



## Evaluation of light regulatory potential of Calvin cycle steps based on large-scale gene expression profiling data

Ning Sun<sup>1,\*</sup>, Ligeng Ma<sup>2,3</sup>, Deyun Pan<sup>1</sup>, Hongyu Zhao<sup>1,4</sup> and Xing Wang Deng<sup>2,3</sup>

<sup>1</sup>Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520, USA (\*author for correspondence; e-mail ning.sun@yale.edu); <sup>2</sup>Peking-Yale Joint Center of Plant Molecular Genetics and Agrobiotechnology, College of Life Sciences, Peking University, Beijing 100871, China; <sup>3</sup>Department of Molecular, Cellular, and Developmental Biology, Yale University, 165 Prospect Street, New Haven, CT 06520, USA; <sup>4</sup>Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520, USA

Received 3 September 2003; accepted in revised form 27 October 2003

**Key words:** Calvin cycle, gene expression, gene regulation, microarray, regulatory potentials

### Abstract

Although large-scale gene expression data have been studied from many perspectives, they have not been systematically integrated to infer the regulatory potentials of individual genes in specific pathways. Here we report the analysis of expression patterns of genes in the Calvin cycle from 95 *Arabidopsis* microarray experiments, which revealed a consistent gene regulation pattern in most experiments. This identified pattern, likely due to gene regulation by light rather than feedback regulations of the metabolite fluxes in the Calvin cycle, is remarkably consistent with the rate-limiting roles of the enzymes encoded by these genes reported from both experimental and modeling approaches. Therefore, the regulatory potential of the genes in a pathway may be inferred from their expression patterns. Furthermore, gene expression analysis in the context of a known pathway helps to categorize various biological perturbations that would not be recognized with the prevailing methods.

### Introduction

Advances in high throughput methodologies allow researchers to directly observe the gene expression response of an organism to perturbations at the genomic level. Although such large-scale data have been routinely collected, analytical approaches such as selecting differentially expressed genes (Kerr *et al.*, 2000; Efron *et al.*, 2001; Li and Wong, 2001; Dudoit *et al.*, 2002) or clustering genes based on expression profiles over different experimental conditions (Eisen *et al.*, 1998; Den-Dor and Yakhini, 1999; Hastie *et al.*, 2000; Lazzeroni and Owen, 2002) cannot relate gene expression data to the regulatory potential of any given gene in cellular metabolic pathways or signaling processes. In this study, we used the Calvin cycle, the primary pathway of carbon assimilation in C3 plants (Figure 1), as an example to investigate the connection between the gene expressions of pathway enzyme

genes and their regulatory potentials in the control of the pathway.

The Calvin cycle plays a fundamental role in most photosynthetic organisms. The cellular functions of its enzymes have been studied intensively through both experimental and mathematical modeling approaches since Calvin and colleagues identified this pathway in the 1950s (Pettersson and Ryde-Pettersson, 1988; Quick *et al.*, 1991; Stitt *et al.*, 1991; Fichtner *et al.*, 1993; Kossmann *et al.*, 1994; Price *et al.*, 1995; Paul *et al.*, 1995; Muschak *et al.*, 1997; Haak *et al.*, 1998; Harrison *et al.*, 1998; Fridlyand *et al.*, 1999; Miller *et al.*, 2000; Poolman *et al.*, 2000, 2001; Henkes *et al.*, 2001). However, it is still not fully understood how gene expression is coordinated and regulated to modulate the Calvin cycle. We used a large gene expression data set involving 95 experiments and carried out a systematic bioinformatics study to assess the gene expression pattern of the Calvin cycle enzyme

genes and to associate the observed regulation pattern with *cis*-element motifs and, more importantly, with the regulatory potentials of these pathway enzymes. In addition, we found that pathway-based analysis allows us to classify the biological perturbations that would not be possible with other methods.

## Materials and methods

### Gene annotation and biological pathways

We collected and processed the annotations for all cDNA probes from multiple public databases, TAIR (<http://www.arabidopsis.org/>), NCBI (<ftp://ftp.ncbi.nih.gov/>), the TIGR *Arabidopsis thaliana* database (<http://www.tigr.org/tdb/e2k1/ath1/>), Swiss-Prot (<http://us.expasy.org/sprot/>), and MIPS *Arabidopsis thaliana* database (MAtdB, <http://mips.gsf.de/proj/thal/>).

We summarized and processed the information on 140 plant biological pathways from Buchannan *et al.* (2000), ExPaSy (<http://us.expasy.org/tools/pathways/>; <http://us.expasy.org/enzyme/>), TAIR (<ftp://tairpub.tairpub@ftp.arabidopsis.org/home/tair/Pathways/>), and KEGG (<http://www.genome.ad.jp/>). We associated the enzyme names with their enzyme commission numbers (EC numbers), and connected the EC numbers to gene locus IDs and the cDNA GenBank accession numbers. We then used the geometric mean of the gene expression data of the cDNA probes corresponding to the same gene to represent the expression value of this gene. Similarly, we obtained the gene expression value for an enzyme from the geometric mean of the expression values of its enzyme-encoding genes.

### Regulated pathway selection

We applied 2-fold change cutoff to define whether an enzyme gene's expression was altered by the given biological perturbations. For each pathway, we counted the number of genes with changed expressions, the number of genes in the pathway, and the total number of genes in the microarray. Then we performed the Fisher's exact test for the null hypothesis that a particular pathway is not regulated under a biological perturbation. We then used 0.001 as the cutoff for the resulting *p* value to identify the regulated pathways for each given biological perturbation. The pathways that are significantly regulated in most of the 95 experiments are good candidates to study gene regulation patterns under these perturbations.

### Analysis on average rank of fold change

We calculated the average rank based on the fold change and the standard error of the rank over 95 biological perturbations for each enzyme. The null hypothesis is that the gene regulations of the nine enzymes are equally sensitive to the given biological perturbations. Under the null hypothesis, the average rank value equals to the mean of random shuffles of 1 to 9 for 95 times, which is 5. We then computed the *z* statistics based on the observed rank values. The average ranks and their standard errors represent the sensitivities of the gene regulations of the nine enzymes to the biological perturbations, whereas the *p* values show the statistical evidence for the enzymes' sensitivity.

### Projection of experiments (biological perturbations) on a two-dimensional space

Let *N* denote the total number of genes used to measure similarity between two experiments,  $\mathbf{V}_i = (v_{ik})$  is the vector of gene expression for the *i*th experiment, where  $i = 1, \dots, 95$  and  $k = 1, \dots, N$ . The dissimilarity between experiments *i* and *j* is defined as either through Euclidean distance,

$$D_{ij}^E = \left( (\mathbf{V}_i - \mathbf{V}_j)' \cdot (\mathbf{V}_i - \mathbf{V}_j) \right)^{1/2} \\ = \sqrt{\sum_k (v_{ik} - v_{jk})^2},$$

or through 1-Pearson correlation coefficient,

$$D_{ij}^P = 1 - \frac{\sum_k v_{ik} v_{jk} - \frac{\sum_k v_{ik} \sum_k v_{jk}}{N}}{\sqrt{\left( \sum_k v_{ik}^2 - \frac{(\sum_k v_{ik})^2}{N} \right) \left( \sum_k v_{jk}^2 - \frac{(\sum_k v_{jk})^2}{N} \right)}}.$$

Because  $D_{ij}^P$  is a distance between two normalized vectors, it mainly measures dissimilarity of expression patterns between two experiments. Therefore, we call it the 'pattern distance', and call  $D_{ij}^E$  the 'strength distance' in the following.

For the 95 experiments, we calculated all pairwise pattern distances and formed a symmetric  $95 \times$

95 matrix, denoted as PDM (pattern distance matrix). Similarly we obtained a  $95 \times 95$  SDM (strength distance matrix).

To visualize the information embedded in PDM and SDM, we used the 'cmdscale' function in S-Plus to perform metric multidimensional scaling to reduce the  $95 \times 95$  matrix (PDM or SDM) to a one-dimensional vector ( $\mathbf{P}$  or  $\mathbf{S}$  respectively) in a way that the distances between any two experiments represented in this one-dimensional vector  $r$  ( $d_i$  in vector  $\mathbf{P}$  or  $\mathbf{S}$ ) are as close as possible to the distances ( $D_{ij}$ , and  $j = 1, \dots, 95$ , in PDM or SDM respectively) in the original distance matrix.

We then used the elements in vector  $\mathbf{S}$  (95 values) as the x coordinates of the 95 experiments to represent the one-dimensional strength distances. Similarly we used the elements in vector  $\mathbf{P}$  (95 values) as the y coordinates of these experiments to represent the one-dimensional pattern distances. Therefore, each of the 95 experiments corresponds to one point in the figure.

#### *Analysis of cis-element motifs*

We searched the light-regulated *cis* elements from a plant motif database by PLACE (<http://www.dna.affrc.go.jp/htdocs/PLACE/>). We performed the motif pattern search with the upstream sequences (based on the genome sequences and genome annotations from the MIPs ftp site: <ftp://ftpmips.gsf.de/cress/>) and the program AlignACE (Roth *et al.*, 1998; Hughes *et al.*, 2000; <http://atlas.med.harvard.edu/>). Then we mapped the three light-regulated *cis* elements and one identified novel motif pattern back to the upstream sequences of 20,536 genes with locus IDs (MIPs: <ftp://ftpmips.gsf.de/cress/>), the 6133 unique genes in the microarray, and the 12 genes encoding the nine Calvin cycle enzymes. We did binomial tests based on the proportions of the existence of a motif pattern in three groups of genes to evaluate the specificity of the motif to the Calvin cycle.

## Results

#### *Calvin cycle and microarray data*

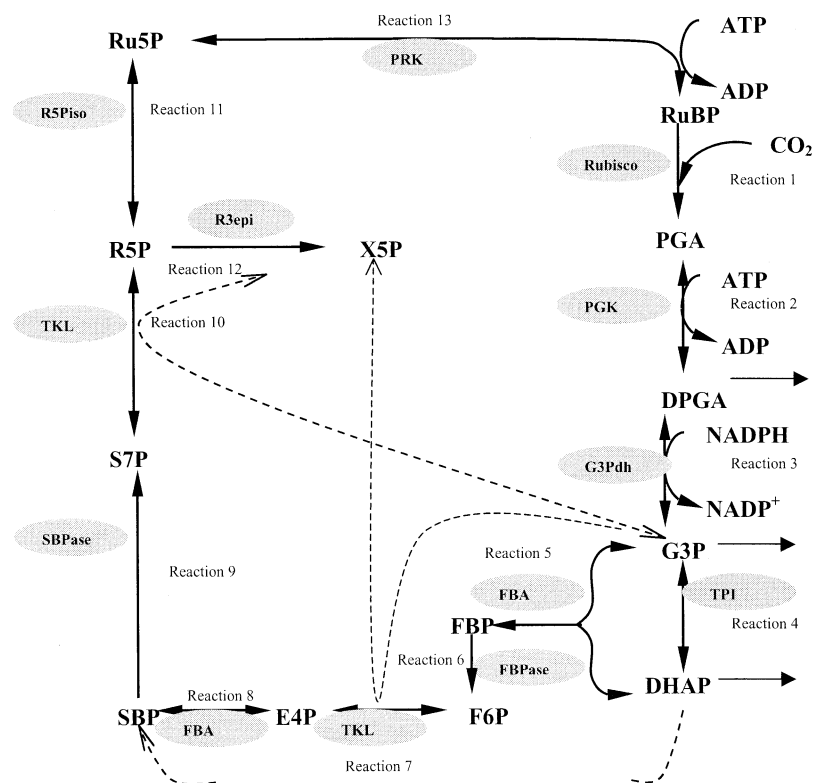
We examined 410 microarray data sets representing 95 different experiments (4 or more replicates per experiment) on light regulation in *Arabidopsis* (Supplemental Table 1). There were 8828 cDNA probes on our *Arabidopsis* microarray used to collect this data set. These probes were annotated with

gene locus IDs. We also built 140 known plant biological pathways into our *Arabidopsis* database PathMAPA (<http://zhao.med.yale.edu/pathmapa.htm>). In this database, all the enzymes were associated with their enzyme commission numbers, the enzyme proteins, and the gene locus IDs of these enzyme proteins. This database allowed us to use gene locus IDs to join gene expression data with these biological pathways.

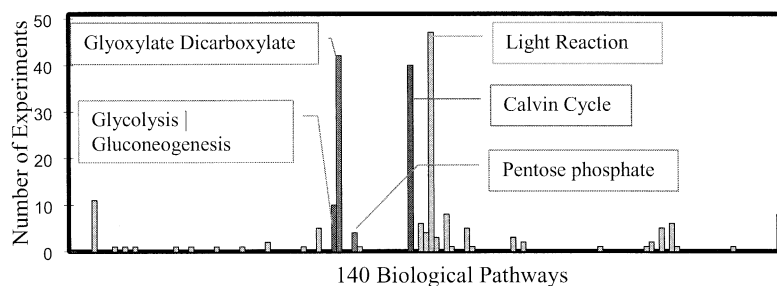
An initial evaluation of the effects of the 95 experiments, representing distinct biological perturbations, on the regulations of the 140 known pathways with the microarray expression data set indicated that the expression of Calvin cycle genes was one of the most sensitively regulated among all the pathways (Figure 2). Further, there is substantial literature on the kinetics of the Calvin cycle reactions involving both experimental and modeling approaches, which can be related to the observations from the microarray experiments. Therefore, we selected the Calvin cycle to explore the relationship between the expression of the enzyme genes and their regulatory potentials in a given pathway.

#### *Average expressions of only the confirmed Calvin cycle enzyme gene family members are utilized*

There are 11 enzymes that catalyze the 13 biological reactions of the Calvin cycle (reactions 5 and 8 share one enzyme, and reactions 7 and 10 share another enzyme, see Figure 1). Although we were able to find 269 cDNA probes representing the Calvin cycle enzyme genes on our microarray, we only selected the genes (or locus IDs) with multiple cDNA representatives on the microarray that have similar gene expression patterns across 95 biological perturbations. Using this criterion, we identified 86 cDNA probes representing 12 Calvin cycle genes, of which four encode the small subunits of Rubisco. The average expression of these four genes was used to represent the Rubisco expression level. The other eight genes encode eight enzymes: glyceraldehyde-3-phosphate dehydrogenase (G3Pdh), fructose biphosphate aldolase (FBA), fructose biphosphatase (FBPase), transketolase (TKL), sedoheptulose-1,7-biphosphatase (SBPase), ribose-5-phosphate isomerase (R5Piso), ribulose-p-3-epimerase (R3epi), and phosphoribulokinase (PRK) of the Calvin cycle (Supplemental Table 2). Phosphoglycerate kinase (PGK) and triose phosphate isomerase (TPI) were not included in our analysis due to the lack of representative cDNA probes. The re-sequencing of these 86 cDNA



*Figure 1.* The Calvin cycle has 13 reactions starting at Ribulose and ending with phosphoribulokinase. There are 11 enzymes to catalyze the 13 reactions, where reactions 5 and 8 share the same enzyme (FBA), and reactions 7 and 10 share the other same enzyme (TKL). In our analysis, we studied 9 enzymes because there were not consistent cDNAs representing PGK and TPI on our microarray. The abbreviations are given in the text.



*Figure 2.* The regulated pathways assessed in this work. We assessed whether each of the 140 biological pathways was regulated under each biological perturbation through Fisher's exact test as explained in the text. We used 0.001 as the cutoff for  $p$  values to define whether a specific pathway was regulated under the given biological perturbation. The horizontal axis represents 140 biological pathways in our database, and the vertical axis is the number of experiments in which a specific pathway was significantly regulated. The Calvin cycle was significantly regulated in many perturbations, whereas the pathways sharing some enzymes with the Calvin cycle was regulated in many fewer experiments. For example, glycolysis/gluconeogenesis and pentose phosphate pathways were regulated in a limited number of experiments (shown in the figure), and pentose & glucuronate interconversions and fructose & mannose metabolism were not regulated in any experiment.

Table 1. Statistical assessment of the specificity of three light-related binding *cis* elements and one novel motif pattern to the Calvin cycle genes.

	Number of genes	GATA or I-box (GATAA)	G-box (SACRTGG)	GT1 (GGTTAA)	Novel (WKNGTGWGG)
Predicted genes	20563	17892	4247	7562	2970
Proportion		0.87	0.21	0.37	0.14
Microarray genes	6133	5384	1424	2249	901
Proportion		0.88	0.23	0.37	0.15
Calvin cycle genes	12	12	10	8	8
Proportion		1	0.83	0.67	0.67
<i>p</i> value		0.38	0	0.067	0.0001

S: C or G; R: G or A; W: A or T; K: G or T; N: A, T, C, or G.

probes confirmed their expected identities. We assessed the potential for cross-hybridization based on the cDNA sequences and found that the cDNA sequences encoding the same enzyme (e.g. Rubisco small subunit) highly likely to cross hybridize with each other, whereas the cDNA sequences from the genes encoding the proteins for different enzymes should not cross hybridize at all. Therefore, gene expression analysis based on the selected cDNA probes should provide a reliable assessment of expression patterns of all genes encoding for each individual Calvin cycle enzymatic activity, but not differentiate specific enzyme isoforms.

It should be pointed out that in the Calvin cycle, there are only three enzymes (Rubisco, SBPase, and PRK) that are specific to the Calvin cycle, while other enzymes are involved in additional pathways within chloroplasts. For example, FBA, R5Piso, R3epi, and TKL are involved in the pentose phosphate pathway. Thus it is of great interest to investigate how this specific pathway among the interconnected pathways is regulated in plants.

*There is a strong correlation between the gene regulation pattern and the regulatory potentials of the enzymes in the Calvin cycle*

Each microarray experiment yielded estimates of relative gene expression levels between two biological samples (biologically perturbed sample versus reference sample) for every cDNA probe on the microarray. For these nine enzymes, we calculated the estimated expression ratios and changed the ratios to the folds. The fold value equals to the ratio if the ratio is larger than 1 otherwise equals to 1/ratio. We then ranked the genes with an ascending order of the fold changes.

A higher rank corresponds to a higher fold change. Figure 3 summarizes the rank distribution of each enzyme's gene expression within the Calvin cycle and across all 95 experiments. This figure shows that the regulation of these nine enzymes' gene expressions exhibits differential sensitivity under the 95 biological perturbations. The regulations of FBA, SBPase, and Rubisco were most sensitive to biological perturbations, R5Piso was least sensitive, and G3Pdh, FBPase, TKL, R3epi, and RPK had intermediate sensitivity.

This gene regulation pattern, which represents the relative responsiveness of the Calvin cycle's enzyme gene expressions to biological perturbations, is remarkably consistent with the rate-limiting roles of these enzymes discovered through both mathematical modeling and experimental studies. The three enzymes most sensitive to biological perturbations, SBPase, Rubisco, and FBA, were found to have important rate-limiting or near-rate-limiting roles in controlling the carbon assimilation flux of the Calvin cycle (Quick *et al.*, 1991; Stitt *et al.*, 1991; Fichtner *et al.*, 1993; Haak *et al.*, 1998; Harrison *et al.*, 1998; Fridlyand *et al.*, 1999; Miller *et al.*, 2000; Poolman *et al.*, 2001). R5Piso was least sensitive to perturbations, and was found to have no rate limiting roles in the Calvin cycle in a previous study (Fridlyand *et al.*, 1999). G3Pdh, FBPase, and RPK, which had intermediate sensitivities, played intermediate rate-limiting roles (Kossmann *et al.*, 1994; Price *et al.*, 1995; Paul *et al.*, 1995; Muschak *et al.*, 1997; Fridlyand *et al.*, 1999). While both R3epi and TKL had intermediate sensitivity, R3epi was thought to have no rate-limiting role (Fridlyand *et al.*, 1999) whereas TKL was thought probably to have a near rate-limiting role in the Calvin cycle (Henkes *et al.*, 2001). Both the agreement and

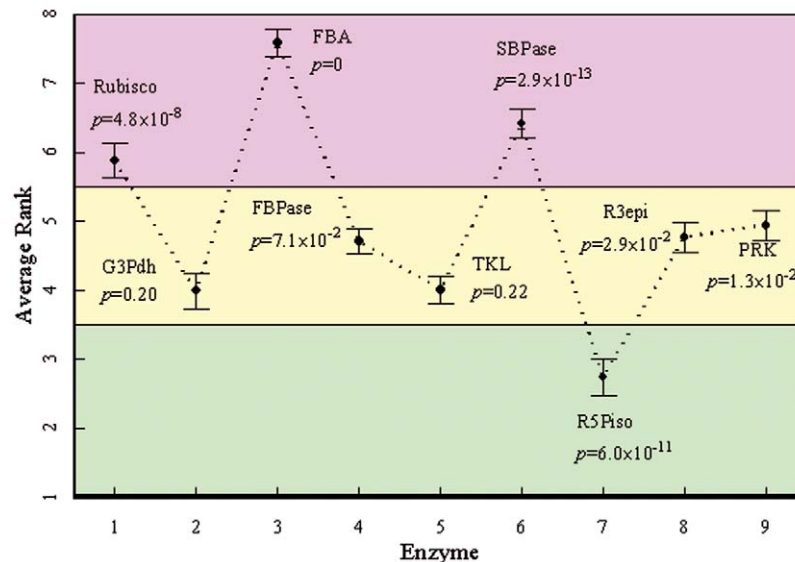


Figure 3. The average rank of the nine Calvin cycle enzyme genes over 95 experiments. The average rank for a gene reflects the relative sensitivity of this gene to biological perturbations in these 95 experiments. The  $p$  value for each gene summarizes the statistical evidence of this gene regulation's sensitivity. The regulations of these 9 enzyme genes are not equally sensitive to biological perturbations, with Rubisco, FBA, and SBPase most sensitive to biological perturbations (upper region), G3Pdh, FBPase, TKL, R3epi, and PRK having moderate sensitivity (middle region), and R5Piso showing the least sensitivity (lower region).

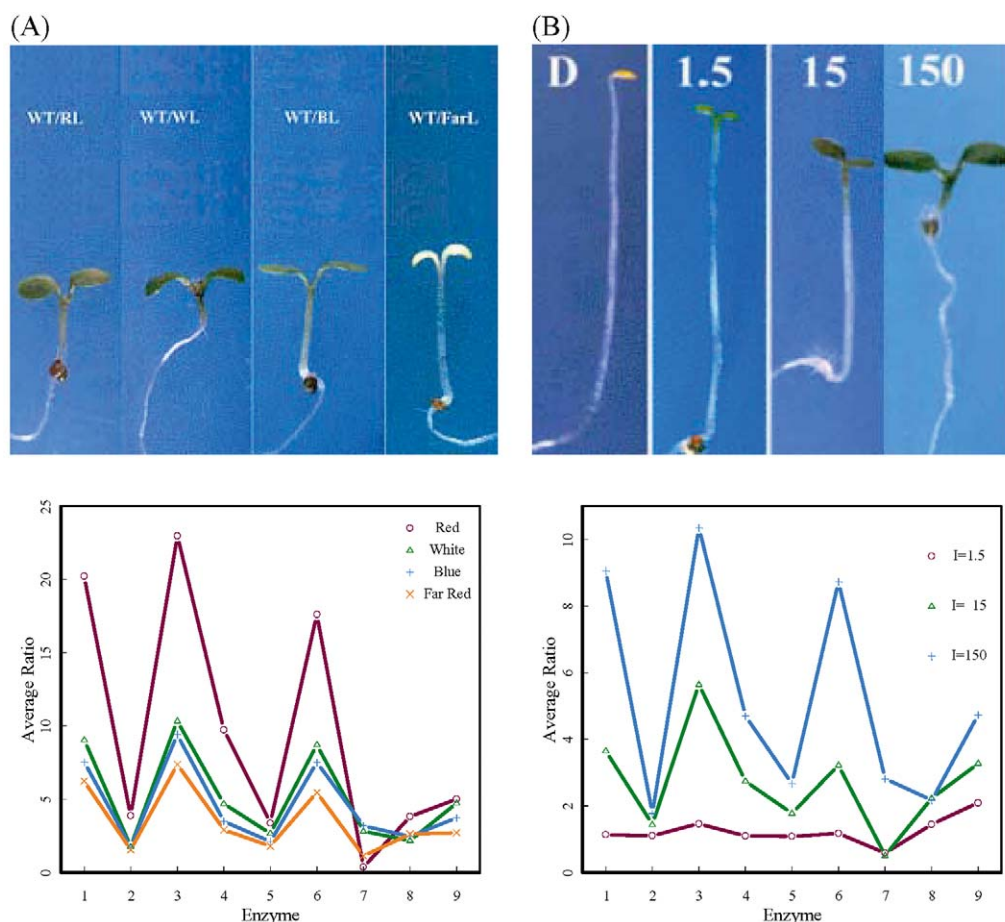
the discrepancy between our observations and the previous reports were also examined through the regulatory motif search, which is described later in this paper.

#### *Light signaling processes likely confer the identified gene regulation pattern*

Among the 95 different biological perturbations, most (ca. 70%) were alterations in light signals or mutations in possible light-signaling components, while the rest were performed as various controls. We found that it was this majority of perturbations that caused the identified gene regulation patterns. For example, the identified gene regulation pattern was preserved over white light with difference light intensities (1.5, 15, 150  $\mu$ E) (Figure 4B), various types of lights (white, red, blue, and far-red) (Figure 4A), different degrees of activation of the *cop1-6* mutants under dark (1, 2, 3, 4, 4.5, 5, 5.5, and 6 days at 22 °C to be active and 5, 4, 3, 2, 1.5, 1, 0.5, and 0 days at 30 °C to be inactive), different mutants (e.g. *cop1-1*, *cop4-1*, *cop9-1*, and *cop10-1*), and various ecotypes (Supplemental Figure 1). The experiments that did not follow the identified gene regulation pattern included those involving lethal *cop* mutants (*cop1-5*, *cop1-8*, and *det1-6* under darkness), strong light receptor mutants (*cry1cry2* double mutant under blue,

and *phyA* mutant under far red), and mutants like *det2* and *coil* under dark. In these experiments, the Calvin cycle genes generally showed no change in expression level, and the plants exhibited either similar phenotypes to the dark-grown wild-type seedlings or severe growth retardation (Supplemental Table 3).

We also observed that this identified gene regulation pattern was present under far-red light (unusable light for producing the energy source for the Calvin cycle) and in *cop1-6* mutants under dark (only to mimic the gene regulation role of light without any light to drive the Calvin cycle). This suggests that light signaling is responsible for the identified gene regulation pattern, because there is no feedback from the Calvin cycle fluxes under the above experimental conditions. Even for experiments with the identified gene regulation pattern, the strength of this pattern varied across different experiments in which the phenotypes of the plants showed variation as well. Figure 4A shows that the strength of this identified pattern decreased from red, white, blue to far-red light we used for the wild type and that the phenotypes showed similar tendency of severities. This order of light conditions is different from the one based on the expected Calvin cycle flux, white (150  $\mu$ E), red (108.5  $\mu$ E), and blue (16.2  $\mu$ E), because the Calvin cycle in young leaves usually undergoes a slow steady state and the

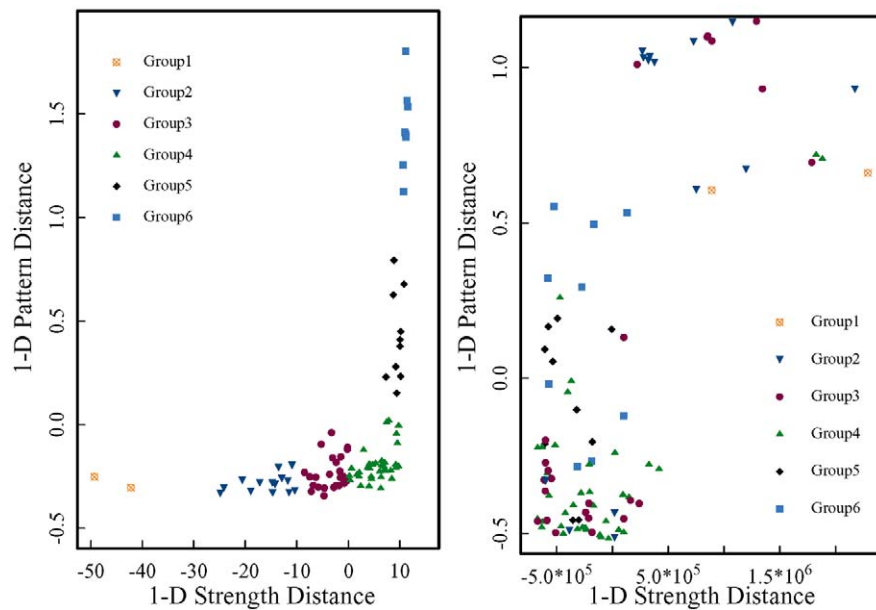


**Figure 4.** The phenotypes and their corresponding gene regulation patterns among the 9 Calvin cycle enzyme genes (the number designation for each genes is the same as Figure 3). A. The wild types under red (108.5  $\mu$ E), white (150  $\mu$ E), blue (16.2  $\mu$ E), and far-red (160.8  $\mu$ E). B. The wild types under white light with three different intensities (0, 1.5, 15, 150  $\mu$ E). Although the gene regulation patterns are similar, the strength of the pattern varied among different experiments or perturbations, which in general correlated with the degree of photomorphogenic phenotype (hypocotyl shortening and cotyledon enlargement) of seedlings used in these experiments except the white-light-grown seedling.

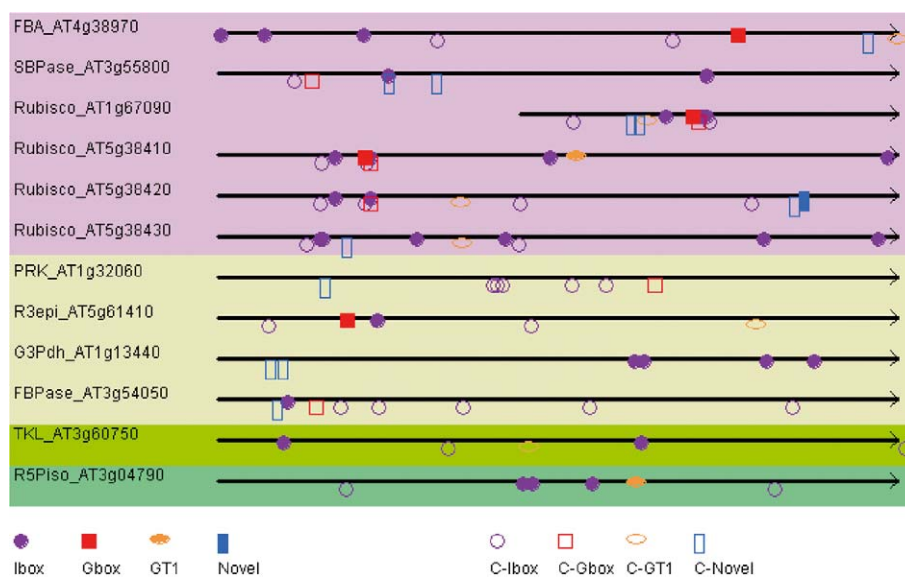
carbon assimilation flux in the Calvin cycle is proportional to light intensities (Poolman *et al.*, 2001). This divergence between the order of light conditions according to the actual gene regulation patterns and the one based on the expected Calvin cycle fluxes was also observed for the *cop* mutants under varying light conditions (Supplemental Figure 1). This further suggests that light signaling contributes more than the feedback of the Calvin cycle fluxes to the strength of the identified gene regulation pattern. Hence, the identified gene regulation pattern likely directly relates to the light signaling process.

*The strength of identified gene regulation pattern correlates with the phenotype consequences of the biological perturbations*

The above results suggest that the identified gene regulation pattern may be an important property of the regulation of the Calvin cycle by light-signaling processes, and the strength of the identified gene regulation pattern may be useful to more sensitively classify various biological perturbations in these processes. Because both the pattern of gene regulations and the magnitude of this pattern provide useful information on the effects of biological perturbations on a given pathway, both types of information can be incorporated to define similarity or distinction among perturbations or experiments. Note that the most com-



*Figure 5.* The projection of 95 experiments onto a 2-dimensional space based on gene expression data of (A, left) the genes for 9 Calvin cycle enzymes and (B, right) all genes on the microarray. The experiments are divided into six groups in A, in which the 2-D space was based on the Calvin cycle gene expression data. These groups are labeled with different symbols and colors. The same symbol and color is used for each experiment when it was projected onto the 2-D space in B, which was generated with gene expression data from all the genes on the microarray. The comparison between these two figures suggests that the information based on the Calvin cycle genes is qualitatively different from that based on all genes to categorize biological perturbations or experiments.



*Figure 6.* The locations of four motifs in the upstream sequences of 12 Calvin cycle genes. The solid black lines represent the available upstream sequences up to 1 kb, with the right end of each line indicating the starting site of transcription and the arrow indicating the direction of transcription. The motif patterns are searched along both strands of the upstream sequences. The motifs on the gene-coding strand are denoted with filled symbols, and those on the other strand are denoted with unfilled symbols and their motif names are labeled with the prefix C. These 12 genes are arranged by the sensitivities of their regulations with the pink region having the most sensitivity, the yellow region and light green regions having intermediate sensitivity, and the dark green region having the least sensitivity. The Calvin cycle genes in the pink and yellow regions have either G-box or the novel motif in their upstream regions, AT3g04790 (R5iso) in the dark green region have neither motifs, and AT3g60750 (TKL) in the light green region, with intermediate sensitivity, does not have G-box nor the novel motif. In contrast, I box (or GATA motif) is present in almost all promoters while GT1 box is distributed among half of those promoters without preference.

mon approach to clustering gene expression data is to use the Pearson correlation coefficient to measure similarity. This measurement mostly captures the pattern but not its magnitude. To visualize experiment classifications based on both types of information, we projected 95 experiments onto a 2-dimensional space, with the X-axis representing the strength distance of the identified pattern and the Y-axis representing the pattern distance. Figure 5A shows that the information carried by the two measurements is different. Because the majority of experiments shared a similar gene regulation pattern in the Calvin cycle but had different phenotypes, the incorporation of the pattern strength information can better correlate these experiments with the observed phenotypes (Supplemental Table 3). Therefore, the integration of gene regulation pattern and pattern strength information is important to evaluate the effects of biological perturbations on the Calvin cycle.

#### *Clustering biological perturbations in the context of the Calvin cycle provides additional insights*

The relationships among the perturbations in our focused analysis of the Calvin cycle genes appear to differ qualitatively from that revealed by the analysis with all genes on the microarray. Based on Figure 5A, we separated the perturbations (experiments) into six groups (labeled with different color symbols in Figure 5). The plants used in each of these six experimental groups in general had similar phenotype consequences (Supplemental Table 3). The six groups identified under the context of the Calvin cycle scattered widely when they were projected onto the 2-dimensional space generated by using all gene expression data of the microarray (Figure 5B). Therefore, the clustering method under the context of a pathway can detect the effect of biological perturbations more sensitively due to the connection between pathway gene regulations and their rate limiting functions and furthermore the phenotypic consequences of the biological perturbations. Although the classification of all perturbations into the six groups based on the pathway is not perfect, it is much more biologically meaningful than that based on the entire gene set on the microarray.

#### *The genes with the identified gene regulation pattern share specific cis-elements*

To correlate gene expression patterns with possible regulatory *cis* elements, we studied three light-related

*cis* elements (G-box, GATA or I box, and GT1; see Donald and Cashmore, 1990; Puente *et al.*, 1996) and a novel motif [WKNGTGWGG] identified through AlignACE (Roth *et al.*, 1998; Hughes *et al.*, 2001) by searching the 1kb upstream sequences of the six regulatory Calvin cycle genes (Rubisco, FBA, SBPase, G3Pdh, FBPase, RPK). The frequencies of these patterns in the upstream regions of the Calvin cycle genes were compared to their frequencies in the upstream regions of 20 563 predicted *Arabidopsis* genes (from the MIPS database, <http://mips.gsf.de/>) and 6133 unique genes on our microarray (Table 1). We found that the GATA box (or I box) and GT1 core are not statistically significant to be specific to the regulated genes, due to the fact that the GATA box and GT1 core are commonly present in promoter regions. However, the G box and the novel motif are statistically significant to be specific to the promoters of those regulated genes in the Calvin cycle. The presence of either or both the G box and the novel motif in the upstream sequences of the Calvin cycle-regulated genes (Figure 6) is consistent with the observed gene regulation pattern. The most sensitively regulated genes tend to have a combination of the G box and the novel motif or multiple copies of one or both motifs, whereas the intermediate responsive genes tend to have fewer instances of the G box or the novel motif. For example, R3epi has one G box and intermediate sensitivity toward perturbations, while TKL has neither the G box nor the novel motif, consistent with its relatively weak sensitivity of gene regulation compared to other genes with intermediate regulation sensitivity. The gene whose regulation was least sensitive to perturbations (R5iso) has neither the G box nor the novel motif.

## Discussion

### *The gene regulation sensitivities of the Calvin cycle enzymes support their regulatory roles in this pathway*

Systematic examination of the average rank of fold changes for the nine Calvin cycle enzyme genes from confirmed representative cDNA probes in the array revealed a clear gene regulation pattern among the Calvin cycle enzyme genes (Figure 3). Due to the fact that most of the biological perturbations are related to light regulation, the Calvin cycle is significantly regulated among the experiments. The biological perturbations seem to cause more sensitive responses to some Calvin cycle enzyme gene expression (Rubisco,

FBA, SBPase) while result in less sensitive response for other enzyme genes (e.g. R5Piso). Consistent with that the genes encoding the enzymes of key regulatory steps in most cellular pathways are highly regulated by various signals (Panda *et al.*, 2002), the enzymes with high gene regulation sensitivities in this study may also indicate their roles in controlling the activities of the Calvin cycle. The enzymes whose transcriptional regulations are less responsive to the given biological perturbations remain a similar level of enzyme gene expressions as the ones in the wild-type seedlings that grew 6 days in darkness. The absolute abundance of these enzymes is not necessarily low because it is possible that they already had high abundance in the reference sample (6-day old dark-grown wild type) due to the need of their activities by other pathways or processes. In response to the light signal, those enzyme activities are sufficient and can be readily utilized by the Calvin cycle. Consequently these enzymes may play a weak role in controlling the Calvin cycle. When we compared the identified gene regulation pattern of the Calvin cycle enzyme genes with the known regulatory potentials of these enzymes in the Calvin cycle from previous experimental and modeling studies, there was a good correlation between the observed gene regulation pattern and the reported regulatory potential. This good correlation supports the above hypothesis that the enzymes with the regulatory roles in the Calvin cycle are the ones whose gene expression levels are most sensitive to the various biological perturbations. For this reason, the high sensitivity of the transcriptional regulation of the genes to the biological perturbations could be used as one of the indicators for the genes encoding the rate-limiting enzymes of the pathway.

*Gene regulation pattern in the Calvin cycle reveals a new cis-element pattern*

It is generally believed that G box, GATA box, and GT1 are among the light-regulated *cis* elements (Donald and Cashmore, 1990; Puente *et al.*, 1996). In our analysis, we have shown that GATA and GT1 are present in the upstream sequences of a large number of genes and that they are not specific to the defined highly regulated genes in the Calvin cycle. On the contrary, the G-box and a novel motif are significantly specific to the regulated genes in the Calvin cycle, and their presence or absence pattern on the upstream sequences of the 12 Calvin cycle genes is consistent with the identified gene regulation pattern. Therefore, the

G box and the novel motif may play an important role in conferring the defined regulatory pattern of those regulated genes in the Calvin cycle.

There were two discrepancies between gene regulation patterns derived from the 95 biological perturbations and previous results. Both R3epi and TKL showed intermediate gene regulation sensitivities in our experiments but were reported to have either no rate-limiting role (Fridlyand *et al.*, 1999) or a near rate-limiting role (Henkes *et al.*, 2001) respectively. Based on both microarray data analysis and the distribution of the promoter elements upstream of those two genes, it indicates that the enzyme R3epi is regulated more sensitively by the given set of biological perturbations than the enzyme TKL. This suggests that R3epi may play a more important role than TKL in controlling the activity of the Calvin cycle through transcriptional regulations under the given conditions. However, this transcriptional control represents only one aspect of the regulatory control of the Calvin cycle, further experimental studies may lead to better understanding of these two discrepancies between our results and those of previous studies.

*A valuable approach: clustering in the context of pathway(s)*

In our analysis, we also found that the application of the identified gene regulation pattern to cluster the biological perturbations in light signaling processes is a valuable approach. We demonstrated that the identified gene regulation pattern and the strength of the pattern carry different information on the biological perturbations (Figure 5). The clusters based on both pattern and strength dissected the role of each biological perturbation in regulating the Calvin cycle better than the existing clustering methods, which overlook this role by using the data of all genes (Figure 5 and Supplementary Table 3). Therefore, clustering under the context of a specific pathway reveals new information embedded in the microarray data. Because the same biological perturbations may have different roles in regulating different pathways, the experiment/perturbation clusters obtained under the context of different pathways may vary. Utilization of clustering information in the context of various pathways is a challenging but promising approach to evaluate the effects and relationships of biological perturbations.

## Conclusion

The results based on our systematic study of the Calvin cycle enzymes are consistent with the notion that genes for enzymes of key regulatory steps in cellular pathways are highly regulated by various signals (Panda *et al.*, 2002). Therefore, a good correlation between enzymes' rate-limiting roles in a pathway and their genes' relative responsiveness to perturbations in regulatory signaling processes may be an important feature of cellular pathways. It suggests that similar analyses on other pathways may lead to insights into the regulatory potentials of their enzymes, when the selected pathways are regulated by the given biological perturbations. Such insights may allow researchers to design experiments to identify or modify the most relevant rate-limiting genes more efficiently. In addition, the exploration of the gene regulation pattern of one or more cellular pathways may lead to a better understanding of various biological perturbations.

## Acknowledgements

We thank Mr Matthew Holford for assisting on the pathway analysis. Our research was supported by grants from American Cancer Society (IRG 58-012-45 to N.S.), National Science Foundation (DMS0241160 to H.Y.Z.), National Institutes of Health (GM-47850 to X.W.D., and GM59507 to H.Y.Z.), and a Strategic International Corporation project (30221120261) from the National Science Foundation of China. L.M. is a long-term postdoctoral fellow of the Human Frontier Science Program.

## References

- Buchanan, B.B., Gruissem, W. and Jones, R.L. 2000. Photosynthesis: Biochemistry & Molecular Biology of Plants. American Society of Plant Physiology, Rockville, MD.
- Den-Dor, A. and Yakhini, Z. 1999. Clustering gene expression patterns. In: S. Istrail, P. Pevzner and M.S. Waterman (Eds.) *Recomb 99*, ACM Press, Washington, DC, p. 188.
- Donald, R.G. and Cashmore, A.R. 1990. Mutation of either G box or I box sequences profoundly affects expression from the *Arabidopsis rbcS-1A* promoter. *EMBO J.* 9: 1717-1726.
- Dudoit, S., Yang, Y.H., Speed, T.P. and Callow, M.J. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.* 12: 111.
- Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. 2001. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Ass.* 96: 1151.
- Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863.
- Fichtner, K., Quick, W.P., Schulze, E-D, Mooney, H.A., Rodermel, S.R., Bogorad, L. and Stitt, M. 1993. Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with 'antisense' *rbcS*. V. Relationship between photosynthetic rate, storage strategy, biomass allocation and vegetative plant growth at three different nitrogen supplies. *Planta* 190: 1.
- Fridlyand, L.E., Backhausen, J.E. and Scheibe, R. 1999. Homeostatic regulation upon changes of enzyme activities in the Calvin cycle as an example for general mechanisms of flux control. What can we expect from transgenic plants? *Photosynth. Res.* 61: 227.
- Haake, V., Zrenner, R., Sonnewald, U. and Stitt, M. 1998. A moderate decrease of plastid aldolase activity inhibits photosynthesis, alters the levels of sugars and starch and inhibits growth of potato plants. *Plant J.* 14: 147.
- Harrison, E.P., Willingham, N.M., Lloyd, J.C. and Raines, C.A. 1998. Reduced sedoheptulose-1,7-bisphosphatase levels in transgenic tobacco lead to decreased photosynthetic capacity and altered carbohydrate accumulation. *Planta* 204: 27.
- Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D. and Brown, P. 2000. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 1, research0003.1-0003.21.
- Henkes, S., Sonnewald, U., Flachmann, R., Badur, R. and Stitt, M. 2001. A small decrease of plastid transketolase activity in antisense tobacco transformants has dramatic effects on photosynthesis and phenylpropanoid metabolism. *Plant Cell* 13: 535.
- Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. 2000. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296: 1205.
- Kerr, M.K., Martin, M. and Churchill, G. 2000. Analysis of variance for gene expression microarray data. *J. Comp. Biol.* 7: 819.
- Kossmann, J., Sonnewald, U. and Willmitzer, L. 1994. Reduction of the chloroplastic fructose-1,6-bisphosphatase in transgenic potato plants impairs photosynthesis and plant growth. *Plant J.* 6: 637.
- Lazzeroni, L. and Owen, A. 2002. Plaid models for gene expression data. *Stat. Sin.* 12: 61.
- Li, C. and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl. Acad. Sci. USA* 98: 31.
- Ma, L., Li, J., Qu, L., Hager, J., Chen, Z., Zhao, H. and Deng, X.W. 2001. Light control of *Arabidopsis* development entails coordinated regulation of genome expression and cellular pathways. *Plant Cell* 13: 2589-2607.
- Ma, L., Gao, Y., Qu, L., Chen, Z., Li, J., Zhao, H. and Deng, X.W. 2002. Genomic evidence for COP1 as a repressor of light-regulated gene expression and development in *Arabidopsis*. *Plant Cell* 14: 2383-2398.
- Ma, L., Zhao, H. and Deng, X.W. 2003. Analysis of the mutational effects of the *COP/DET/FUS* loci on genome expression profiles reveals their overlapping yet not identical roles in regulating *Arabidopsis* seedling development. *Development* 130: 969-981.
- Miller, A., Schlagenhafer, C., Spalding, M. and Rodermel, S.R. 2000. Carbohydrate regulation of leaf development: prolongation of leaf senescence in Rubisco antisense mutants of tobacco. *Photosynth. Res.* 63: 1.

- Muschak, M., Hoffmann-Benning, S., Fuss, H., Kossmann, J., Willmitzer, L. and Fisahn, J. 1997. Gas exchange and ultra-structural analysis of transgenic potato plants expressing mRNA antisense construct targeted to the cp-fructose-1,6-bisphosphate phosphatase. *Photosynthetica* 33: 455.
- Panda, S., Antoch, M.P., Miller, B.H., Su, A.I., Schook, A.B., Straume, M., Schultz, P.G., Kay, S.A., Takahashi, J.S. and Hogenesch, J.B. 2002. Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* 109: 307–320.
- Paul, M.J., Knight, J.S., Habash, D., Parry, M.A.J., Lawlor, D.W., Barnes, A.A., Loynes, A. and Gray, J.C. 1995. Reduction in phosphoribulokinase activity by antisense RNA in transgenic tobacco: effect on CO<sub>2</sub> assimilation and growth in low irradiance. *Plant J.* 7: 535.
- Pettersson, G. and Ryde-Pettersson, U. 1988. A mathematical model of the Calvin photosynthesis cycle. *Eur. J. Biochem.* 175: 661.
- Poolman, M., Fell, D. and Thomas, S. 2000. Modeling photosynthesis and its control. *J. Exp. Bot.* 51: 319.
- Poolman, M.G., Ölcer, H., Lloyd, J.C., Raines, C.A. and Fell, D.A. 2001. Computer modelling and experimental evidence for two steady states in the photosynthetic Calvin cycle. *Eur. J. Biochem.* 268: 2810.
- Puente, P., Wei, N. and Deng, X.W. 1996. Combinatorial interplay of promoter elements constitutes the minimal determinants for light and developmental control of gene expression in *Arabidopsis*. *EMBO J.* 15: 3732–3743.
- Price, G.D., Evans, J.R., Von Caemmerer, S., Yu, J.-W and Badger, M.R. 1995. Specific reduction of chloroplast glyceraldehyde-3-phosphate dehydrogenase activity by antisense RNA reduces CO<sub>2</sub> assimilation via a reduction in ribulose bisphosphate regeneration in transgenic tobacco plants. *Planta* 195: 369.
- Quick, W.P., Schurr, U., Scheibe, R., Schulze, E-D, Rodermel, S.R., Bogorad, L. and Stitt, M. 1991. Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with 'antisense' rbcS. I. Impact on photosynthesis in ambient growth conditions. *Planta* 183: 542.
- Roth, F.R., Hughes, J.D., Estep, P.E. and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nature Biotech.* 16: 939.
- Stitt, M., Quick, W.P., Schurr, U., Schulze, E-D, Rodermel, S.R. and Bogorad, L. 1991. Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with 'antisense' rbcS. II. Flux-control coefficients for photosynthesis in varying light, CO<sub>2</sub>, and air humidity. *Planta* 183: 555.
- The supplemental materials of this paper are also available at <http://zhao.med.yale.edu/Calvin Cycle>.

## Website references

## References

- ExPaSy Molecular Biology Server: <http://us.expasy.org/tools/pathways/>; <http://us.expasy.org/enzyme/>.
- Kyoto Encyclopedia of Genes and Genomes (KEGG): <http://www.genome.ad.jp/>.
- MIPs *Arabidopsis thaliana* database (MAtdB): <http://mips.gsf.de/proj/thal/>.
- National Center for Biotechnology Information (NCBI): <ftp://ftp.ncbi.nih.gov/>.
- Pathway MicroArray Processor for *Arabidopsis* (PathMAPA): <http://zhao.med.yale.edu/pathmapa.htm>.
- PLACE: <http://www.dna.affrc.go.jp/htdocs/PLACE/>.
- The Arabidopsis Information Resource (TAIR): <http://www.arabidopsis.org/>.
- Swiss-Prot Protein Knowledgebase: <http://us.expasy.org/sprot/>.
- TIGR *Arabidopsis thaliana* database: <http://www.tigr.org/tdb/e2k1/ath1/>.