Starch as a major integrator in the regulation of plant growth

Ronan Sulpice,a*, Eva-Theresa Pyla, Hirofumi Ishiharaa, Sandra Trenkampa, Matthias Steinfatha,b, Hanna Wittucka-Wallf, Yves Gibona, Björn Usadela, Fabien Poreea, Maria Conceição Piques, Maria Von Korff, Marie Caroline Steinhausera, Joost J. B. Keurentjesd,e, Manuela Guenthera, Melanie Hoehne, Joachim Selbigb, Alisdair R. Ferniea, Thomas Altmanna, and Mark Stittb

aMax Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam-Golm, Germany; Departments of bBioinformatics and cGenetics, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Strasse 24–25, 14476 Potsdam, Germany; dGenetics Laboratory and Laboratory of Plant Physiology, Wageningen University and Research Centre, NL-6703 BD Wageningen, The Netherlands; and eCentre for BioSystems Genomics, NL-6708 PB Wageningen, The Netherlands

Edited by Joseph R. Ecker, The Salk Institute for Biological Studies, La Jolla, CA, and approved April 29, 2009 (received for review April 3, 2009)

Rising demand for food and bioenergy makes it imperative to breed for increased crop yield. Vegetative plant growth could be driven by resource acquisition or development programs. Metabolite profiling in 94 Arabidopsis accessions revealed that biomass correlates negatively with many metabolites, especially starch. Starch accumulates in the light and is degraded at night to provide a sustained supply of carbon for growth. Multivariate analysis revealed that starch is an integrator of the overall metabolic response. We hypothesized that this reflects variation in a regulatory network that balances growth with the carbon supply. Transcript profiling in 21 accessions revealed coordinated changes of transcripts of more than 70 carbon-regulated genes and identified 2 genes (myo-inositol-1-phosphate synthase, a Kelch-domain protein) whose transcripts correlate with biomass. The impact of allelic variation at these 2 loci was shown by association mapping, identifying them as candidate lead genes with the potential to increase biomass production.

Arabidopsis | association mapping | biomass | metabolites | predictive

Plants use light energy to convert CO2 into carbohydrates. Although we might expect plant growth to be driven by the availability of carbohydrates and other central metabolites, recent studies point to a more complex interaction. Numerous free air CO2 elevation studies show that higher rates of photosynthesis do not lead to a commensurate increase in biomass and yield (1). Studies of natural genetic diversity reveal a negative correlation between the levels of metabolites and biomass or yield (2–4). Although biomass was only weakly correlated with individual metabolites in an Arabidopsis recombinant inbred line (RIL) population, a highly significant prediction was obtained when multivariate analysis was used on the entire metabolite profile (3). These results indicate that much of the genetic variation for biomass production affects the balance between resource availability and developmental programs, which determine how rapidly these resources are used for growth.

Partial Correlation Analysis to Remove Spurious Correlations. Because many metabolic traits correlate with each other (2), some

The following experiments test the hypothesis that starch turnover and C allocation occupy a central role in the network that coordinates metabolism with growth. We first investigate biomass and metabolite levels in 94 Arabidopsis accessions. This species-wide analysis reveals that starch content at the end of the day integrates many other metabolic traits and is negatively correlated with biomass. We then compare the expression of C-responsive transcripts in 21 accessions, identify candidate genes that may contribute to genetic variation in the regulation of metabolism and growth, and test their role by association mapping of sequence polymorphisms.

Results and Discussion

Many Metabolites Are Negatively Correlated to Biomass. Over 400 Arabidopsis thaliana accessions were genotyped with 419 markers (13) to identify a genotypically diverse set of 94 accessions with maximized allelic richness (Table S1). The accessions were grown in short-day conditions (8 h light/16 h dark) in moderate light and well-fertilized soil to apply a moderate C deprivation. They were harvested at the end of the day, 5 weeks after germination when they were still in the vegetative growth phase. Rosette fresh weight (FW) was measured as an indicator of biomass. We have documented a very close relation between rosette FW and rosette dry weight (2). We analyzed starch, total protein, chlorophyll, and 48 low-molecular-weight metabolites, including individual amino acids, organic acids, sugars, lipids, and secondary metabolites (Table S1). Pair-wise Spearman’s correlations were calculated for biomass against every metabolic trait (Table 1). Rosette biomass showed a high negative correlation to starch (R = −0.54); lower but significant negative correlations with protein (R = −0.37), chlorophyll (R = −0.31), and several low-molecular-weight metabolites (sucrose, total amino acids, glycine, alanine, glutamate, threonine acid, benzoic acid, sinapic acid); and nonsignificant negative correlations with other metabolites.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Ferily available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the EMBL Nucleotide Sequence database (accession nos. FM998553–FM998644 for At4g39800 and FM995274–FM995365 for At1g23930).

1To whom correspondence should be addressed. E-mail: sulpice@mpimp-golm.mpg.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0903478106/DCSupplemental.
of the correlations with biomass may be secondary. Partial Correlation Analysis was performed to correct for spurious secondary correlations (Fig. 1A). The analysis confirmed the link between biomass and starch but did not provide evidence for direct links of biomass to any other individual metabolic traits. Some links were found between metabolites; starch was linked to sucrose, glucose was linked to fructose but not to sucrose or starch, several amino acids were linked, and raffinose was linked to galactinol and myo-inositol, which are involved in its synthesis.

The negative correlation between biomass and starch was not due to population structure. Using Structure 2.1 (14) on 419 markers distributed across the whole genome, the smallest K value for highest posterior probability split the population into 7 subpopulations (Table S1). These 7 subpopulations had similar average values for biomass and starch. R values between starch and biomass were less than −0.63 in 3 subpopulations (containing 61 accessions), less than −0.42 in 2 subpopulations (containing 25 accessions), and less than −0.24 in the other 2 subpopulations (containing 11 accessions).

Partial Least Squares (PLS) Regression Reveals That Starch Integrates the Metabolic Status. It has been shown that predictive power can be increased by using multivariate analysis to predict biomass from a linear combination of a set of low-molecular-weight metabolites (3). We investigated whether this was the case in our study. In datasets like ours, where the number of predictors (54) is close to the number of accessions (94), the predictive power of linear models is often improved by dimensionality-reduction methods like PLS regression. PLS identifies combinations of the original predictors that have the maximum covariance with the trait of interest. These orthogonal combinations are then used to predict the trait. The PLS prediction of biomass was significantly (F-test, P = 0.002) but only slightly improved when all of the metabolic traits were used as predictors, compared with the prediction provided by starch alone (R = 0.57 and 0.50, respectively; the value for starch was checked by cross-validation, hence the sign is absent and the value slightly lower than in Table 1).

To investigate the reasons for this lack of additivity, we divided the dataset into 3 univariate traits (biomass, starch, total protein) and 1 multivariate trait class (all other metabolites). The univariate traits were compared using linear regression, and the multivariate trait class was used to predict each univariate trait using PLS regression (Fig. 1B). The predictive accuracy of each pair-wise comparison was assessed by cross-validation (see Materials and Methods). Starch and total protein showed a significant negative correlation with biomass (R = 0.50 and 0.32, respectively) and correlated weakly with each other (R = 0.22). PLS regression on the multivariate metabolite class allowed prediction of starch, protein, and biomass (R = 0.38, 0.30, and 0.32), respectively. Variance importance in the projection (VIP) gives an estimate of the contribution of a given predictor for a PLS regression. Starch, protein, and rosette biomass (Fig. 1C and D) were predicted by the same metabolites, with remarkably similar VIP values (for a full list, see Table S2).

This analysis shows that starch and, to a lesser extent, total protein integrate metabolic status. It also indicates that the regulatory network that determines starch and protein levels contributes to the regulation of biomass. To provide functional information about this network, we subjected the reference genotype Col0 and 10 large and 10 small accessions (listed in Table S3) to a more detailed physiological and molecular analysis.

Table 1. Spearman coefficients of metabolic traits against biomass

<table>
<thead>
<tr>
<th>Structural components</th>
<th>Amino acids and derivatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein content</td>
<td>Glycine</td>
</tr>
<tr>
<td>Chl b</td>
<td>L-Alanine</td>
</tr>
<tr>
<td>Total chlor</td>
<td>Amino acids derivatives</td>
</tr>
<tr>
<td>Chl a</td>
<td>Valine</td>
</tr>
<tr>
<td></td>
<td>Glutamate</td>
</tr>
</tbody>
</table>

Carbohydrates

| Starch                | Arginine                     |
|                      | Aspartate                    |
|                      | Asparagine                   |
|                      | Lysine                       |
|                      | Proline                      |
|                      | Serine                       |
|                      | β-Alanine                    |
|                      | Leucine                      |
|                      | Proline                      |
|                      | Tryptophane                  |
|                      | Isoleucine                   |
|                      | Tyrosine                     |
|                      | Methionine                   |

Other metabolites

| Threonine            | Urea                         |
|                      | Guanidine                    |

Sugar-Responsive Gene Network. Changes in the C supply modify the transcript levels for hundreds of genes during the diurnal cycle (10, 11, 15–17). We asked whether C-responsive genes show coordinated changes of expression between the 21 accessions and whether any of their transcripts correlate with the levels of major metabolites or biomass. Two nonoverlapping sets of C-responsive genes were selected (Fig. S2) from published data, one including 52 genes whose transcripts change >log21.4 within 30 min of adding sucrose to C-starved seedlings (18) and one containing another 42 genes that show changes of expression during a diurnal cycle and an extended night (15, 17). Transcript levels were measured at the end of the night (Table S3), when changes in the C status have the largest impact (15, 17). Principal component analysis (PCA) gen-

Sulpice et al.
erated a similar separation of the accessions (R = 0.93) in the first component, irrespective of which gene set was used (see Fig. S3). There was a significant correlation between the separation of accessions along the first component and biomass (R = 0.53, \( p = 0.013 \)). The finding that transcript levels of C-regulated genes possess predictive power for biomass and that an almost identical result is given by 2 nonoverlapping sets of genes encouraged us to pursue the analysis.

**Combined Network for Transcripts, Metabolites, and Biomass.** We used an algorithm that identifies functional modules within complex networks (19) to generate a correlation network that combines transcript levels, metabolite levels, and biomass (Fig. 2). This algorithm defines a module as a subset of nodes that are more connected to each other than to nodes in other modules. Starting from the initial state, in which each node represents a module, it performs iterations of merging, splitting, and transferring nodes between modules to maximize the interconnectivity of edges within modules and, thus, the modularity of the network. The resulting network contained 71 of the 94 genes investigated in this study. They were organized in 2 large, well-connected modules, a smaller module and several nodes that are only connected by 1 to 2 edges. One large module contains mainly C-repressed genes (23 of 28), and the second contains 16 C-repressed and 12 C-induced genes. Thus, most C-responsive genes show coordinated changes of expression across this set of 21 accessions.

**Comparison with a Transcript Network Obtained by Perturbing C Status in Col0.** We compared the correlation network in Fig. 2 with a correlation network for the same 94 transcripts, which we generated from data obtained in earlier studies where we subjected 5-week-old rosettes of the reference accession Col0 to 23 treatments that alter endogenous C levels (15, 17, 18). The network that was obtained will be termed the ‘C-perturbation’ network. When we compared the R value for each gene–gene pair in the 2 correlation networks, data shuffling revealed a significant enrichment of shared positive (\( R = 6 \times 10^{-3} \)) and shared negative (\( R = 2.4 \times 10^{-3} \)) correlations (Fig. S4A). There was a lower, but still significant, enrichment of shared correlations with networks generated from a Col0 leaf developmental series (\( R = 3.5 \times 10^{-3} \)) (Fig. S4B) and a rosette abiotic stress series (\( R = 2 \times 10^{-4} \)) (Fig. S4C), but no significant enrichment in a root abiotic stress series (\( R = 0.21 \)) (Fig. S4E).

We identified 26 genes where >30% of the significant gene–gene correlations are shared in the accession and C-perturbation networks (Table S5). They include many potential regulatory components, including bZIP transcription factors, F-box proteins, ATG8e, and 2 BTB/POZ proteins. Recent evidence implicates trehalose-
6-phosphate (Tre6P) (17, 20, 21) and AKN10/11 (17, 20) in C signaling. The 26 genes included 3 members of the trehalose-phosphate synthase gene family (TPS8, TPS10, TPS11) and 19 genes whose transcripts respond to AKN10 overexpression, including bZIP1/AT5G49450, a direct target of AKN10 (20). This overlap with known upstream components in sugar signaling supports the validity of our network approach. This set of genes overlap with known upstream components in sugar signaling including IPS1/AT5G49450, a direct target of AKN10 (20). This overlap with known upstream components in sugar signaling supports the validity of our network approach.

Candidate Gene Allelic Variation. This correlation indicated that sequence diversity in IPS1/At4g39800 and At1g23390 might influence biomass. Genomic DNA, containing 1065 base pairs of the promoter and the entire transcribed region of At1g23390, was amplified and sequenced from 92 of the 94 acces-sions. We detected 95 SNPs, 9 insertion-deletion polymorphisms (Indels), and 1 simple sequence repeat (SSR) in Atlg23390 and 79 SNPs, 7 Indels, and 1 in SSR in At4g39800 (Table S6). The frequency of single, rare (<5), minor (>5), and major (>15) polymorphisms was 34%, 19%, 15%, and 32% for At1g23390 and 38%, 28%, 17%, and 17% for At4g39800, respectively. Linkage disequilibrium (LD) was calculated by using Graphical Genotype (23). Rare polymorphisms were excluded for the calculation. Both genes contained a set of polymorphisms in strong LD in the transcribed region and further polymorphisms in weaker LD in the promoter region (Table S6; a summary is provided in Fig. 3). The decay of LD to below 0.2 in 1–2 kb is in the range reported in ref. 24. Repeat number for the SSR in the At1g23390 promoter varied independently of loci in the remainder of the gene, probably because changes in repeat number can occur independently of, and more rapidly than, recombination.

Sequence diversity in these genes is not closely tied to population structure. When we K-clustered the accessions into 7 classes based on the polymorphisms in each of the genes and compared these with the 7 classes obtained after K-clustering (see Materials and Methods and Fig. S1), the overlap was only slightly larger than that expected by chance (2.7%).

Candidate Gene Association Mapping. Using these polymorphisms and the trait values for rosette biomass, amino acid, starch, sucrose, and protein, marker trait association (Fig. 3) was determined in TASSEL (25) with a general linear model (GLM) using the population structure from Fig. S1 to control for population-structure effects (see Materials and Methods). Adjusted P values were calculated by using 10,000 permutations to correct for multiple testing.

For At1g23390, the SSR in the promoter (2 classes were formed with <12 and ≥12 repeats) was significantly associated with FW and starch. Ten SNPs in the transcribed region showed significant associations with protein and sucrose. They were in strong LD, and 7 caused nonsynonymous changes in the coding region (285, Asn/Lys; 680, Pro/Leu; 761, Val/Gly, 775, Met/Val; 826, Glu/Lys; 977, Gly/Asp; 1074, Lys/Asn). They were also in LD with 3 SNPs in the promoter (−1084, −845, −299) that showed associations with sucrose or amino acids and an SNP in the 3’ UTR (position 1380) that associated with sucrose.

For At4g39800, 3 adjacent SNPs in the promoter (−786, −785, and −784) associated with rosette biomass, amino acids, and starch.
Another 4 SNPs in exon/intron regions of the gene and a SNP in the downstream region (2298) associated with rosette biomass, protein, and starch. These 5 SNPs were in strong LD, and one led to a nonsynonymous change in the coding sequence (699, Asp/Asn). As a control, we performed association mapping for the 419 markers used to genotype the population (see Materials and Methods), using the same trait and structure population datasets. No significant associations were found to any trait.

The large number of polymorphisms prevented association mapping against all individual haplotypes. Instead, we built haplotypes from the polymorphisms that showed associations and led to nonsynonymous changes. For At4g39800, we used position −784 (promotor) and position 699 (Asp/Asn). There were significant associations for FW (P = 0.0065; R²M = 4.8%), starch (P = 0.043; R²M = 3.1%), protein (P = 0.029; R²M = 3.1%), and amino acids (P = 0.0003; R²M = 7%). For At1g23390, we used the SSR in the promoter and a set of 9 SNPs in strong LD in the ORF (680, 761, 775, 826, 864, 972, 977, 1026, 1074). They generated 6 haplotypes, with significant associations for FW (P = 0.008; R²M = 7.4%), protein (P = 0.048; R²M = 4.9%), sucrose (P = 0.0004; R²M = 9.7%), and, at a lower level of significance, starch (P = 0.085; R²M = 4.8%).

These results indicate interactive functions for the promotor and ORF of these 2 genes. In all cases where associations were found, a given allele or haplotype displayed an opposite effect on biomass and metabolic traits.

Finally, we checked whether any of the polymorphisms in the promotor correlated with transcript levels. Genotype information was available for 20 of the accessions in which transcript levels were determined. For At4g39800/IPS1, the minor (non-Col0) alleles at −785, −784, and −410 in the promoter region were found in 6, 8, and 2 of the accessions. They correlated with higher IPS1 transcript levels (P = 0.02, 0.05, and 0.02, respectively) (Table S6). For At1g23390, no significant correlations were found at P > 0.05, but several polymorphisms in the promotor region correlated at P < 0.1.  

Concluding Remarks. There is increasing interest in the possibility of using biomarkers to predict plant biomass. Meyer and colleagues showed that biomass can be predicted by a set of low-molecular-weight metabolites (3), but their study was restricted to a single biparental RIL population and did not reveal why this set of metabolites have predictive power. We show that metabolite levels change reciprocally to biomass across a large set of genotypically diverse Arabidopsis accessions. Further, and importantly, the changes of metabolites are integrated as changes in the level of starch and, to a lesser extent, protein. This finding has the practical advantage that starch can be easily extracted and assayed. Using robotized systems (8), we can precisely measure starch levels in 400 samples per day. This will make it possible to identify genotypes where changes in biomass production are, and are not, connected to changes in central metabolism.

It also points to a biological explanation for the negative relation between biomass and metabolites, namely, that large accessions have a modified balance between the C supply and growth, which is integrated as a change in starch levels. In agreement, profiling of C-regulated transcripts revealed coordinated changes of many C-responsive transcripts between Arabidopsis accessions, including genes involved in Tre6P and Akin10 signaling (16, 20, 21, 26). This hypothesis-driven approach also identified 2 candidate genes, encoding a myo-inositol-1-phosphate synthase and an unusual Kelch repeat-containing F-box family protein, whose transcript levels correlate with rosette biomass. Association mapping revealed polymorphisms in these genes that are related to rosette biomass and show opposite allelic effects on metabolites, including starch and protein. It has already been shown that antisense inhibition of a homolog to IPS1/At4g39800 in potato leads to increased levels of sucrose and starch, altered leaf morphology, precocious senescence, and decreased tuber yield (27), as expected if this enzyme or its products contribute to the regulation of C partitioning and growth.

Starch is a C-storage polymer without demonstrated regulatory activities. It is more likely that regulators of starch metabolism or signals derived from starch act as integrators of plant metabolism and growth. It is intriguing that starch and protein are correlated and are predicted by the same set of metabolites. A strong correlation between starch turnover, protein content, and biomass is also found when Col0 is grown in different photoperiods (9). The conserved correlation might reflect the large energy costs associated with protein synthesis and maintenance (see ref. 9 and references therein). The TOR/RSK pathway is known to regulate ribosome numbers, protein level, and growth in response to the nutrient status in yeast and animals (28, 29), and evidence is emerging for an analogous role in plants (30). It will be interesting to investigate whether this signaling pathway contributes to the close link between starch, protein, and biomass.

Thus, multilevel metabolic and molecular phenotyping can be used to systematically identify metabolic traits and genes that correlate species-wide with growth and, combined with deep genotyping, to identify allelic variation that underlies these relationships. This work identifies candidate genes and polymorphisms that may be used directly or through the isolation of homologs to modulate biomass production in crops and provides precedence for an efficient strategy for future use to identify (crop-) species-specific lead genes.

Materials and Methods

Plant Material and Growth. A. thaliana accessions were obtained as in ref. 13 and grown in soil as in ref. 2. They were grown in at least 2 independent experiments. Each experiment contained 3 replicates of 5 pooled plants, with full randomization in growth cabinets to avoid microenvironmental effects. Material was harvested at the end of the light period. Samples typically contained 5 rosettes (~800 mg of FW). They were powdered in liquid N2, subaliquoted, and stored at −80 °C.

Metabolite Assays. Analysis of total amino acids, glucose, fructose, sucrose, starch, total protein, and chlorophyll was performed as in ref. 2 and GC-MS as in ref. 31, identifying metabolites by comparison with database entries of authentic standards (32).

Design and Validation of qRT-PCR Primers, RNA Preparation, and RT-PCR Assays. Primers were designed and synthesized at MWG Biotech AG using the PRIME program of GCG Wisconsin Package, version 10.2. Global alignments of suggested primer sequences with genomic and transcript sequences were performed using NCBI-BLASTN (33) to ensure unique oligonucleotide sequences. All primers were checked for nonspecific signals arising from primer dimers or template contamination by measuring a water control. Sequences of primers are in Table S3. RNA preparation, real time PCR, data analysis, and procedures for cDNA synthesis were as in ref. 34. Cut-off Ct values for all primers were set to 35 cycles. Ct values were normalized to 4 reference genes (34), At2g28390 (SAND family protein), At1g53090 (HECT-domain-containing protein), At5g08290 (yellow-leaf specific protein 8), and At1g25760 (ubiquitin-conjugating enzyme), by subtracting the average Ct value of the 4 reference genes from the Ct value of the gene of interest for each accession. Data were normalized based on the gene-wise average of all accessions including Col0, so that ΔΔCt represents ΔCtM minus ΔCtA to, ΔCtM values of technical and biological replicates within one accession were averaged, if they did not differ by more than 0.8 and 1.5, respectively. Genes were excluded if <11 accessions gave valid ΔΔCtM values. Of the 92 genes in set 1 and set 2, 42 and 52 were retained, respectively. Average transcript levels for all genes in the accessions are provided in Table S3.

Genotyping and Analysis of Population Structure. Selected accessions were genotyped with 460 SNP markers: 149 framework SNPs assembled in the frame of the A. thaliana “HapMap” project (J. Borevitz, personal communication; see http://naturalvariation.org/hapmap) and 311 SNPs with intermediate allele frequency selected by Wirthmann et al. (35) (Table S1). Genotyping was carried out using real-time PCR to amplify the genomic DNA, followed by sequencing the amplicons. SNP genotypes were determined by alignment of the amplicons using a model-based clustering method for inferring population structure that uses genotypic data from unlinked markers and accounts for the presence of LD by introducing population structure and attempting to find population groupings.
that are not in disequilibrium (14). An ancestry model allows population admixture. From the 460 SNP markers, 419 were used that had <25% missing data. Allele frequencies were assumed to be correlated (i.e., allele frequencies were likely to be similar due to shared ancestry or migration). The optimal number of subpopulations was simulated by setting K (number of subpopulation) from 1 to 15. The length of burn-in period as well as Markov Chain Monte Carlo iterations was simulated by setting 15. The length of burn-in period as well as Markov Chain Monte Carlo iterations improved steadily until 1.5. Usadel B, et al. (2008) Global transcript levels respond to small changes of the carbon status during a progressive exhaustion of carbohydrates in Arabidopsis rosettes. From the 460 SNP markers, 419 were used that had <25% missing data. Genomics sequencing was performed on both strands using LargeDye terminator chemistry on ABI 3730 sequencers (Applied Biosystems) by the Automatic DNA isolation and Sequencing unit at the Max Planck Institute for Plant Breeding. The sequences were assembled by using a Sequencer 4.8 (GeneCode) (see Table S6).

**Candidate Gene Association Testing:** Association mapping was performed using the sequence-verified SNPs that occurred with a minimum allele frequency of 0.05 in the entire accession panel. The GLM function of the TASSEL (25) program was applied with the STRUCTURE results used to control for population structure, with 10,000 permutations to determine significance.

**ACKNOWLEDGMENTS.** We thank Dr. Nengyi Zhang for his valuable advice about the TASSEL software. This work was supported by a Netherlands Genomics Initiative Genomics Fellowship 030-72-412, the Max Planck Society, the European Commission under the 6th Framework Programme (contract LSHT-G-2006-037704) and the Federal Ministry of Education and Research within the German Plant Genome Initiative (Genome Analysis of the Plant Biological System, grants 0311222B and 0310580E) and the GfM Forschungs-seinheiten zur Systembiologiemprogram.

Fig. S1. Population structure of the population used in the study. The program Structure 2.1 (1) was used to determine population structure and assign accessions to subpopulations, using 419 markers that were randomly distributed and had <25% missing data. The optimal number of subpopulation was simulated by setting K (number of subpopulation) from 1 to 15. The length of burn-in period as well as MCMC reps after burn-in were set to 100,000 for each run, and each run was iterated 10 times. An accession was assigned to the subpopulation or group to which it showed the highest probability of membership. When K was varied from 1 to 15 the posterior probability [Ln P(D)] improved steadily for K = 7–8 and reached plateau for K ≥ 8.

Fig. S2. Principle component analysis (PCA) of transcript levels at the end of the night for set 1 and set 2 of the C-responsive genes in 21 accessions. Set 1 contained 42 genes that show marked changes of transcripts at different times during the diurnal cycle or during the first hours of an extended night in wild-type plants. All these transcripts also respond to changes of the CO2 concentration in wild-type Col0 and show accentuated responses in the starchless pgm mutant (1, 2). The other set included 52 genes whose transcripts change log2 1.4 within 30 min of adding sucrose to carbon-starved seedlings (3). These presumably represent upstream components in the transcriptional response to sucrose. (A) PCA, using the 42 genes from for set 1. The PC1 and PC2 explained 41% and 11% of the variance, respectively. (B) PCA, using the 52 genes from for set 2. The PC1 and PC2 explained 34 and 14% of the variance, respectively. (C) Comparison of the separation of the 21 accessions in the PC1 of the PCA with genes of set 1 and set 2. (D) Relation between the loadings of the accessions in the PC1 and their rosette FW. Accessions were separated in 2 groups of slow (red circles) and fast (green circles) growing accessions from their behavior in an 8 h light/16 h dark regime. Col0 is depicted with a blue circle.

Fig. S3. Selection of the genes for sets 1 and 2. Selection criteria for genes in set 1 were based on the changes in transcript level in Col0 (WT) and pgm mutant in the normal diurnal cycle and in the extended night in WT. (A) Idealized representation of the behavior of the 5 classes of genes that were included in gene set 1. Five groups of sugar-responsive genes were designed very low range (a), low range (b), intermediate range (c), high range (d), and very high range (e) in respect to their response to changes in the C status of the rosettes. Genes that are induced under high sugar status show a curve shape that is labeled with (+), whereas genes that are repressed under high sugar conditions show a curve that is labeled with (−). A time axis is indicated beneath the figure, and periods of illumination and darkness are depicted as white and gray backgrounds, respectively. (B) Mathematical criteria used to fulfill the conditions described in A. All numerical values refer to a log2 scale used for display and comparison of ATH1 array data. Maximal (max) and minimal (min) correspond to the highest and lowest log2 values per gene, respectively. To generate the second and nonoverlapping set, the genes in set 1 were removed and a second set was then identified based on the response 30 min after adding sucrose to C-starved seedlings (1). The transcripts that increased or decreased most strongly in response to sucrose after 30 min were selected as set 2. They are presumably representatives of an early transcriptional response to sucrose. All genes selected for set 1 and 2 are listed in Table S3. ED, end of day; EN, end of night; ExN extended night.

Fig. S4. Comparison of gene correlation networks generated by measuring transcript levels at the end of the night in 21 accessions, with networks obtained in Col0 containing different levels of endogenous sugars (C-perturbation network), and 3 publicly available datasets in the database csbdb (1). (A) Scatter plot of the R values of all pair-wise gene–gene correlations obtained by comparing the accession gene network and the C-perturbation network. The accession gene network is computed from the data shown in Fig. 2 and Table S5. The C-perturbation network is calculated from published ATH1 experiments in which 5-week-old rosettes of Col0 were subjected to 23 treatments that alter endogenous C, harvest at different times during the diurnal cycle in wild-type Col0 and the starchless pgm mutant, and a short extension of the night and illumination in the presence of different CO₂ concentrations in Col0 (2, 3). Changes of C make a major contribution to the global changes of transcript levels in this data set (2, 3). R values were calculated for all pair-wise correlations of the 94 genes in each correlation network using R (4). Gene pair-wise correlations that were significant in both networks at \( P < 0.01 \) and \( R > 0.7 \) are depicted in blue and red circles when they have the same (Quadrant I and III) or opposite (Quadrant II and IV) signs, respectively. (B–E) Evaluation of the significance of the enrichment of pair-wise correlations in the accession network (Fig. 2) compared with correlation networks obtained from in house and publically available expression array datasets for Col0. The datasets used to generate a C-perturbation network are listed in the legend of Table S3. R and \( P \) values were calculated for each gene pair in each network (A) and conserved (QI and QIII) and reversed (QII and QIV) significant gene pair-wise correlations identified using a threshold of \( P < 0.01 \) and \( R > 0.7 \). The number of significant correlations was compared with the randomly created population of correlations generated by shuffling 100,000 times the data in the correlation matrix of the C-responsive genes in the 21 accessions to calculate the \( P \) values for the enrichment in these quadrants. A similar approach was taken to compare the accession network with for 3 further data series, publicly available on the csbdb database (5). (B) Accession network versus the carbon perturbation network. (C) Accession network versus a developmental series for wild-type Col0. (D) Accession network versus an abiotic stress series for Col0 aboveground organs. (E) Accession network versus an abiotic stress series for Col0 roots.

Other Supporting Information Files

Table S1
Table S2
Table S3
Table S4
Table S5
Table S6