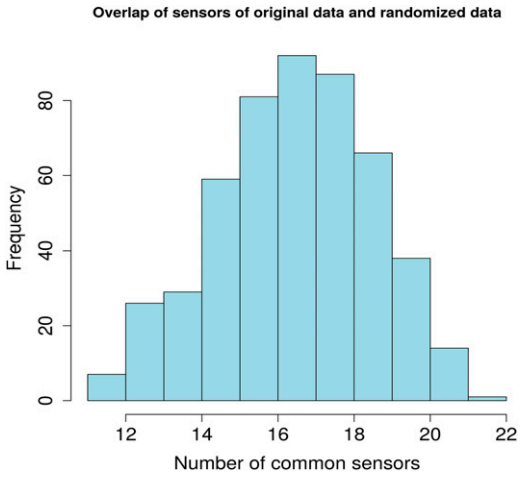

Figure 6. Distribution for the size of the overlap of identified sensors. The distribution is obtained after randomizing the reversibility assignment in in the AraCORE model. The *x* axis displays the number of sensors overlapping with the original analysis. The *y* axis displays the frequencies of common sensors in 500 shufflings of the reversibility assignment. Originally, 23 sensors were detected.

supported this finding by the distributions of the three properties in every interval over the considered conditions, which could not be distinguished between the classes of metabolites and metabolite pairs (Supplemental Fig. S3).

It has already been observed that the sensors predicted by the approach we used may change upon alterations of the reaction directionality (Liu et al., 2013). Therefore, in the second analysis or robustness, we tested the effect of randomizing the reversibility of reactions considered in the model. The 500 randomizations were performed while preserving the number of reversible and irreversible reactions in the network together with the set of metabolites they interconvert. The original sensors in AraCORE consisted of 23 metabolites, whereas after randomization we found between 25 and 42 sensors, of which 11–22 (i.e. at least 48%, see Fig. 6) were also present in the original set of sensor metabolites. Moreover, each of the sensors was identified in at least one randomization. Therefore, the results supported the robustness of the identified sensors and were in line with existing studies, which have pointed out that reversibility of biochemical reactions had a small effect on the identified sensors (Liu et al., 2011). Similar results were obtained for the second model; here, after randomization, we found between 108 and 153 sensors, of which 57–90 (i.e. at least 28.79%) were identical with the 198 sensors in the original network. The overlap with the original sensors was lower, compared to the other two models; nevertheless, we could capture, in more than half of the permutations, a >36% overlap. These results were also partly in support of our claim for the robustness of sensors.

Test on Kinetic Model of Central Carbon Metabolism

To further validate the finding that sensor metabolites are more correlated with each other than nonsensor metabolites, we repeated the analysis with a synthetic data set generated from a medium-size kinetic model of plant central carbon metabolism. The model included the Calvin-Benson-Cycle, triose phosphate transport, Suc biosynthesis and degradation, starch biosynthesis and degradation, photorespiration, ATP synthesis, and the photosynthetic electron transport distributed over five compartments. It comprised 78 metabolites and 112 reactions, representing the largest kinetic model of plant central metabolism to date (Hahn, 1986; Singh and Ghosh, 2006). This model, however, does not contain the TCA cycle and the vast majority of amino acids. The reaction rates were modeled according to mass action kinetics (see Supplemental Kinetic Model for the stoichiometric matrix and reaction parameters).

In this case, we identified six sensor metabolites solely using the approach based on the network structure, including 2-oxoglutarate, Ser in the mitochondrion, Suc in the cytosol, and the vacuole, as well as hydrogen peroxide H_2O_2 and ammonia. Based on the simulated data profiles (by varying the initial conditions), we again found that the correlation within sensors was higher than within nonsensor metabolites, for both day (d) and night conditions (Figs. 7 and 8). The SD of the sensors was in both cases higher than for the nonsensors, similar to the results of the AraSEED model. The results of the CV differed between d and night simulations. The d simulation showed a pattern that was comparable to AraCORE (see Fig. 7), while the night simulations showed similarities to the AraSEED model results (see Fig. 8). Altogether, the findings from the simulated data profiles from a medium-size kinetic model were in line with the data gathered from experiments, particularly with respect to the observed ordering of correlations within and between groups of metabolites.

Implications of the Findings

In this study we demonstrated that metabolites identified as sensors were more correlated than nonsensor metabolites based on data profiles gathered from wet-lab experiments as well as in silico simulations. Most of the findings were independently reproduced for two well-curated models of Arabidopsis. Furthermore, we showed that this was not due to an artifact of the used data by an extensive robustness analysis. By randomly assigning the labels "sensor" and "nonsensor" to the metabolites in the analyzed data set, we demonstrated that the correlation of sensors and nonsensors, and between sensors and nonsensors, could no longer be observed. Additionally, we tested the influence of reversible reactions in metabolic networks. Importantly, we could reproduce these results on a kinetic model of medium size used for simulating a synthetic data set. Using a

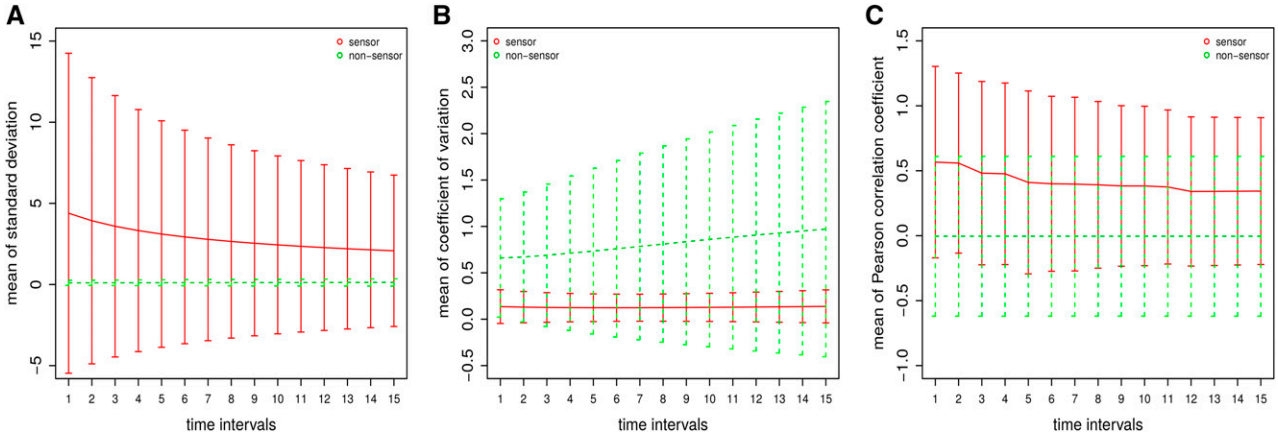

Figure 7. Statistical comparison of the sensors and nonsensors in the kinetic *d*-time model of plant central carbon metabolism. The *x* axis represents the investigated time interval, from 1 to 15. The *y* axis represents values for the three statistics, respectively: A, SD; B, CV; and C, Pearson correlation of sensors and nonsensors. The red line corresponds to the values for the statistics between sensor metabolites, while the green line corresponds to values between nonsensor metabolites. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and nonsensors. Bars represent the range ± 1 SD from the mean value, for five simulations.

random but physiologically viable initial condition for the simulation of d and night cycles, we found the same relationship between sensor metabolites and nonsensor metabolites as in the Arabidopsis large-scale models.

The identified sensors in all models were metabolites that act as major building blocks of biomass. In the smaller AraCORE, we found cellulose for cell wall synthesis and most of the amino acids for the protein biosynthesis, as well as nucleotides for DNA and RNA replication. These were in agreement with the results of the genome-scale Arabidopsis model, AraSEED, in which, in addition to the mentioned metabolite classes, we also identified Coenzyme A and related metabolites playing important roles in the tricarboxylic acid cycle (Fatland et al., 2002).

Our results largely depend on the quality of the networks employed. Therefore, we critically investigated the network models used and found that a large number of sensor metabolites were in fact dead-end metabolites, created upon removal of the biomass reactions. By consideration of the respective biomass reaction, the identified metabolites were not dead-end metabolites just in AraCORE.

In AraSEED with a biomass reaction, 4 of the 15 sensors were dead-end metabolites. Investigation of the large-scale metabolic networks used in the study of

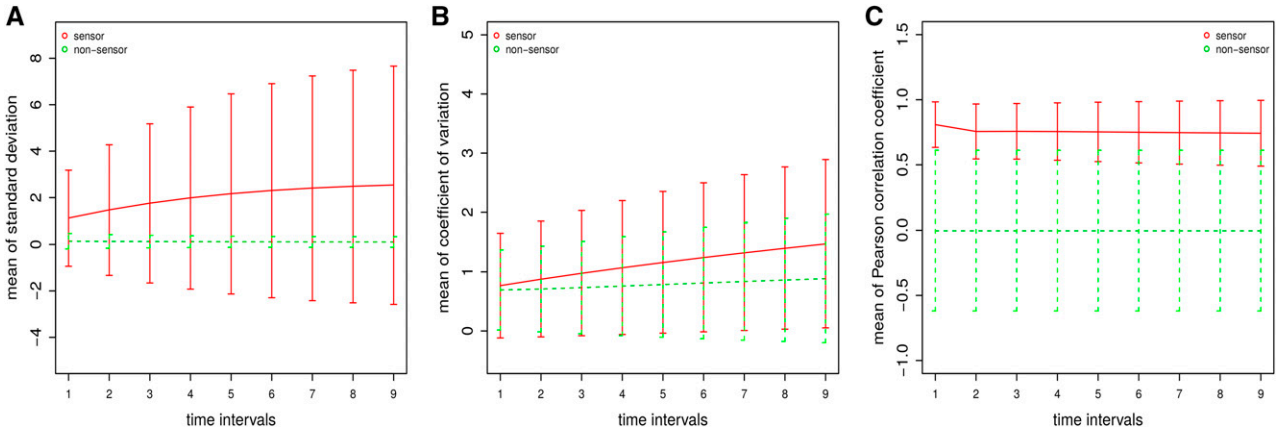

Figure 8. Statistical comparison of the sensors and nonsensors in the kinetic night-time model of plant central carbon metabolism. The *x* axis represents the investigated time interval, from 1 to 9. The *y* axis represents values for the three statistics, respectively: A, SD; B, CV; and C, Pearson correlation of sensors and nonsensors. The red line corresponds to the values for the statistics between sensor metabolites, while the green line corresponds to values between nonsensor metabolites. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and nonsensors. Bars represent the range ± 1 SD from the mean value, for five simulations.

Liu et al. (2013) showed similar results: In the human RECON1 model, yeast, and *Escherichia coli* models, 57.04%, 76.92%, and 59.81% of the sensors were dead-end metabolites. This is in line with a claim of Liu et al. (2013) that all pure products, i.e. metabolites that do not act as reactants in a single reaction, can serve as sensors. A potential explanation of these high numbers of blocked reactions is that most models contain only an incomplete set of catabolic reactions. Therefore, more metabolites may be predicted as sensors by this approach, as degrading reactions might be missing.

While our empirical tests were built around time-courses within single genotypes, we demonstrated that similar relationships among our predicted sensors could be found in genetic populations of Arabidopsis. Analogs to these results have also been observed in correlation-based network analysis of metabolic profiles from a tomato (*Solanum lycopersicum*) introgression line-mapping population, where five amino acids (i.e. Gly, Ile, Ser, Thr, and Val) were significantly more correlated (average value of 0.84) in comparison to the average correlation between any other measured metabolites (Toubiana et al., 2012, 2015). Thus, our predicted sensors may be useful to understand the correlations arising in genetically variable populations.

CONCLUSION

In this work, we aimed to identify if there are features of the data profiles of sensor metabolites, identified with well-established network-based approaches, that separate them from the rest of the metabolites in a given large-scale plant metabolic network. Methods from observability theory allow computationally feasible identification of sensor metabolites; however, the existing studies have not investigated the extent to which the data profiles of sensors may differ from those of nonsensor metabolites. By employing experimentally and in silico generated time-series metabolomics data together with large- and medium-scale structural and kinetic models of Arabidopsis central metabolism (Dall'Osto et al., 2012), we demonstrated that sensor metabolites are, on average, more correlated than nonsensor metabolites across employed models and data sets. Our analyses of robustness further confirmed that these results were due to the position of the sensor metabolites in the network, and complement the implications from other approaches. These correlations tend to persist irrespective of the conditions as long as the underlying functionality of the network, a result of the set of the operational biochemical reactions, remains largely unchanged, as illustrated on data from natural variation. As a result, our study suggests that relatively few key metabolites could be measured to potentially characterize the entire metabolic network, opening the possibility for applications of targeted metabolite analyses guided by predictions from large-scale models as a means of providing a rapid yet accurate synopsis of the metabolic status of a plant system.

MATERIALS AND METHODS

Our analysis is based on the graphical approach of Liu et al. (2013). The sensor metabolites can be determined by building the inference graph obtained from a given network of biochemical reactions under the assumption that their rates are described by mass action kinetics. The nodes in the inference graph are given by the metabolites. For instance, the network in Figure 1A contains seven metabolites, denoted by A–G, transformed via nine reactions with rate constants $k_1 - k_9$. A node (i.e. metabolite) u is connected by directed edge to node v if metabolite v occurs in the differential equation for metabolite u. To illustrate the building of the inference graph, we again turn to the network of biochemical reactions in Figure 1: Because A appears on the right-hand side of the differential equation for A (i.e. dA/dt on Fig. 1B), there is a directed edge from A to itself (Fig. 1C), Similarly, there is a directed edge from node A to node B because A appears in the differential equation for metabolite B, The inference graph can be decomposed into its SCCs. An SCC is the maximal subgraph for which there are directed paths from every node to all others. For instance, nodes B and D form an SCC because there is an edge from B to D as well as from D to B. However, E and C are not in an SCC because there is no directed path from C to E, although there is a path from E to C (Fig. 1C). If an SCC does not have an incoming edge, it is referred to as a "root" SCC. In our toy example, B and D as well as E and F form two root SCCs, while A, C, and G form three nonroot SCCs.

Liu et al. (2013) showed that the sensors are located in this set of nodes in the root SCCs. The set of nodes obtaining by selecting at least one node from each root SCC then allows complete observability of the system. A similar framework has also been applied and discussed in Rios et al. (2013). The approach can be readily applied to any genome-scale metabolic network because the inference graph can be built only from the stoichiometric matrix, as input. To determine the edges that start at a node u, it suffices to identify the substrate metabolites of the reactions in which the metabolite u participates as a substrate or product. The substrates of a reaction are readily given by the negative entries of the corresponding reaction vector in the stoichiometric matrix. For instance, node B participates in reactions with B, D, and A as a substrate, and, thus, there are directed edges to these nodes from B. We used the R package Igraph (Csárdi and Nepusz, 2006) to build the inference graph and to find its (root) SCCs.

A root SCC may not consists of a single metabolite, as is the case on the toy network in Figure 1C. In this case, for the root SCC consisting of B and D, any of the two can serve as a sensor. We applied the graphical approach to two genome-scale metabolic networks of Arabidopsis (*Arabidopsis thaliana*), the bottom-up assembled Arabidopsis core model, AraCORE (Arnold and Nikoloski, 2014), and the Arabidopsis model from PlantSEED (Seaver et al., 2014), referred to as AraSEED. Both networks cover pathways of plant primary metabolism. We analyzed these models, whose characteristics appear in Supplemental Table S7, with and without consideration of biomass and sink reactions. The sensor metabolites were selected from the root SCCs as those that could be mapped to the available metabolic profiles. In our study, this resulted in a single sensor node identified per root SCC (see Supplemental Table S1 and 3 for lists of identified sensors in the two models and Supplemental Tables S5 and S6 for lists of mapped metabolites).

To relate the predicted sensors to metabolic measurements, we obtained metabolic profile data from Caldana et al. (2011) generated by gas chromatography-mass spectroscopy (GC-MS). This metabolic data set consists of 91 metabolites measured under the following conditions: 21°C at 75 μE m^{-2} s^{-1}; 150 μE m^{-2} s^{-1} light intensity and darkness; 4°C at 85 μE m^{-2} s^{-1} light intensity and darkness; and 32°C at 150 μE m^{-2} s^{-1} and darkness. Therefore, the analyzed data set consisted of metabolic time series covering 20 time points and gathered under seven conditions. In addition, to augment the set of tested conditions, we used metabolic data profiles from a study of natural variation in central carbon metabolism of Arabidopsis (Sulpice et al., 2013). In this study, the data profiles of 45 metabolites were measured in 97 Arabidopsis lines in three conditions, namely 8 h of light with high N supply, 12 h of light with high N supply, and 12 h of light with low N supply. Metabolite data were acquired using GC-MS technology. A detailed description of the plant growth conditions and experimental design can be found in the "Materials and Methods" of Sulpice et al. (2013). The two studies whose data sets we used here performed their GC-MS experiments as outlined in Lisec et al. (2006).

These data sets allow first insights, to our knowledge, into how the sensors relate to the rest of the measured metabolome under a variety of genotypes, environmental conditions, and over time. For the statistical analysis, we tested for differences in the means of correlation values between the two groups of sensor and nonsensor metabolites by two-sided t test at a significance level of $\alpha = 0.05$. A graphical representation of the complete workflow applied in this study is visualized in Supplemental Figure S1.

Genome-scale metabolic networks are open systems, in contrast to the closed systems (i.e. without in- and out-flux reactions) considered by Liu et al. (2013) and Rios et al. (2013). Because an open system has additional self-edges at the output metabolites in the inference graph, these have no effect on the identification of root SCCs (see Supplemental Material and Liu et al. (2013)). Figure 1 illustrates an open system in which the root SCCs remain unaffected if the import and export reactions are removed. To identify dead-end metabolites given a large-scale network, we used the COBRA toolbox function *removeDeadEnds* in MATLAB (Schellenberger et al., 2011).

Supplemental Materials

The following supplemental materials are available.

Supplemental Figure S1. Schematic overview of the procedure.

Supplemental Figure S2. Comparison of Pearson correlation values of sensors (red), nonsensors (green), and between sensors and nonsensors (blue).

Supplemental Figure S3. Statistical comparison after randomizing the sensors and nonsensors.

Supplemental Figure S4. Comparison of AraSEED sensors depending on nonroot SCC connection.

Supplemental Table S1. Sensor metabolites in the AraCORE model.

Supplemental Table S2. Sensor metabolites in the AraCORE model with biomass function.

Supplemental Table S3. Sensor metabolites in the AraSEED model.

Supplemental Table S4. Sensor metabolites in the kinetic model.

Supplemental Table S5. AraCORE—mapped sensors and nonsensors.

Supplemental Table S6. AraSEED—mapped sensors and nonsensors.

Supplemental Table S7. Overview of the models used.

Supplemental Material. Kinetic model.

Supplemental Material. Data for the kinetic model.

Received June 6, 2016; accepted August 23, 2016; published August 26, 2016.

LITERATURE CITED

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci **19:** 52–61

Arnold A, Nikoloski Z (2014) Bottom-up metabolic reconstruction of *Arabidopsis* and its application to determining the metabolic costs of enzyme production. Plant Physiol **165:** 1380–1391

Caldana C, Degenkolbe T, Cuadros-Inostroza A, Klie S, Sulpice R, Leisse A, Steinhauser D, Fernie AR, Willmitzer L, Hannah MA (2011) High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions. Plant J **67:** 869–884

Chang Y, Suthers PF, Maranas CD (2008) Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments. Biotechnol Bioeng **100:** 1039–1049

Chaves M, Sontag ED (2002) State-estimators for chemical reaction networks of Feinberg-Horn-Jackson zero deficiency type. Eur J Control **8:** 343–359

Cosgrove DJ (2005) Growth of the plant cell wall. Nat Rev Mol Cell Biol **6:** 850–861

Csárdi, G, Nepusz, T (2006). The igraph software package for complex network research. Int J Complex Sys M/S no 1695 2006 http://igraph.org

Dall'Osto L, Holt NE, Kaligotla S, Fuciman M, Cazzaniga S, Carbonera D, Frank HA, Alric J, Bassi R (2012) Zeaxanthin protects plant photosynthesis by modulating chlorophyll triplet yield in specific light-harvesting antenna subunits. J Biol Chem **287:** 41820–41834

de Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK (2010a) AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. Plant Physiol **152:** 579–589

de Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK (2010b) C4GEM, a genome-scale metabolic model to study C4 plant metabolism. Plant Physiol **154:** 1871–1885

Fatland BL, Ke J, Anderson MD, Mentzen WI, Cui LW, Allred CC, Johnston JL, Nikolau BJ, Wurtele ES (2002) Molecular characterization of a heteromeric ATP-citrate lyase that generates cytosolic acetyl-coenzyme A in *Arabidopsis*. Plant Physiol **130:** 740–756

Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet **25:** 39–48

Fiehn O (2002) Metabolomics—the link between genotypes and phenotypes. Plant Mol Biol **48:** 155–171

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol **22:** 245–252

Gu L, Jones AD, Last RL (2010) Broad connections in the *Arabidopsis* seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. Plant J **61:** 579–590

Gu L, Jones AD, Last R (2012) Rapid LC–MS/MS profiling of protein amino acids and metabolically related compounds for large-scale assessment of metabolic phenotypes. *In* MA Alterman, P Hunziker, eds, Amino Acid Analysis: Methods and Protocols. Humana Press, New York

Hahn BD (1986) A mathematical model of the Calvin cycle: analysis of the steady state. Ann Bot (Lond) **57:** 639–653

Jha S, van Schuppen JH (2001). Modelling and control of cell reaction networks. PNA-R0116 (CWI). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.4201

Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotechnol **14:** 491–496

Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat Protoc **1:** 387–396

Liu Y-Y, Slotine J-J, Barabási A-L (2011) Controllability of complex networks. Nature **473:** 167–173

Liu Y-Y, Slotine J-J, Barabási A-L (2013) Observability of complex systems. Proc Natl Acad Sci USA **110:** 2460–2465

Lu Y, Savage LJ, Larson MD, Wilkerson CG, Last RL (2011) Chloroplast 2010: a database for large-scale phenotypic screening of *Arabidopsis* mutants. Plant Physiol **155:** 1589–1600

Nöh K, Grönke K, Luo B, Takors R, Oldiges M, Wiechert W (2007) Metabolic flux analysis at ultra short time scale: isotopically non-stationary ^{13}C labeling experiments. J Biotechnol **129:** 249–267

Rios DF, Shirin A, Sorrentino F (2013) The network observability problem: detecting nodes and connections and the role of graph symmetries. ARXIV arXiv:1308.5261

Saha R, Suthers PF, Maranas CD (2011) *Zea mays* iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. PLoS One **6:** e21784

Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, et al (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc **6:** 1290–1307

Schönbrodt FD, Perugini M (2013) At what sample size do correlations stabilize? J Res Pers **47:** 609–612

Seaver SMD, Gerdes S, Frelin O, Lerma-Ortiz C, Bradbury LMT, Zallot R, Hasnain G, Niehaus TD, El Yacoubi B, Pasternak S, Olson R, Pusch G, et al (2014) High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. Proc Natl Acad Sci USA **111:** 9645–9650

Singh VK, Ghosh I (2006) Kinetic modeling of tricarboxylic acid cycle and glyoxylate bypass in *Mycobacterium tuberculosis*, and its application to assessment of drug targets. Theor Biol Med Model **3:** 27

Sulpice R, Nikoloski Z, Tschoep H, Antonio C, Kleessen S, Larhlimi A, Selbig J, Ishihara H, Gibon Y, Fernie AR, Stitt M (2013) Impact of the carbon and nitrogen supply on relationships and connectivity between metabolism and biomass in a broad panel of *Arabidopsis* accessions. Plant Physiol **162:** 347–363

Templeton GW, Moorhead GBG (2004) A renaissance of metabolite sensing and signaling: from modular domains to riboswitches. Plant Cell **16:** 2252–2257

Toubiana D, Batushansky A, Tzfadia O, Scossa F, Khan A, Barak S, Zamir D, Fernie AR, Nikoloski Z, Fait A (2015) Combined correlation-based network and mQTL analyses efficiently identified loci for branched-chain amino acid, serine to threonine, and proline metabolism in tomato seeds. Plant J **81:** 121–133

Toubiana D, Semel Y, Tohge T, Beleggia R, Cattivelli L, Rosental L, Nikoloski Z, Zamir D, Fernie AR, Fait A (2012) Metabolic profiling of a mapping population exposes new insights in the regulation of seed metabolism and seed, fruit, and plant relations. PLoS Genet **8:** e1002612

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiol **144:** 32–42