Analysis of 101 nuclear transcriptomes reveals 23 distinct regulons and their relationship to metabolism, chromosomal gene distribution and co-ordination of nuclear and plastid gene expression

Alexander Biehl[a,1], Erik Richly[a,b,1], Christos Noutsos[a,b], Francesco Salamini[a], Dario Leister[a,b,*]

[a]*Abteilung für Pflanzenzüchtung und Ertragsphysiologie, Germany*
[b]*Abteilung für Pflanzenzüchtung und Genetik; Max-Planck-Institut für Züchtungsforschung, Carl-von-Linné Weg 10, D-50829 Köln, Germany*

Abstract

Post-endosymbiotic evolution of the proto-chloroplast was characterized by gene transfer to the nucleus. Hence, most chloroplast proteins are nuclear-encoded and the regulation of chloroplast functions includes nuclear transcriptional control. The expression profiles of 3292 nuclear *Arabidopsis* genes, most of them encoding chloroplast proteins, were determined from 101 different conditions and have been deposited at the GEO database (http://www.ncbi.nlm.nih.gov/geo/) under GSE1160–GSE1260. The 1590 most-regulated genes fell into 23 distinct groups of co-regulated genes (regulons). Genes of some regulons are not evenly distributed among the five *Arabidopsis* chromosomes and pairs of adjacent, co-expressed genes exist. Except regulons 1 and 2, regulons are heterogeneous and consist of genes coding for proteins with different subcellular locations or contributing to several biochemical functions. This implies that different organelles and/or metabolic pathways are co-ordinated at the nuclear transcriptional level, and a prototype for this is regulon 12 which contains genes with functions in amino acid and carbohydrate metabolism, as well as genes associated with transport or transcription. The co-expression of nuclear genes coding for subunits of the photosystems or encoding proteins involved in the transcription/translation of plastome genes (particularly ribosome polypeptides) (regulons 1 and 2, respectively) implies the existence of a novel mechanism that co-ordinates plastid and nuclear gene expression and involves nuclear control of plastid ribosome abundance. The co-regulation of genes for photosystem and plastid ribosome proteins escapes a previously described general control of nuclear chloroplast proteins imposed by a transcriptional master switch, highlighting a mode of transcriptional regulation of photosynthesis which is different compared to other chloroplast functions. From the evolutionary standpoint, the results provided indicate that functional integration of the proto-chloroplast into the eukaryotic cell was associated with the establishment of different layers of nuclear transcriptional control.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Chloroplast; DNA array; Gene expression regulation; Photosynthesis; Plastid ribosome; Regulon

1. Introduction

Most proteins in chloroplasts are encoded by nuclear genes (hereafter indicated as nuclear chloroplast genes) and translated as precursors on cytosolic ribosomes. While the regulation of plastid-encoded proteins includes transcriptional (Pfannschmidt et al., 1999), posttranscriptional (Rochaix, 2001), translational (Choquet and Wollman, 2002) and post-translational (Vothknecht and Westhoff,

2001) mechanisms, the plastid-to-nucleus signalling (Barkan and Goldschmidt-Clermont, 2000; Surpin et al., 2002) involves transcriptional regulation of nuclear chloroplast genes (Gray et al., 2003; Pfannschmidt et al., 2003; Richly et al., 2003a). Nuclear transcriptional control of chloroplast functions concerns, among many others, plastid ribosomal proteins (Kurth et al., 2002), subunits of the photosystems (Kurth et al., 2002; Zhong and Burns, 2003), the Calvin cycle and amino acid biosynthesis (Kleffmann et al., 2004). A DNA-array-based analysis of transcription of nuclear chloroplast genes, under 35 different conditions, highlighted the existence of a general regulatory mechanism, which acts in a binary mode by either inducing or repressing the same large set of nuclear chloroplast genes (Richly et al., 2003a).

While the transcription-based analysis of altered genetic or environmental states of an organism is a useful and frequent outcome of the study of mRNA expression profiles, also the definition of clusters of co-expressed genes, known as regulons, is a primary research target. Proteins encoded by members of a regulon can have a coherent function, or even be physically associated, like the components of the ribosome, proteasome and nucleosome (Brown and Botstein, 1999). In human, *Drosophila*, *Caenorhabditis* and budding yeast, co-regulated genes have been shown to cluster in the genome (Cohen et al., 2000; Caron et al., 2001; Roy et al., 2002; Spellman and Rubin, 2002).

We report on the definition and characterization of regulons of nuclear chloroplast genes of *A. thaliana* based on their mRNA expression data, recorded under 101 different environmental or genetic conditions. Some co-regulated nuclear chloroplast genes are organized as pairs of adjacent genes, and their products can, in cases, be associated with particular subcellular compartments or functions. More frequently, specific functions are distributed among regulons, supporting the existence of additional layers of gene regulation. Detailed analysis of functions associated with specific regulons shows that many photosynthetic and ribosome-protein-encoding genes are co-expressed, suggesting a nuclear transcriptional control of plastid ribosome abundance. This type of control should contribute to a co-ordinated expression of plastome- and nucleus-encoded proteins of the photosynthetic machinery.

2. Materials and methods

2.1. Plant material and growth conditions

Arabidopsis WT and mutant plants, propagated either under controlled greenhouse conditions or in sterile culture, were treated according to the procedures listed in the Supplementary Material 1. In general, total RNA was isolated from 4-week-old leaves, 2 h after the start of the day period, according to Kurth et al. (2002).

2.2. Data generation and analysis

Generation and use of a 3292-GST nylon array, enriched for nuclear chloroplast genes, has been described (Richly et al., 2003a). At least three experiments with cDNA probes from independent plant pools were carried out for each condition. cDNA synthesis was primed by a oligonucleotide mixture matching the 3292 genes in antisense orientation, and hybridized to the array as described (Kurth et al., 2002; Richly et al., 2003a). Hybridization images were read by phosphorimager (Storm 860; Molecular Dynamics, Sunnyvale, CA, USA), imported into *ArrayVision* (version 6.0; Imaging Research, St. Catharines, Ontario, Canada), and statistically evaluated using *ArrayStat* (version 1.0 Rev. 2.0; Imaging Research) as described (Pesaresi et al., 2003; Richly et al., 2003a). Data were normalized with reference to all spots on the array (Kurth et al., 2002) and average expression ratios derived from at least three independent experiments were analyzed by hierarchical clustering, k-means clustering or self-organizing-map clustering, using *Genesis* (version 1.5.0 b1) (Sturn et al., 2002).

2.3. Functional classification and assignment of subcellular location

Grouping of genes, based on the previous classification of Pesaresi et al. (2003) which was complemented with classes derived from the KEGG (http://www.genome.jp/kegg/) and AraCyc (http://www.arabidopsis.org/tools/aracyc/) databases, resulted into 15 functional classes: (i) amino acid metabolism; (ii) carbohydrate metabolism; (iii) secondary metabolism; (iv) other metabolic functions; (v) photosynthesis (dark reaction); (vi) photosynthesis (light reaction); (vii) protein modification and fate; (viii) protein phosphorylation; (ix) sensing, signalling and cellular communication; (x) stress response; (xi) transport; (xii) transcription; (xiii) protein synthesis; (xiv) other functions; and (xv) unknown function. Assignment of subcellular location to the encoded proteins was based on experimental evidence (Kruft et al., 2001; Friso et al., 2004; Heazlewood et al., 2004; Kleffmann et al., 2004) or on stringent computational prediction of subcellular targeting (Richly et al., 2003b; Richly and Leister, 2004). A list of all 1590 genes contained in the 23 regulons, as well as their assignment to functional classes and the subcellular location of the encoded proteins, is provided as Supplementary Material 2.

2.4. χ^2 comparisons

To compare the actual distribution of chromosomal or subcellular locations of genes (or their products) in each of the 23 regulons with the ones of chance distributions, the significance of the differences between regulons and the control group (all 3292 genes present in the array) was

determined using the χ^2 test. The degree of freedom (*df*) for the analysis of the distribution among five chromosomes was 4; for the subcellular distribution (chloroplast, mitochondrion, or other) it was 2.

2.5. Comparison of random and actual genomic distributions of co-expressed genes

The distribution of co-expressed genes in the genome of *A. thaliana* was compared by a randomization test to a random expectation. This was conducted by creating simulated genomes in which the chromosomal locations of co-regulated genes were randomly re-assigned. In each simulated genome, the number of co-expressed genes for each regulon was the same as in the real genome. The distribution of co-expressed genes in the actual genome was compared with that of 10,000 simulated genomes.

2.6. Data deposition and supplementary information

Complete data sets for each of the 101 expression profiles were deposited at the GEO database (http://www.ncbi.nlm.nih.gov/geo/) under accession numbers: GSE1160–GSE1260. Additional information, including a list of the 101 genetic and environmental conditions tested (Supplementary Material 1), and a list of all 1590 genes contained in the 23 regulons, as well as their assignment to functional classes and the subcellular location of the encoded proteins (Supplementary Material 2), is available at http://mpiz-koeln.mpg.de/~leister/101.html.

3. Results

3.1. Identification of regulons

The mRNA expression of 3292 nuclear genes, most of them coding for chloroplast proteins (Richly et al., 2003a), was determined under a total of 101 different environmental or genetic conditions (see Supplementary Material 1). Conditions known to affect chloroplast metabolism included mutations affecting photosynthesis, or protein import into plastids, or plastid-to-nucleus signalling, as well as treatments with different light quantities and qualities or herbicides. In addition, WT transcripts present in cells of different organs or of consecutive developmental stages of WT plants, as well as changes induced by other conditions, such as alterations in atmospheric CO_2 concentration, application of redox-active substances, wounding, or biotic stresses, were studied.

A set of 1590 genes was differentially regulated in at least 95 of the 101 expression profiles (Fig. 1A), and hierarchical clustering allowed their classification into between 21 and 27 groups of co-expressed genes, indicated as regulons. Basically the same set of regulons was also found by k-means or self-organizing-map clustering, and the

relatedness of the mean expression profiles of those 23 regulons which were identified by self-organizing-map clustering is shown in Fig. 1B. Genes of each of these 23 regulons were then analyzed for the chromosomal distribution of their genetic loci, their potential to code for proteins targeted to particular subcellular locations, and their participation to specific biochemical functions.

3.2. Chromosomal location of regulons

For each regulon, the distribution of the genetic loci encoding the concerned genes on the five *Arabidopsis* chromosomes was compared with the distribution of all 3292 genes contained in the DNA array. Employing a χ^2 test ($p < 0.01$), seven regulons were shown to include genes with a non-random distribution on chromosomes, compared to the 3292-gene-set (Fig. 2A).

A systematic, large-scale analysis of adjacent genes for their co-expression was complicated by the limited number of genes present on the 3292-gene-array. Nevertheless, 532 pairs and 62 triplets of adjacent genes were included in our array. Of those, 113 pairs and 6 triplets were also contained in the 1590-gene-set used for the definition of the 23 regulons. None of the triplets but eight pairs of adjacent genes were found in specific regulons. Five of the eight pairs of co-expressed genes consisted of highly homologous genes (*At1g29910/29920*; *At2g34420/34430*; *At5g38420/38430*; *At2g02380/02390*; *At4g12490/12500*) which most likely derived from tandem duplication and for which cross-hybridization as a cause for the observed co-expression cannot be excluded. The other three pairs contained non-homologous genes (*At2g04530/04540*; *At2g34460/34470*; *At3g52140/52150*), indicating that—similar to the case of budding yeast (Cohen et al., 2000)—co-expression of adjacent genes occurs in *A. thaliana*.

Although only eight pairs of adjacent, co-expressed genes existed in the 23 regulons, the actual fraction of such co-regulated gene pairs in the entire transcriptome of *Arabidopsis* might be much higher but remained undetected in our analysis due to the limited data set. To test whether larger clusters of co-expressed genes contributed to the non-random distribution between chromosomes shown in Fig. 2A, we compared the genomic organization of co-expressed genes for each regulon with a hypothetical chromosomal organization which would result from a completely random chromosomal distribution of co-regulated genes. As a matter of fact, the genomic distribution of co-expressed genes differed markedly from a random distribution (Fig. 2B). In particular, co-location of co-regulated genes was substantially less frequent in DNA regions spanning between 0 and 150 kbps compared to their random distribution situation. This indicates, that large, continuous clusters of co-regulated genes, such as in human (Caron et al., 2001) and *Drosophila* (Spellman and Rubin, 2002), apparently do not exist in *Arabidopsis*; on the contrary, the observed actual distribution is compatible
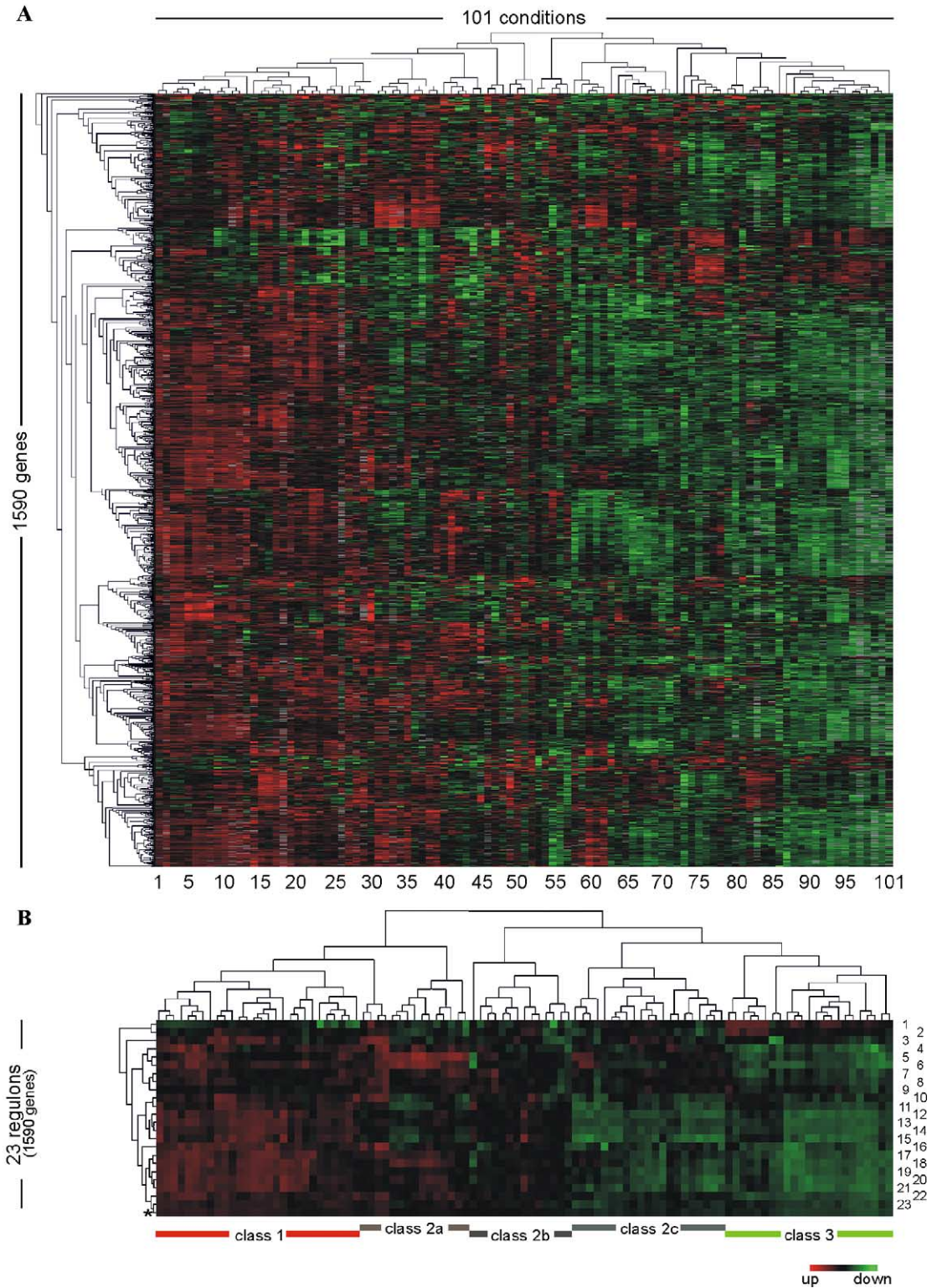
Fig. 1. Hierarchical clustering of expression profiles under 101 different conditions. (A) Hierarchical clustering of the expression profiles of 1590 genes that show differential expression under at least 95 of the 101 conditions tested. Colours indicate up- (red) or down- (green) regulation, grey lines indicate missing values. The cladograms on top and at left describe the relatedness of transcriptome responses in terms of conditions or as reaction of individual genes, respectively. The numbers at the bottom of the cluster refer to the 101 conditions (see Supplementary Material 1). (B) Hierarchical clustering of the mean expression profiles of 23 regulons, identified by self-organizing-map clustering. The cladograms on top and at left summarize the relatedness of the regulons. Designation of regulons from 1 to 23 is displayed at the right. The last row (asterisk) indicates the average of all 1590 genes contained in the 23 regulons. Regulons 1 and 2 are most distantly related to the average of all regulons (last row). Designation of response classes refers to the nomenclature defined by Richly et al. (2003a) (see Section 3.5).

with the existence of a certain fraction of pairs of adjacent, co-expressed genes which would be under-represented in a 3292-gene-sample with respect to the entire genome.

3.3. Subcellular protein location associated to regulons

Protein subcellular location in chloroplast, mitochondria, or other compartments, was assigned to each gene of the 23

regulons. Assignment was based on experimental evidence (Kruft et al., 2001; Friso et al., 2004; Heazlewood et al., 2004; Kleffmann et al., 2004) or on stringent computational prediction of subcellular targeting (Richly et al., 2003b; Richly and Leister, 2004). Because the location of only a minority of chloroplast and mitochondrial proteins has been reliably experimentally determined, and because the computational prediction applied was of high specificity but relatively low sensitivity (Richly et al., 2003b; Richly and Leister, 2004), our classification underestimates the number of proteins located in chloroplasts or mitochondria (it follows that in Fig. 2C the number of proteins in the category "other location" is overestimated). Regulons exclusively containing genes that code only for one specific subcellular compartment were not observed, supporting the view that nuclear chloroplast genes are integrated into large regulatory networks (Richly et al., 2003a). Four regulons, however, differed significantly (χ^2 test, $p<0.01$), when the distribution of the total of 3292 genes present on the array was taken as control (Fig. 2C), suggesting that those regulons have a higher degree of specificity for particular organelles than for others. This concerns, in particular, regulon 1 which consists to >80% of genes that are predicted or known to code for chloroplast proteins.

3.4. Regulons and their relationship to chloroplast functions

To each of the 1590 genes included in the 23 regulons, 1 of 15 different biochemical functions was assigned (see Section 2.3). The largest functional group (~40%) included proteins with unknown function (class xv). In Fig. 3A, for 8 of the 15 functional classes the distribution of genes among the 23 different regulons is shown, whereas Fig. 3B displays the distribution of functional classes within individual regulons. With exceptions, regulons were more or less
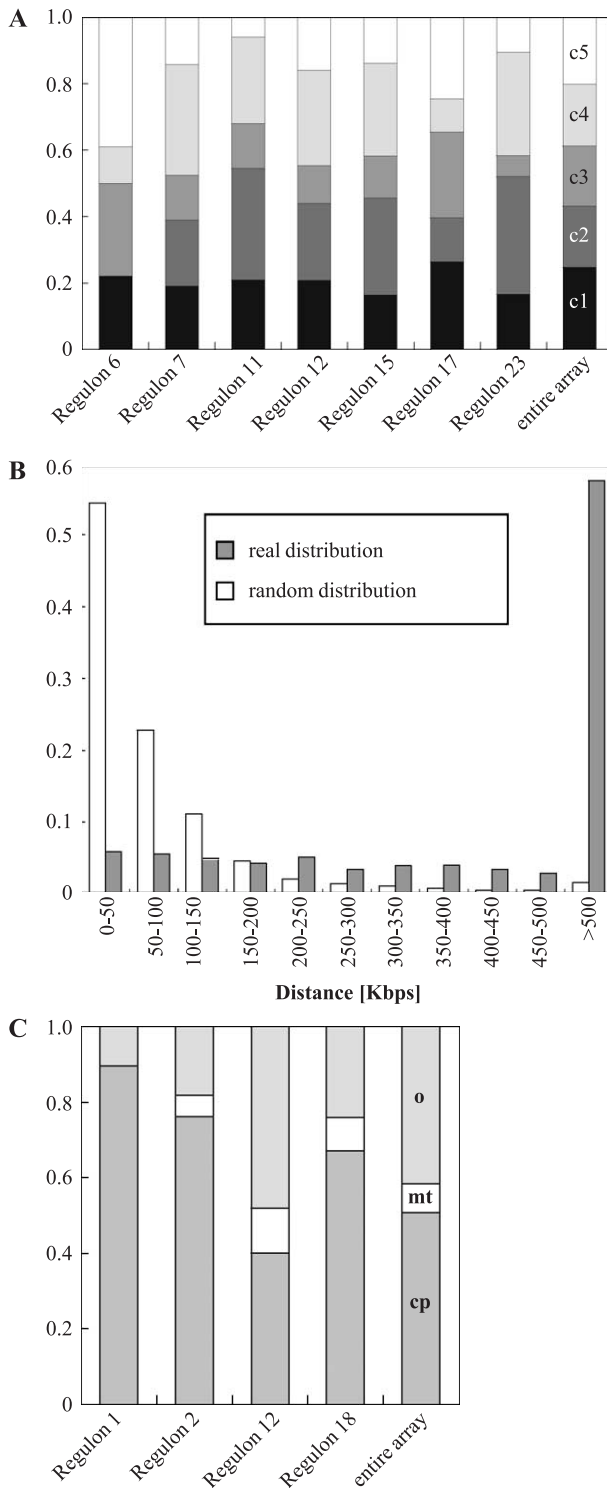
Fig. 2. Relationship among co-regulated genes with respect to their chromosomal location and the subcellular location of corresponding gene products. (A) Distribution of genes of seven regulons among the five *A. thaliana* chromosomes (c1 to c5), as well as of all 3292 genes contained in the array (last row). Frequencies correspond to the height of columns. For the seven regulons shown here, the χ^2 test provided statistical evidence ($p<0.01$) for a distribution different from the one of all 3292 genes present on the array. (B) Spatial distribution of co-expressed genes in the genome of *A. thaliana*. Co-regulated genes of each of the 23 regulons, shown in Fig. 1B, were analyzed. The actual distribution of co-expressed genes is displayed as relative frequency of distances between adjacent genes in 50-kbps intervals, ranging from a distance of between 0 and 50 kbps to a distance larger than 500 kbps. The actual distribution of co-regulated genes was compared with a random expectation by creating 10,000 simulated genomes in which the chromosomal locations of co-expressed genes were randomly re-assigned. In each simulated genome, the number of co-regulated genes in each regulon was the same as in the real genome. (C) Distribution of proteins encoded by genes of four specific regulons among different subcellular locations: cp, chloroplast; mt, mitochondria; o, other location. Conservative estimates were applied in the assignment of proteins to chloroplasts or mitochondria. The four regulons are shown for which χ^2 test provides statistical evidence ($p<0.01$) for a distribution different from the one for all genes represented by the 3292-GST array (last column).
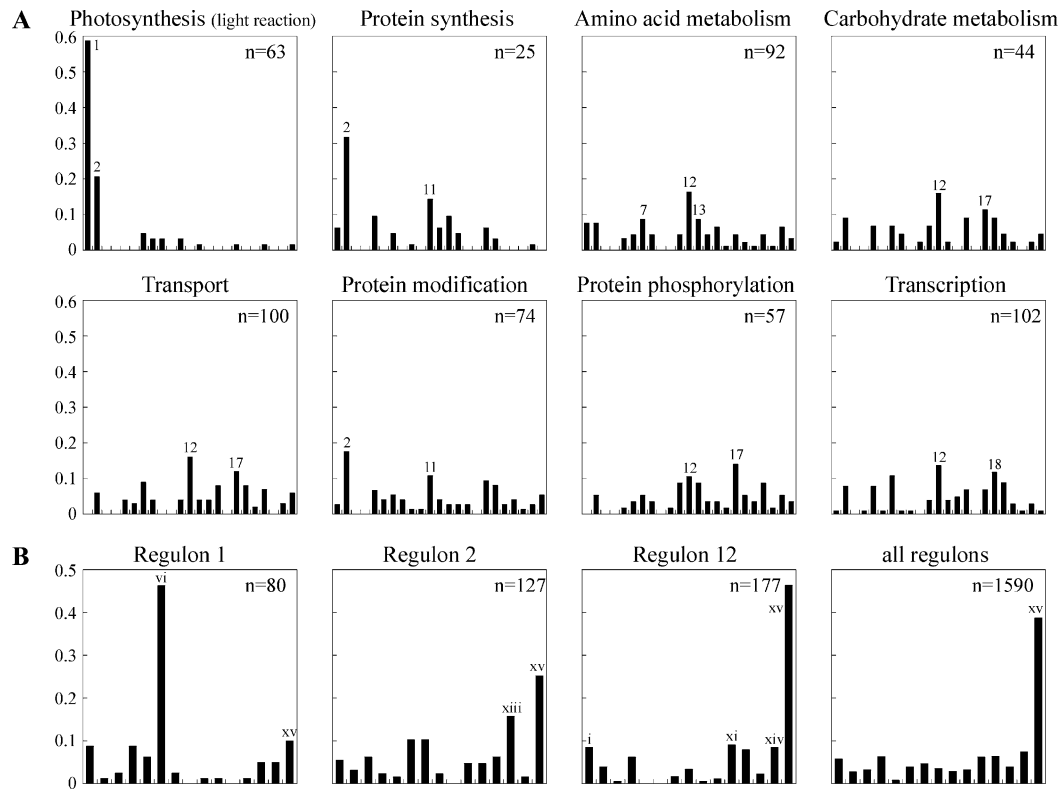
Fig. 3. Functional classification of regulons. (A) Histograms display for 8 of the 15 functional classes considered the distribution of genes among the 23 regulons. Column height corresponds to the number of genes contained in each regulon relative to the total number in each functional class (given in the left corner of each histogram). The designation of regulons from 1 to 23 is as in Fig. 3. Regulons over-represented in a particular functional class are indicated by numbers on the top of the appropriate column. (B) Distribution of genes among the 15 functional classes for selected regulons and for the total of 1590 genes included in regulons 1 to 23. Frequencies are symbolized by columns, and the number given in the right top corner of each histogram indicates the total number of genes contained in the concerned regulon. The 15 functional classes are in roman numbers, corresponding to the list provided in Materials and methods. For functional classes which are over-represented in a particular regulon, numbers are indicated at the appropriate column.

equally represented in each functional class, like if the co-ordination of different biochemical functions is, at least in part, regulated at the transcriptional level. Exceptions were amino acid (functional class i) and carbohydrate (ii) metabolism, protein synthesis (xiii), and the dark (class v) and light (vi) reactions of photosynthesis. Amino acid and carbohydrate metabolism, as well as transport (xi) and transcription (xii), were over-represented in regulon 12 (Fig. 3A; Table 1), suggesting a common co-ordination of genes belonging to the former two different pathways, in agreement with the finding that in plants nitrogen and carbon metabolism are, in part, interconnected (Foyer et al., 2003; Halford and Paul, 2003).

Regulons 1 and 2 have a very similar expression profile (see Fig. 1), which is most distantly related to the average expression profile of the complete 1590-gene-set (indicated by the asterisk in Fig. 1B). Regulon 2 is enriched for genes involved in chloroplast gene transcription and translation (xiii) (Fig. 3), in particular for genes encoding chloroplast ribosome proteins, as well as for genes associated with photosynthesis. Regulon 1 is exceptional, because it contains to a great extent proteins of known function, around half of them associated with the light reaction of photosynthesis. To a lesser extent, also in regulon 2 genes

for photosystem proteins are represented, and the contrary (plastid ribosomal genes in regulon 1) is evident, too. The co-regulation of genes for proteins of the plastid ribosome

Table 1
Regulon 12: exemplary genes encoding proteins of four different functional classes and with different subcellular location

Functional class	Examples
Amino acid metabolism	At3g49680: amino-acid transaminase (cp)
	At2g29690: anthranilate synthase (cp)
	At3g13110: serine acetyltransferase (mt)
Carbohydrate metabolism	At1g32900: putative starch synthase (cp)
	At5g03650: 1,4-α-glucan branching enzyme (cp)
	At5g51460: trehalose-6-phosphate phosphatase (o)
Transport	At2g29650: Na$^+$-dependent inorganic phosphate co-transporter (cp)
	At5g59250: D-xylose-H$^+$-symporter (cp)
	At2g29870: putative aquaporin (o)
Transcription	At2g25640: putative transcription elongation factor (cp)
	At3g54220: SCARECROW1 (cp)
	At2g34830: putative WRKY-type DNA-binding protein (o)

Subcellular location: cp, chloroplast; mt, mitochondrion; o, other location.

(mostly regulon 2) or of the photosystems (mostly regulon 1) can be interpreted as a mechanism to ensure that the subunits of those multi-protein complexes are synthesized in matching amounts. Co-variation of the expression of genes for photosynthetic proteins and of genes for the plastid ribosome (and to a lesser extent also of genes coding for proteins involved in the transcription of plastome genes) implies that nuclear transcriptional control of chloroplast ribosome abundance might be necessary to co-ordinate expression of plastome- and nucleus-encoded proteins of the photosynthetic machinery (see Discussion).

3.5. Five general types of nuclear transcriptome responses

The hierarchical clustering of the 1590 genes (Fig. 1A), as well as their grouping into 23 regulons (Fig. 1B), revealed the existence of five different types of nuclear transcriptome response. A group of about 28 treatments or genetic states induced a preferential up-regulation of almost all 1590 genes (class 1, bottom of Fig. 1B). A second group of conditions (class 3) supported down-regulation of largely the same genes. These two classes reflect the existence of a master-switch of the nuclear chloroplast transcriptome (Richly et al., 2003a). Only the regulons 1 and 2 escape the master switch by exhibiting markedly different directions of regulation. The response induced by the remaining conditions, which was previously described as class-2 response (Richly et al., 2003a) and which is characterized by about equal numbers of up- and down-regulated genes, can be assigned to three distinct classes (designated as 2a to 2c; Fig. 1B). This classification, which is based on the analysis of a much larger set of conditions than before (Richly et al., 2003a), is supported by the behaviour of groups of genes assigned to certain regulons. In Fig. 1B, class 2a is mainly defined by the up-regulation of regulons 5, 6, 17 and 18, whereas for class 2c strong down-regulation of regulons 11–15 is characteristic. In class 2b, regulons 17–23 are neither strongly up-, nor down-regulated, which is in contrast to classes 2a and 2c, where the concerned regulons are either up- or down-regulated, respectively.

4. Discussion

Control of transcript abundance is an effective strategy for regulating gene function. In eukaryotes, transcription rates are modulated by transcription factors, by remodelling of local chromatin structure through histone modification, as well as by transcriptional activation of euchromatic chromosomal regions. Finally, regulation of sets of adjacent genes has been demonstrated in a number of cases, including groups of co-expressed pairs or triplets of adjacent genes (Cohen et al., 2000), or larger clusters of co-regulated genes (Caron et al., 2001; Roy et al., 2002; Spellman and Rubin, 2002). From the evolutionary standpoint, the case of transcriptional control of gene expression presented in this paper has wide relevance for the understanding of how gene co-ordination has been achieved during the co-evolution of chloroplasts and their host cells. Nuclear chloroplast genes are, in fact, the final outcome of complex events, following the initial endosymbiosis of a cyanobacterial-like prokaryote by a eukaryotic cell. Available data indicate that several thousands of genes of the endosymbiont were gradually moved to the eukaryotic nucleus (reviewed in Timmis et al., 2004). This transfer went beyond mere physical integration of cyanobacterial genes into the eukaryote genome, and it had to involve a regulatory integration in terms of gene co-ordination, as described in its final state in this paper.

The expression analysis of a 3292-gene-set of *A. thaliana*, particularly enriched for nuclear chloroplast genes, indicates that some co-regulated genes are unevenly distributed among chromosomes. One could speculate that chromosomal clustering of genes for proteins targeted to a specific subcellular compartment, as described for some nuclear genes encoding mitochondrial proteins (Elo et al., 2003), might be responsible for the observed uneven distribution of co-expressed genes in *Arabidopsis*. This is not the case for the present data, because comparison of the actual and a randomized distribution of co-expressed genes along the five *Arabidopsis* chromosomes did not reveal a significant physical clustering to specific chromosomal regions (see Fig. 2B). A second possibility concerns the genomic clustering of genes associated with specific pathways, which exists to a varying extent in some eukaryotes (Lee and Sonnhammer, 2003; Spellman and Rubin, 2002). For regulons 1 and 2, however, which exhibit the strongest concentration of genes belonging to the same pathway, a distribution bias of the concerned genes across chromosomes (see Fig. 2A) or towards specific chromosomal regions (data not shown) was not detected. This raises the question of how the uneven inter-chromosomal distribution of genes of some regulons has evolved. Our data indicate the existence of very small and tightly linked groups of co-regulated genes, such as gene pairs, which might be unevenly distributed between chromosomes. It has been proposed, in fact, that acquisition of promoter sequences was the rate-limiting step for the successful integration of chloroplast genes into the nucleus (Martin and Herrmann, 1998). If organization as adjacent pairs in the nucleus represents a major mode of transcriptional co-regulation in plants, spatial and transcriptional co-evolution of groups of nuclear chloroplast genes should have been severely limited by their need to integrate or translocate, after insertion, nearby suitable loci.

The 1590 most-regulated genes were assigned to 23 regulons, which only marginally exhibited an association with specific biochemical functions and/or subcellular location. Most regulons contained genes involved in different biochemical functions or with products present in different subcellular compartments. This indicates the existence of complex networks of transcriptional regulation, such that different pathways and functions are co-ordinated

within and between different compartments. This highlights again the necessity that transferred genes acquired correct promoter sequences to support their proper integration into existing metabolic pathways. Many associations of regulons to specific functions or subcellular locations might have gone, however, undetected in our study due to incomplete *A. thaliana* gene annotation. The experimental and computational prediction of subcellular protein targeting is also still error-prone (Richly et al., 2003b; Richly and Leister, 2004).

The positive side of our analysis, despite these limitations, is that a number of relationships between co-expression and metabolic function of gene products became evident. Prominent examples are the regulons 1, 2 and 12. Regulon 12 contains genes associated with transport, transcription, amino acid and carbohydrate metabolisms— this is in line with previous suggestions that the later two pathways are in plants, to a certain degree, co-ordinated (Foyer et al., 2003; Halford and Paul, 2003). Regulon 1 is highly specific for genes encoding proteins of the photosynthetic apparatus, whereas a large fraction of genes in regulon 2 encodes proteins making up the chloroplast ribosome or are involved in other steps of chloroplast gene transcription/translation (see Fig. 3). In the future, this marked concentration of genes for proteins either involved in photosynthesis or in protein synthesis in the concerned regulons will enable an expression-profile-based approach for gene function discovery: any of the more than 40 genes with so far unknown function contained in these two regulons, in fact, is a bona fide candidate for a novel photosynthetic protein or a polypeptide with a function in chloroplast gene expression.

Regulons 1 and 2, as a matter of fact, show quite similar expression profiles, which are, on average, most distantly related to those of the other regulons. This establishes that nuclear genes for photosystem proteins or for proteins of the plastid ribosome have a distinct chloroplast-specific mode of transcriptional regulation, which involves only few proteins targeted to mitochondria or to other locations. Furthermore, co-regulation of genes encoding proteins of the photosynthetic machinery and of subunits of the plastid ribosome implies the existence of a mechanism that co-ordinates the expression of plastome- and nucleus-encoded proteins existing in the photosystems. It appears plausible that in such complexes, when nuclear-encoded photosystem proteins increase, this has to be matched by a similar modification of the level of plastome-encoded proteins. Our data indicate that this is achieved as a nucleus-dependent increase in plastid ribosomal proteins—ultimately of the number of ribosomes—in support of an increase of the translational capacity of chloroplasts.

Our analysis confirms the existence of a transcriptional master switch, previously described to regulate the expression of largely overlapping sets of genes in opposite directions (Richly et al., 2003a) (class 1 and class 3 in Fig. 1B). A mixed response to alterations in the environ-

ment or genetic state, all eliciting a more balanced transcriptional activity, can now be subdivided into the three subclasses 2a–2c (Fig. 1B). This underlines that additional levels of gene expression regulation exist in the plant cell. Expression of nuclear photosystem and of chloroplast ribosome proteins escapes the general response supported by the master switch. Thus, a distinct type of nuclear transcriptional regulation should exist for these specific groups of genes, demonstrating that the functional integration of the cyanobacterial endosymbiont into the eukaryotic cell was associated with fundamental changes in its mode of regulation (Martin and Herrmann, 1998), which included the establishment of different layers of nuclear transcriptional control.

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at 10.1016/j.gene.2004.09.009.

References

Barkan, A., Goldschmidt-Clermont, M., 2000. Participation of nuclear genes in chloroplast gene expression. Biochimie 82, 559–572.

Brown, P.O., Botstein, D., 1999. Exploring the new world of the genome with DNA microarrays. Nat. Genet. 21, 33–37.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., Heisterkamp, S., van Kampen, A., Versteeg, R., 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291, 1289–1292.

Choquet, Y., Wollman, F.A., 2002. Translational regulations as specific traits of chloroplast gene expression. FEBS Lett. 529, 39–42.

Cohen, B.A., Mitra, R.D., Hughes, J.D., Church, G.M., 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. 26, 183–186.

Elo, A., Lyznik, A., Gonzalez, D.O., Kachman, S.D., Mackenzie, S.A., 2003. Nuclear genes that encode mitochondrial proteins for DNA and RNA metabolism are clustered in the *Arabidopsis* genome. Plant Cell 15, 1619–1631.

Foyer, C.H., Parry, M., Noctor, G., 2003. Markers and signals associated with nitrogen assimilation in higher plants. J. Exp. Bot. 54, 585–593.

Friso, G., Giacomelli, L., Ytterberg, A.J., Peltier, J.B., Rudella, A., Sun, Q., van Wijk, K.J., 2004. In-depth analysis of the thylakoid membrane proteome of *Arabidopsis thaliana* chloroplasts: new proteins, new functions, and a plastid proteome database. Plant Cell 16, 478–499.

Gray, J.C., Sullivan, J.A., Wang, J.H., Jerome, C.A., MacLean, D., 2003. Coordination of plastid and nuclear gene expression. Philos. Trans. R. Soc. Lond., B Biol. Sci. 358, 135–144 (Discussion 144–135).

Halford, N.G., Paul, M.J., 2003. Carbon metabolite sensing and signalling. Plant Biotechnol. J. 1, 381–398.

Heazlewood, J.L., Tonti-Filippini, J.S., Gout, A.M., Day, D.A., Whelan, J., Millar, A.H., 2004. Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. Plant Cell 16, 241–256.

Kleffmann, T., Russenberger, D., von Zychlinski, A., Christopher, W., Sjölander, K., Gruissem, W., Baginsky, S., 2004. The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. Curr. Biol. 14, 354–362.

Kruft, V., Eubel, H., Jansch, L., Werhahn, W., Braun, H.P., 2001. Proteomic approach to identify novel mitochondrial proteins in *Arabidopsis*. Plant Physiol. 127, 1694–1710.

Kurth, J., Varotto, C., Pesaresi, P., Biehl, A., Richly, E., Salamini, F., Leister, D., 2002. Gene-sequence-tag expression analyses of 1800 genes related to chloroplast functions. Planta 215, 101–109.

Lee, J.M., Sonnhammer, E.L., 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13, 875–882.

Martin, W., Herrmann, R.G., 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 118, 9–17.

Pesaresi, P., Gardner, N.A., Masiero, S., Dietzmann, A., Eichacker, L., Wickner, R., Salamini, F., Leister, D., 2003. Cytoplasmic N-terminal protein acetylation is required for efficient photosynthesis in *Arabidopsis*. Plant Cell 15, 1817–1832.

Pfannschmidt, T., Nilsson, A., Allen, J.F., 1999. Photosynthetic control of chloroplast gene expression. Nature 397, 625–628.

Pfannschmidt, T., Schutze, K., Fey, V., Sherameti, I., Oelmuller, R., 2003. Chloroplast redox control of nuclear gene expression—a new class of plastid signals in interorganellar communication. Antioxid. Redox Signal. 5, 95–101.

Richly, E., Leister, D., 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. Gene 329, 11–16.

Richly, E., Dietzmann, A., Biehl, A., Kurth, J., Laloi, C., Apel, K., Salamini, F., Leister, D., 2003a. Covariations in the nuclear chloroplast transcriptome reveal a regulatory master-switch. EMBO Rep. 4, 491–498.

Richly, E., Chinnery, P.F., Leister, D., 2003b. Evolutionary diversification of mitochondrial proteomes: implications for human disease. Trends Genet. 19, 356–362.

Rochaix, J.D., 2001. Posttranscriptional control of chloroplast gene expression from RNA to photosynthetic complex. Plant Physiol. 125, 142–144.

Roy, P.J., Stuart, J.M., Lund, J., Kim, S.K., 2002. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature 418, 975–979.

Spellman, P.T., Rubin, G.M., 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. J. Biol. 1, 5.

Sturn, A., Quackenbush, J., Trajanoski, Z., 2002. Genesis: cluster analysis of microarray data. Bioinformatics 18, 207–208.

Surpin, M., Larkin, R.M., Chory, J., 2002. Signal transduction between the chloroplast and the nucleus. Plant Cell 14, S327–S338. Suppl.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev., Genet. 5, 123–135.

Vothknecht, U.C., Westhoff, P., 2001. Biogenesis and origin of thylakoid membranes. Biochim. Biophys. Acta 1541, 91–101.

Zhong, G.V., Burns, J.K., 2003. Profiling ethylene-regulated gene expression in *Arabidopsis thaliana* by microarray analysis. Plant Mol. Biol. 53, 117–131.