Cell Research (2008) 18:1007-1019.
© 2008 IBCB, SIBS, CAS All rights reserved 1001-0602/08 $ 30.00
www.nature.com/cr

ORIGINAL ARTICLE

Construction of a chloroplast protein interaction network and functional mining of photosynthetic proteins in *Arabidopsis thaliana*

Qing-Bo Yu[1,5,*], Guang Li[2,4,*], Guan Wang[1,*], Jing-Chun Sun[2,6], Peng-Cheng Wang[1], Chen Wang[1], Hua-Ling Mi[3], Wei-Min Ma[3], Jian Cui[4], Yong-Lan Cui[1], Kang Chong[5], Yi-Xue Li[7], Yu-Hua Li[4], Zhongming Zhao[6], Tie-Liu Shi[2,7], Zhong-Nan Yang[1]

[1]College of Life and Environmental Sciences, Shanghai Normal University, Shanghai 200234, China; [2]Shanghai Information Center for Life Science, Shanghai Institutes for Biological Sciences, The Chinese Academy of Sciences, Shanghai 200032, China; [3]National Key Laboratory of Plant Molecular Genetics, Institute of Plant Physiology and Ecology, The Chinese Academy of Sciences, Shanghai 200032, China; [4]College of Life Sciences, The Northeast Forestry University, Harbin, Heilongjiang 150040, China; [5]Research Center for Molecular and Developmental Biology, Key Laboratory of Photosynthesis and Environmental Molecular Physiology, Institute of Botany, The Chinese Academy of Sciences, Beijing 100093, China; [6]Department of Psychiatry and Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23298, USA; [7]Bioinformation Center, Key Laboratory of System Biology, Shanghai Institutes for Biological Sciences, The Chinese Academy of Sciences, Shanghai 200232, China

Chloroplast is a typical plant cell organelle where photosynthesis takes place. In this study, a total of 1 808 chloroplast core proteins in *Arabidopsis thaliana* were reliably identified by combining the results of previously published studies and our own predictions. We then constructed a chloroplast protein interaction network primarily based on these core protein interactions. The network had 22 925 protein interaction pairs which involved 2 214 proteins. A total of 160 previously uncharacterized proteins were annotated in this network. The subunits of the photosynthetic complexes were modularized, and the functional relationships among photosystem I (PSI), photosystem II (PSII), light harvesting complex of photosystem I (LHC I) and light harvesting complex of photosystem I (LHC II) could be deduced from the predicted protein interactions in this network. We further confirmed an interaction between an unknown protein AT1G52220 and a photosynthetic subunit PSI-D2 by yeast two-hybrid analysis. Our chloroplast protein interaction network should be useful for functional mining of photosynthetic proteins and investigation of chloroplast-related functions at the systems biology level in *Arabidopsis*.

***Keywords*:** *Arabidopsis*, chloroplast protein, network, functional linkage, photosynthesis
Cell Research (2008) **18**:1007-1019. doi: 10.1038/cr.2008.286; published online 23 September 2008

Introduction

Chloroplast is an essential organelle in plant cells and green algae. Most proteins in the chloroplast are encoded by the nuclear genome; only 88 proteins are encoded by the chloroplast genome (http://www.arabidopsis.org). In all, 4 255 nuclear-encoded proteins in the plant plastid database (PPDB) were predicted to be likely targets to *Arabidopsis* chloroplasts based on combinatory evidence from TargetP (http://www.cbs.dtu.dk/services/TargetP/), LumenP [1] and the predictor of a-helical TM proteins, TM-HMM version 2.0 (http://www.cbs.dtu.dk/services/

*These three authors contributed equally to this work.
Correspondence: Zhong-Nan Yang[a], Tie-Liu Shi[b]
[a]Tel: +86-2164324650; Fax: +86-2164322142
E-mail: znyang@shnu.edu.cn
[b]Tel: +86-2154922980; E-mail: tlshi@sibs.ac.cn

Abbreviations: TAIR (the *Arabidopsis* information database); PPDB (plant plastid database); MS (mass spectrometry); PSI (photosystem I); PSII (photosystem II); LHC I (light-harvesting complex of photosystem I); LHC II (light-harvesting complex of photosystem II); GO (gene ontology); GFP (green fluorescent protein); GOM (global optimization method); FDR (false discovery rate); CCs (correlation coefficients); BLAST (basic local alignment search tool)

Received 15 October 2007; revised 10 January 2008; accepted 7 April 2008; published online 23 September 2008

TMHMM/) [2]. Another prediction program, Predotar (v1.03), predicted 1 591 plastid proteins [3]. It was estimated that the *Arabidopsis* chloroplast might contain 2 100-3 600 distinct proteins [4]. So far, only 1 284 nonredundant nuclear-encoded chloroplast proteins have been identified by proteomic techniques based on available published data and Plprot database, a plastid protein database [1, 5-17]. Thus, the *Arabidopsis* chloroplast proteome is largely undetermined.

Protein interactions can be classified into two major categories: physical and functional. The former refer to physical association between two proteins, whereas the latter refer to the proteins in a biochemical or a signaling pathway [18]. In fact, protein interactions are very important for cellular and organelle functions. Global protein interaction networks have been constructed for several eukaryotes, including *Saccharomyces cerevisiae* [19, 20], *Drosophila melanogaster* [21], *Caenorhabditis elegans* [22] and *Homo sapiens* [23], by high-throughput yeast two-hybrid screens. However, the assembly of protein interaction networks in plants considerably lags behind the efforts in animals [24]. Only a few interaction networks in *Arabidopsis thaliana* have been reported [25, 26]. Among them, Geisler-Lee *et al*. [27] constructed a genome-wide protein interactome in *Arabidopsis*, but it was based only on the interologs data in other model organisms.

Large-scale experiments, such as yeast two-hybrid analysis, expression profiling and 2-D gels with mass spectrometry (MS), have accumulated mass of data allowing for the development of a number of bioinformatics approaches for predicting protein-protein interactions [28]. Those approaches include text mining [29], computational methods based on conserved protein interactions (i.e., interologs) [30], gene expression data [31, 32] and correlation of genomic context (i.e., gene neighbor algorithm) [33, 34], gene fusion (Rosetta Stone method) [35, 36], phylogenetic profiles [37, 38], and gene clustering (operon) [39, 40]. Since each computational method is inherently biased towards and limited to the biological features it uses [41], new strategies or algorithms have been developed, such as Bayesian network modeling [42], to integrate protein interaction and related biological data. The classifier in Bayesian network modeling integrates all data based on the gold standard positives (GSPs) and gold standard negatives (GSNs) by applying machine learning theory. The Bayesian method has recently been successfully applied to integrate different data types, yielding probabilistic measurements of yeast and human interactomes [42-45].

In this study, we applied an integrative genomics strategy to identify a set of reliable chloroplast proteins in *Arabidopsis*. Based on these proteins, a chloroplast protein interaction network was constructed for *Arabidopsis thaliana*, which can be used for searching high-confidence protein interaction data. We use this network to identify and annotate previously uncharacterized proteins. In particular, we captured the photosynthetic complexes and some novel interacting proteins using our constructed network. Thus, this network provides a useful resource for further investigation of chloroplast protein functions and a comprehensive understanding of the photosynthetic mechanism.

Results

Identification of chloroplast core proteins through integrative genomics in Arabidopsis thaliana

In order to integrate a set of reliable chloroplast proteins, currently available data were collected and extensively analyzed. The *Arabidopsis* chloroplast proteins were collected from Predotar [46], Plprot [17], PPDB [2], and extracted from The *Arabidopsis* Information database (TAIR) based on the gene ontology (GO) cellular-component annotation. Also, the chloroplast proteins were predicted based on BLASTP analysis, co-expression analysis, the unique Pfam domain analysis and published chloroplast proteomics data. A total of 7 592 nonredundant putative chloroplast proteins were obtained from the nine different datasets (Table 1, Supplementary information, Table S1).

To identify a set of reliable chloroplast proteins (core data or core proteins) from the nine different resources, a naïve Bayesian classifier [47] was used. As the Plprot and MS method, as well as PPDB data and gene ontology (GO) method are not independent based on the analysis of Pearson correlation coefficients (CCs), these data (Plprot and MS, PPDB and GO) were regarded as one feature, respectively. Using a conservative threshold ($LR_{protein}$>1.14, Figure 1), 1 534 proteins were reliably determined to be chloroplast proteins. Finally, 1 720 nuclear-encoded chloroplast proteins were identified after further integration of data from MS, Swiss-prot, missed GSP, and dual targeting proteins (Supplementary information, Figure S1 and Supplementary information, Table S2). Considering the 88 chloroplast genome-encoded proteins, the set of reliable chloroplast proteins amounts to 1 808. The remaining 5 784 putative chloroplast proteins were classified as candidate chloroplast proteins.

A chloroplast protein interaction network for Arabidopsis thaliana

A chloroplast protein interaction network was built based on the chloroplast core and candidate proteins.

Table 1 Nine genome-wide datasets used to integrate for reliable chloroplast proteins

Dataset method	Data source	Protein predicted	FDR[1] (%)
1 PPDB data	http://ppdb.tc.cornell.edu/	4 255	30.3
2 Predotar	http://urgi.infobiogen.fr/predotar/	1 000	42.2
3 Plprot	http://www.plprot.ethz.ch/	807	31.3
4 Gene ontology	http://www.arabidopsis.org	4 073	30.3
5 Bacteria homology	All bacteria photosynthetic proteins ortholog	243	14.3
6 Plant homology	Other plant chloroplast proteins ortholog	1 846	47.9
7 Coexpression	Coexpression with known chloroplast genes in *Arabidopsis* tissue atlases	1 585	33.4
8 Protein domain	Pfam domain found only in eukaryotic chloroplast proteins (SwissProt)	408	20.2
9 MS/MS	*Arabidopsis* chloroplast proteomics	1 205	24.9
This work		1 534	4.77

Nine individual data sources and an integrated approach in this work were used to predict chloroplast localized proteins in the *Arabidopsis thaliana* proteome. The genome-wide false discovery rate was estimated from a large gold standard training dataset.
[1]FDR: false discovery rate.

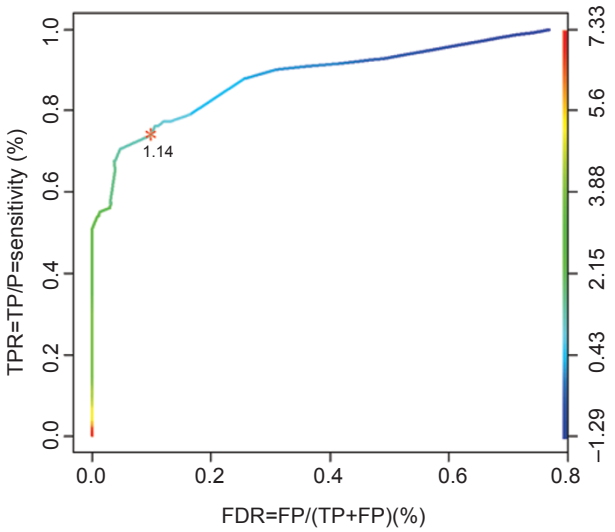

Figure 1 The sensitivity and false discovery rate of chloroplast protein prediction by integrative genomic methods. Using training datasets of 594 gold positive proteins and 2 460 gold negative proteins, the sensitivity and false discovery rate of the integrative methods were estimated. The specific thresholds ($LR_{protein}$>1.14) are marked by an asterisk.

First, a genome-wide protein interaction network was constructed in *Arabidopsis thaliana* as described by Cui *et al.* [48]. Protein interaction pairs between core proteins and core/candidate proteins were then screened from the genome-wide interaction data (Supplementary information, Table S3). Since some unknown core protein interactions cannot be predicted with these methods, the relative CCs of all genes from gene chip data were also calculated. Of the interaction pairs with a CC greater than 0.95, pairs of core proteins and core/candidate proteins were picked up (Supplementary information, Table S4). By integrating the above data, the chloroplast protein interaction network was built with 22 925 interaction pairs involving 2 214 proteins. Of these proteins, 1 043 were chloroplast core proteins and the remaining 1 171 were candidate chloroplast proteins. Confidence values for each interaction pair were assigned (LR_{pair} value) as described in Materials and Methods. A total of 5 784 protein pairs, involving 967 different proteins, exhibited high confidence interactions, with an LR_{pair} value > *f*850.

Previous studies in yeast revealed that proteins from the same functional group tend to have connections [49]. Under this assumption, the degrees of functional linkages among the characterized proteins were evaluated as a measure of the reliability of the network. In the network, 2 010 of the 2 214 proteins were assigned to 33 broad functional categories in the Mapman database. We generated 100 scrambled networks with these 2 010 chloroplast proteins as a control. In the scrambled networks, the average ratio of the proteins that is consistent with actual annotation was merely 6.95%. However, our network allowed a correct prediction of 57.6% of 2 010 characterized proteins with at least one annotated partner. This result indicated that the composition of the network is consistent with the proposed biological functions, and that it can provide information for further experimental research.

Topological properties of the chloroplast protein interaction network

Network topology offers a quantifiable description of the networks that characterize various biological systems [50]. The chloroplast protein interaction network was

analyzed with topological tools, including departing the whole network, calculating the average distance, degree distribution and clustering coefficient. Computational analysis revealed one main network of 3 109 protein interactions between 309 proteins, and another 84 small isolated networks of more than two proteins. For the largest interaction network, the mean shortest path length between any two proteins in the network is 4.07 links (Figure 2A), an indication that most of the proteins were very closely linked [51].

In order to measure the probability that a given protein interacts with *k* other proteins, we calculated the degree distribution *P*(*k*) of the proteins involved in the network. As shown in Figure 2B, the degree distribution decreased slowly, approximately following a power-law, meaning that the interaction map has scale-free properties. The average clustering coefficient can be used to measure the tendency of proteins to form clusters or groups, which gives an indication of hierarchal character in a network. Figure 2C shows the distribution of the clustering coefficient *C*(*k*) on a log-log plot. The average *C*(*k*) diminished as the number of interactions per protein increased, indicating that the network had a potential hierarchical modularity [50, 52].

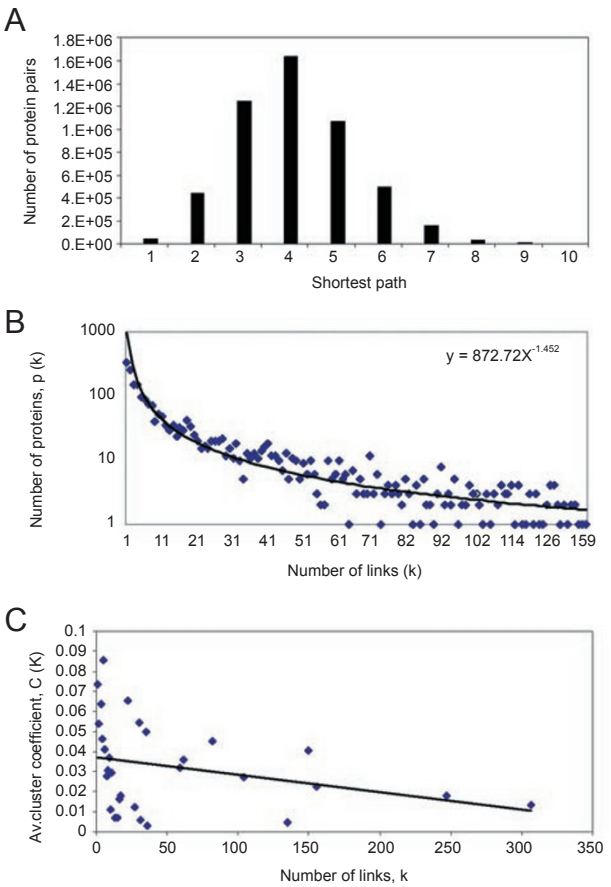

Figure 2 Properties of the chloroplast protein interaction network in *Arabidopsis thaliana*. **(A)** The distribution of the shortest path between pairs of proteins in the chloroplast protein interaction network. On average, any two proteins in the network are connected via 4.07 links. **(B)** The degree distribution of the proteins in the network. The number of proteins with a given link (*k*) in the network approximates the power-law ($P(k) \sim k^{g}$; g = 1.40). **(C)** The degree distribution of the clustering coefficients of the network proteins. The average clustering coefficient of all nodes with *k* links was plotted against the number of links ($CC_p = 2n/k_p(k_p-1)$), with *n* as the number of links connecting the k_p neighbors of node p to each other.

Annotation of unknown proteins

Approximately 13.9% (307) of the proteins in the network have not been characterized. Among these, 160 proteins were annotated based on our protein interaction network, and were assigned to functional categories in the Mapman database using the global optimization method (GOM) as described by Vazquez *et al.* [53]. Of these, 57 were assigned to one functional category, and the rest (103) were placed into two or more functional categories (Supplementary information, Table S5). Of the 160 newly annotated proteins, all belonged to the chloroplast core proteins, and 53 were assigned to Bin 1, the functional category of photosynthesis.

The photosynthetic complexes in the chloroplast protein interaction network

Chloroplasts in green algae and higher plants contain five major protein complexes, four of which are in the thylakoid membrane (photosystem I [PSI], photosystem II [PSII], ATP synthase and cytochrome b6f complexes). All of these complexes play important roles in photosynthesis. Thus, we set out to determine if these complexes could be detected in the chloroplast protein interaction network. A sub-network with only predicted core protein interactions was extracted. In this sub-network, most of the subunits of the five known complexes grouped together to form modules (Figure 3A). This clustering suggested strong functional relationships among these proteins.

Light-harvesting complex of photosystem II (LHC II) located in the thylakoid membrane collects energy from sunlight and transfers it to the PSII reaction center in the form of excitation energy. Light-harvesting complex of photosystem I (LHC I) proteins form peripheral monomeric antennas of PSI with chlorophyll and absorb light energy. We extracted a sub-network using all proteins predicted to interact with LHC II subunits. In this sub-network, LHC II subunits connected not only with PSII, but also with PSI and LHC I. This clustering suggested

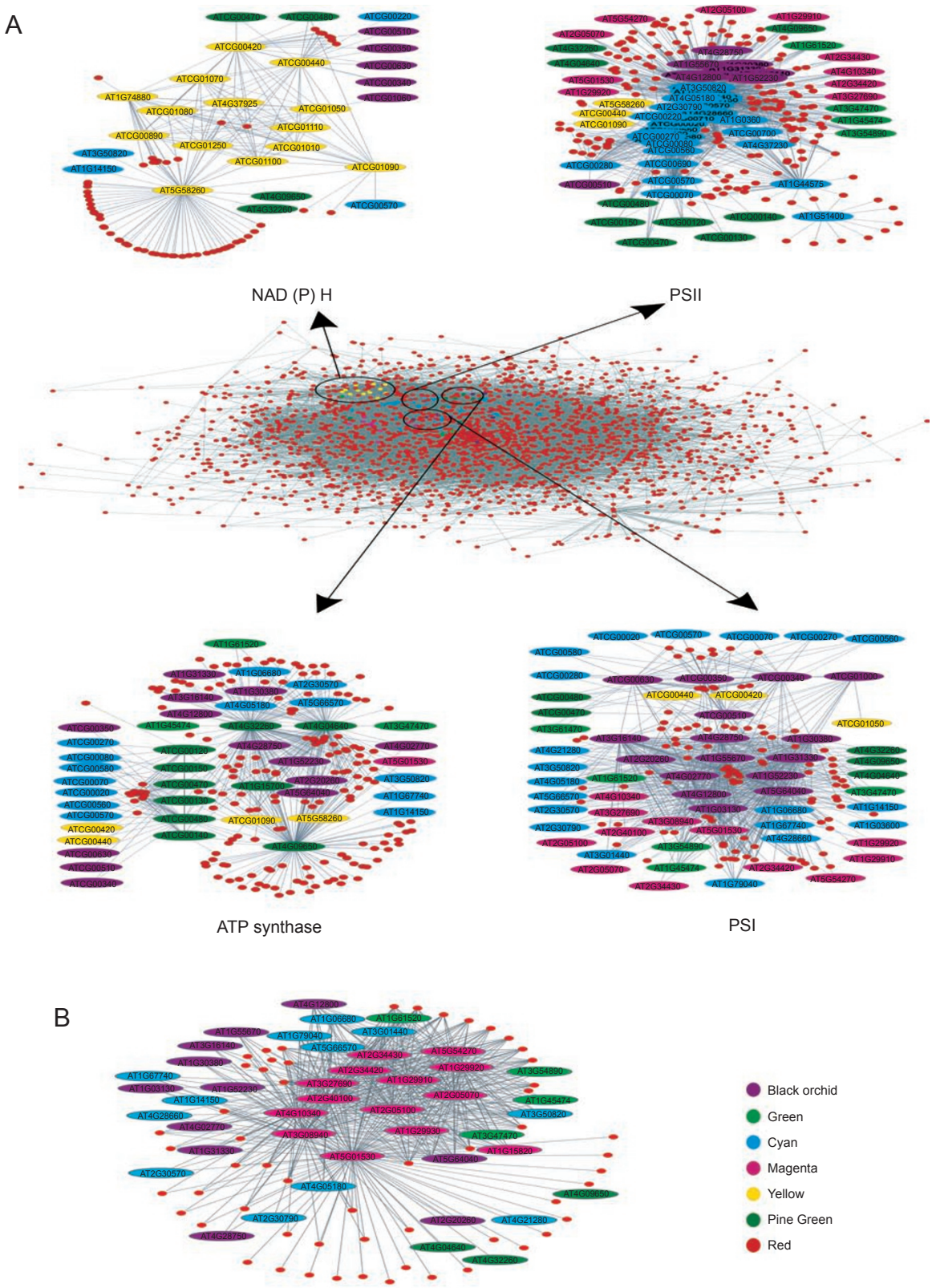

Figure 3 The photosynthetic complexes in the chloroplast protein interaction network. **(A)** Overviews and magnified views of the four photosynthetic complexes in the chloroplast protein interaction network; **(B)** the LHC II complex functionally relates to the LHC I complex, Photosystem I core subunits and the Photosystem II core subunits. A sub-network with only predicted photosynthetic complexes was extracted and shown. Each color represents the components of a photosynthetic complex in the interaction network (DarkOrchid: Photosynthetic complex I proteins; Green: LHC I complex proteins; Cyan: Photosynthetic complex II proteins; Magenta: LHC II complex proteins; Pine Green: ATP synthase; Yellow: NAD(P)H complex proteins; Red: other proteins).

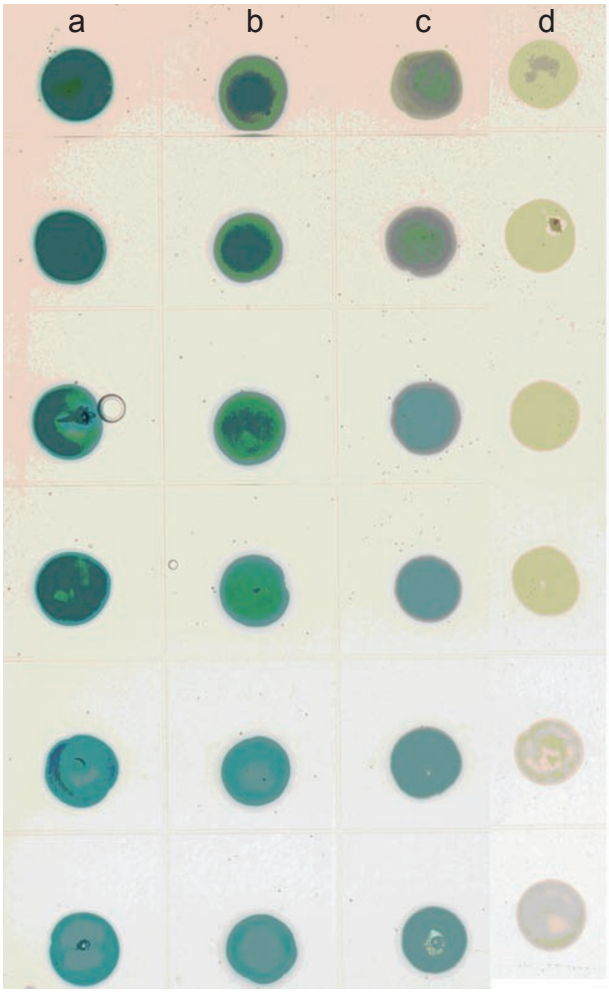

Figure 4 The results of yeast two-hybrid experiments for two protein interactions. **(A)** Positive control; **(B)** PSI-D2 and AT1G52220; **(C)** PSI-L and AT2G46820; **(D)** negative control.

that those complexes are fully functionally linked (Figure 3B).

A novel physical interaction validated by yeast two-hybrid experiments

The predicted protein interaction suggests their functional linkage including physical interactions. To test whether physical interactions exist, we randomly chose 12 novel interactions involving photosynthesis related proteins from the network and tested them by yeast two-hybrid analysis (Supplementary information, Table S6). Our results indicated that AT1G52220 and AT2G46820 (TMP14) interacted with PSI-D2 and PSI-L, respectively (Figure 4). One pair (AT2G46820 interacts with PSI-L) was previously verified by immunoblot analysis by Khrouchtchova *et al.* [54]. In order to obtain the subcellular localization information of AT1G52220, its full-length coding region was fused with green fluorescent protein (GFP) and subsequently introduced into the wild-type Columbia plant. Confocal laser microscopy analysis showed that GFP fluorescence was localized in the thylakoid of *Arabidopsis* chloroplast (Figure 5), suggesting that AT1G52220 is a thylakoid protein. The subcellular localization information supports that AT1G52220 physically interacts with PSI-D2.

Discussion

The chloroplast core proteins in Arabidopsis thaliana

A chloroplast is a typical semi-autonomous plant cell organelle and its proteins are encoded by both the nuclear genome and the plastid genome. Although the *Arabidopsis* genome has been completed for several years, there still exists uncertainty on the total chloroplast proteins. To date, several methods have been developed to predict protein subcellular localization on the basis of protein sequence, structure and evolution [55]. However, each method has its intrinsic biases. In this paper, we extensively collected and predicted chloroplast proteins. A total of 7 592 proteins were obtained and catalogued as chloroplast candidate proteins. These chloroplast candidate proteins are up to 27% of the *Arabidopsis* total proteome. We believe that these proteins should cover almost all *Arabidopsis* chloroplast proteins. However, these candidate data were only considered as data source, and the purpose of the process was to cover the chloroplast proteins as completely as possible for the next step: integration.

Obtaining reliable chloroplast proteomic data was the premise for construction of the protein interaction network. After integration, a total of 1 808 proteins are considered as reliable chloroplast proteins. Of these proteins, 1 720 were nuclear encoded. These proteins were compared with the 4 255 proteins from PPDB and the 1 284 proteins integrated from proteomics data (Figure 6). Of the 1 284 proteins, 710 were also included in the PPDB data. These proteins should be reliable chloroplast proteins since they were included in both the PPDB and integrated proteomics data. Of the 710 overlapping proteins, 705 were also present in our core data of chloroplast proteins, further demonstrating the reliability of our core data. However, 354 of the 1 284 integrated proteomics proteins were not included in PPDB data and our core data. Recently, Friso *et al.* [1] identified 4 255 *Arabidopsis* chloroplast candidate proteins based on TargetP prediction, which relies on the N-terminal signal sequence of a protein (PPDB, http://cbsusrv01.tc.cornell.edu/users/ppdb/). However, it has been known that some proteins that are located on the chloroplast outer membrane do not contain the signal sequence and those

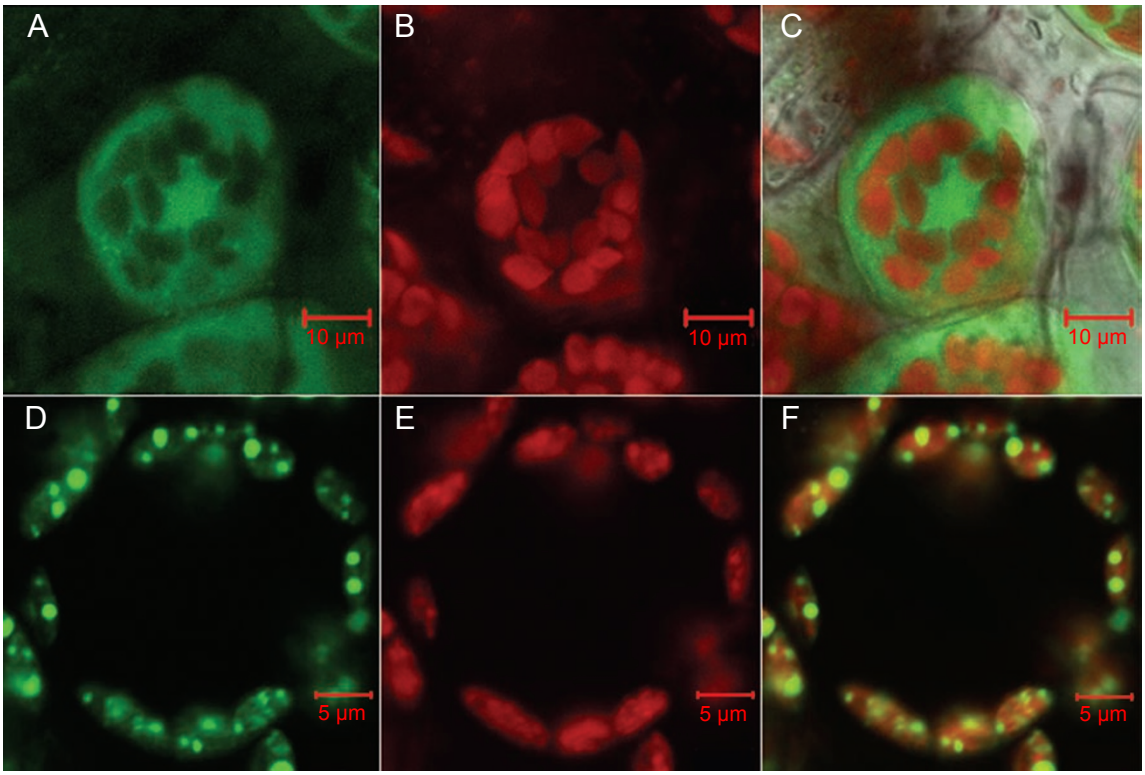

Figure 5 Localization of the p530-EGFP-52220 fusion protein to the chloroplast; **A**-**C** show the subcellular localization of 35S-EGFP fusion protein in transgenic *Arabidopsis* young leaf cells; **D**-**F** show the subcellular localization of p530-EGFP-52220 fusion protein in transgenic *Arabidopsis* young leaf cells; **(A** and **D)** green fluorescence; **(B** and **E)** chlorophyll autofluorescence; **(C** and **F)** overlapping of the green florescence and chlorophyll autofluorescence.

proteins cannot be predicted by TargetP. Thus, some of these 574 proteins should be true chloroplast proteins. But MS-based proteomics analysis cannot tell which of these proteins are true chloroplast proteins or contamination. In contrast, our core data suggest that 220 of them are reliable chloroplast proteins with 43 GSP proteins. Further analyzing the 43 GSP proteins with TargetP found only 4 of them predicted to be chloroplast proteins; this demonstrated that most of these 220 proteins cannot be recognized as chloroplast proteins by TargetP. In addition, our core dataset also contains 176 proteins that were neither included in PPDB data nor detected by proteomics. Of them, 70 proteins were found in the gold standard positives. Therefore, the extensive integration of chloroplast proteins in this work has helped to correctly identify more chloroplast proteins. The 1 720 chloroplast core nuclear proteins identified in this work are reliable and comprehensive. Furthermore, our analysis process provides a new effective approach to integrate data from different resources.

A chloroplast protein interaction network in Arabidopsis thaliana

Before the construction of chloroplast protein interaction network, a genome-scale protein interaction network was firstly built. To obtain a comprehensive *Arabidopsis* protein interaction network, we widely collected different data and predicted protein interactions by several methods including interologs [30], phylogenetics [41], biological process [42], gene co-expression [42] and enriched domain pairs [23]. Geisler-Lee *et al.* [27] reported a predicted *Arabidopsis* interactome with 19 979 interaction pairs of 3 617 proteins by the interolog method. In the protein interaction network of this work, 19 368 interaction pairs of 3 565 proteins were also predicted based on the interolog method. These proteins and their interaction pairs are included in Geisler-Lee *et al.*'s data [27]. This demonstrates the prediction data of Geisler-Lee *et al.* [27] are reliable. Each prediction method has its inherent bias, which has been demonstrated in previous study [41]. Therefore, in our protein interaction network we used six different methods, of which the interolog method is the only one to increase the prediction coverage for protein-protein interactions, and then used the Bayesian classifier

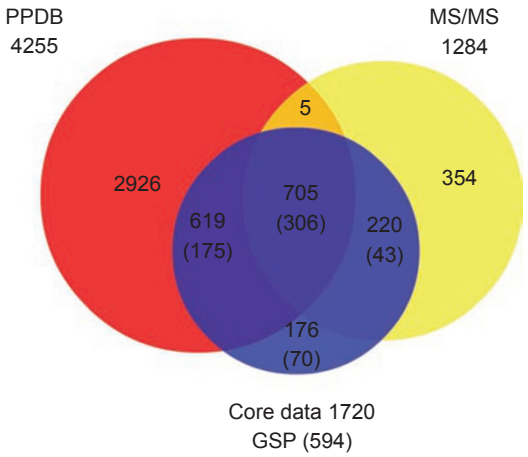

Figure 6 Comparisons of chloroplast protein data from PPDB, MS and the core data of this work. The PPDB data (4 255) were from the Cornel plastid proteome database; MS data (1 284) were integrated from available published data; the core data (1 720) were generated in this work. The number in the brackets indicates the gold positive data in different blocks.

to integrate those data.

The Bayesian classifier algorithm requires prior experience probability to determine the post probability. Therefore, the gold standard positives and negatives are prerequisite when performing Bayesian classifier. An effort was made to find experimentally validated chloroplast protein interactions that could be used as gold standard positives, and to find information that one chloroplast protein does not interact with another that could be used as GSN. Failure to find GSNs via extensive text mining led to an alternative strategy: we identified the chloroplast protein interactions based on genome-wide protein interactions. For GSNs, it is hard to pick up the protein pairs without interaction; we circumvented this problem following a previous paper [23] which has been widely used in the bioinformatics field. In the GSN, proteins with multiple locations have been discarded to reduce the potential false positive data.

Chloroplast protein interaction pairs were then screened with core proteins and core/candidate proteins from the genome-wide protein interaction data. Many of the chloroplast proteins had no annotation other than the gene chip information. A previous study indicated that a certain correlation exists between the gene expression profile and the protein functional linkage [56]. Therefore, protein interactions between core proteins and core/candidate proteins with a relative expression coefficient of more than 0.95 were also integrated into the network. This process extended the protein interaction data and provided more information for the study of protein functions.

Our network contains 2 214 proteins with 22 925 protein interaction pairs. Of these proteins, 1 043 are core proteins and the other 1 171 are candidate proteins. These candidate proteins were selected based on their predicted interactions with core chloroplast proteins, and we believe that the majority of them also belong to the chloroplast. As these candidate proteins cannot be integrated into the core data, and the core data contain 21 dual targeting proteins localized in both chloroplasts and mitochondria, certain nonchloroplast proteins might be mixed in these candidate proteins. These candidate proteins extend the number of identified chloroplast proteins to 2 979. This dataset should cover most of the chloroplast proteomes since previous research estimated that the chloroplast proteome of *Arabidopsis* contains 2 100-3 600 proteins [4]. Of the 22 925 interaction pairs, 15 295 interaction pairs were between two core proteins. The other 7 630 interaction pairs were comprised of core and candidate protein interactions. Therefore, although over half (52.9%) of the proteins in the network are candidate chloroplast proteins, the majority of the network is formed from core protein interactions. In order to minimize the false positive data, the interactions between candidate proteins were not included in the interaction network. Overall, we believe that the protein interaction network built here for the chloroplast is quite reliable.

In order to further assess the reliability of the network, the quality of the chloroplast protein interaction network was evaluated through the analysis of degree distribution and topological properties as described in previous research [23]. The degree of the functional linkages among the characterized proteins is much higher than that of the scrambled networks. This reveals that the network shows strong functional correlation between the interacting protein pairs. It is quite in agreement with the results of protein interactions from other organisms [21, 23]. The mean shortest path length between any two proteins is 4.07 links. This result indicates that most proteins are very closely linked, in accordance with a phenomenon described as a small-world property of networks [51] (Figure 2A). The degree distribution $P(k)$ of interactions per protein decays slowly, approximating a power law (Figure 2B). Meanwhile, the average clustering coefficient, $C(k)$, diminishes with an increase in the number of interactions per protein, implying that this network possesses a potential hierarchical organization (Figure 2C). All of these statistical analyses reflect that the network, in contrast to a scrambled network, is a biologically meaningful small-world network that displays scale-free hierarchical organization properties. This conclusion is in agreement with previous protein-protein interaction

network studies for model organisms [21, 23]. All of the data show that this network is reliable and provides a solid basis for interpretation of protein functions and interactions in the chloroplast.

Applications of the chloroplast protein interaction network

Functional annotation of unknown proteins The sequence of the *Arabidopsis* genome was reported, and its genes were annotated several years ago [57, 58]. However, the functions of about 30% of these gene products could not be assigned based on sequence homology and remain uncharacterized [59]. Protein interaction provides useful information of functional linkage between interacting partners, and has been used for functional annotation [49]. Among the 2 214 proteins in this chloroplast network, 307 are uncharacterized. Of them, 160 proteins were annotated in this work. This annotation approach could have certain limitations since it is heavily dependent on the accuracy and completeness of databases. Furthermore, this method could also miss some information on protein-specific functions [49, 60]. However, the annotated information for these proteins should provide useful hints for further functional analysis of these genes.

Photosynthetic complexes in the network The major function of the chloroplast is photosynthesis, and photosynthetic complexes play essential roles in this process. In the subnetwork extracted from the predicted protein interactions, most of the subunits of each complex grouped together (Figure 3A). The groupings indicated that those components for each complex are tightly linked in a manner that is consistent with their functional role in photosynthesis. Among the photosynthetic complexes, LHC I and LHC II distribute absorbed light energy to the cores of PSI and PSII for photosynthesis. It was discovered more than 30 years ago that plants can balance the excitation energy distribution between the two photosystems [61]. In the subnetwork extracted from predicted protein interactions between LHC II subunits and core proteins, LHC II functionally relates to LHC I, the PSI core and the PSII core (Figure 3B). These results captured the inter-relationships among the complexes. At the same time, many previously unknown functional linkages were detected among those complexes. These new interactions among or within photosynthetic complexes provide valuable information for future investigation.

The interactions between proteins can be classified into two major categories: physical and functional. The former refer to physical association between two proteins, whereas the latter refer to the proteins in a biochemical or a signaling pathway [18]. Indeed, most of these methods predict protein interactions that mean functional linkages, which include physical interactions and general involvement in the same biological process. We used this concept for protein interactions in this work. Based on the predicted protein interactions, we experimentally validated two physical interactions from 12 randomly selected pairs. Our network offers protein function clues and it also provides a platform or important resource for the research on chloroplast systems biology.

Materials and Methods

Candidate chloroplast proteins in Arabidopsis thaliana

A total of nine datasets of putative chloroplast proteins were obtained from various sources (Table 1). Datasets 1-4 were collected from the different databases including PPDB data [1], Predotar [46], Plprot [17] and GO (http://www.geneontology.org/). Datasets 5-9 were predicted in this work and briefly described as follows. Datasets 5 and 6 were generated based on BLASTP [62] searches using photosynthetic bacterial and other plant chloroplast proteins, respectively. Dataset 7 was from an analysis of microarray co-expression of genes in *Arabidopsis thaliana* leaves. Dataset 8 was predicted based on unique chloroplast Pfam domains (see Supplementary information, Data S1). In dataset 9, we included 1 205 nonredundant chloroplast proteins that were available from published *Arabidopsis* chloroplast proteomics data [1, 5-16]. Based on these nine datasets, we obtained a total of 7 592 nonredundant putative chloroplast proteins. Finally, we downloaded 88 chloroplast genome-encoded proteins from the TAIR database (http://www.arabidopsis.org). Detailed information of these datasets was included in the Supplementary information, Data S1.

The gold positive and negative data

A total of 594 nonredundant chloroplast proteins were categorized as the gold positives. The gold negatives were generated from the Swiss-Prot database (Release 49.7). After we excluded 'by similarity', 'potential', 'probable', and 'possible' entries and experimental data for individual proteins, we had 2 460 nonchloroplast proteins. Detailed information of these standard datasets is included in the Supplementary information, Data S1.

Integration of genome-scale datasets in a naïve Bayesian integration and identification of chloroplast core proteins

Before the naïve Bayesian integration, the independency of the nine datasets was evaluated by calculating the Pearson CCs between each comparison pair of features (Supplementary information, Table S7), and two sets of the dependent data (Plprot and MS, GO and PPDB) were regarded as one feature, respectively, based on the CC values. After the naïve Bayesian integration [47], 1 534 chloroplast core proteins were obtained with a 4.77% estimated false prediction rate (Supplementary information, Data S1). Under this threshold, the Bayesian classifier missed 135 of the 594 gold positive proteins. These missed proteins were also regarded as core proteins. To establish a set of reliable chloroplast proteins (core data), other proteins were integrated into the core data as

shown in Supplementary information, Figure S1. Currently a total of 1 284 nonredundant chloroplast proteins have been identified by proteomic technique. Of them 925 were integrated into the dataset of the 1 534 core proteins by Bayesian classifiers. We analyzed the remaining 359 proteins and found that 59 were previously identified by at least two different experiments. Our text mining of these 59 proteins revealed that 26 of them are not localized in the chloroplast. The remaining 33 proteins were included into the core data. Besides, our text mining also found 21 proteins localized in both the chloroplast and mitochondria. These proteins were also categorized as the core data of chloroplast proteins, but not as gold positive data. Of the 288 chloroplast proteins from Swiss-port, 233 were supported by our available published papers and were categorized as gold positive data. For two of the 288 proteins, the homologs cannot be found in *Arabidopsis*. The remaining 53 proteins were categorized as chloroplast core data, although no supporting references were found. In total, the core data of chloroplast proteins are 1 720. Considering the 88 chloroplast genome-encoded proteins, the set of reliable chloroplast proteins contains 1 808 nonredundant proteins.

Construction of a chloroplast protein interaction network for Arabidopsis thaliana

Protein interaction prediction Six computational methods were exploited to detect protein interactions: phylogenetic profiling [38, 63], interolog analysis [30], co-expression [42, 56], biological process (SSBP) [23], enriched domain pairs [23] and gene fusion [35]. In phylogenetic profiles method, we used the optimized parameters (LR_{pair} value L¦850) described in Sun *et al*. [63], and constructed and assessed profiles using the method described in Date and Marcotte [36]. To obtain ortholog interactions, we used the protein interaction data in four organisms (*Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Homo sapiens*), which were downloaded from the DIP database (http://dip.doembi.ucla.edu/dip/Download.cgi), and ortholog map files, which were downloaded from the Inparanoid database (http://inparanoid.cgb.ki.se/). *Arabidopsis thaliana* protein-protein interaction data were then obtained according to ortholog function transfer as described in Matthews *et al.* [30] Co-expressed genes were found by using publicly available microarray data from the TAIR database (ftp://ftp.arabidopsis.org), and microarray data from leaves were selected for further analysis. The Pearson correlations for each pair of genes were calculated between any two genes. Protein interaction pairs with a correlation value larger than 0.95 were extracted. A total of 23 493 nonredundant proteins in 1 389 GO annotations were downloaded from the Gene Ontology Consortium (http://www.geneontology.org). All proteins were paired, and the SSBP value of each pair was calculated. The gold standard positive interactions (4 902) involving 1 254 proteins were extracted from the published literatures, KEGG and IntAct database (see Supplementary information, Data S1 and Supplementary information, Table S8) for *Arabidopsis thaliana*. To find the enriched domains, the domains in these GSP protein interaction pairs were analyzed [23]. There were 5 337 pairs involving 438 proteins found at the genome level. The gene fusion method [35] was used. All the predictive data sources were integrated with a naïve Bayesian Networks approach as described in Rhodes *et al.* [42]. All the networks were drawn by software Pajek [64].

The GSP and GSN data The GSP of the protein interactions means that two proteins in a pair have true interaction relationship or functional linkages. The GSP interactions were collected from available published literatures, KEGG and IntAct database. We got 4 902 *Arabidopsis* protein interactions involving 1 254 proteins from these methods. Of them, 1 339 were chloroplast protein-protein interaction pairs.

The gold standard negative (GSN) of the protein interactions means that any physical interaction or functional relationship does not exit in the protein pairs. Due to few negative data available, we created the GSN in the following way. It was defined in this work as a protein pair with one located in the plasma membrane and the other located in the nucleus, as assigned by Gene Ontology Consortium (http://www.geneontology.org). In all, 19 GO IDs were assigned to the plasma membranes and 478 GO IDs were assigned to the nucleus. We mapped corresponding GO IDs to the *Arabidopsis thaliana* proteins, and created unique 364 161 negative protein interactions involving 6 461 proteins. No overlapped protein-protein pairs were found between the GSP and the GSN.

The Bayesian network approach To integrate different data, a naive Bayesian framework was adopted [47]. The essence of the approach is to provide a mathematical rule, given some predictive evidence, to explain how to adjust the odds that a pair of proteins interact, either in a true interaction instance (GSP) or, correspondingly, in negative protein interactions, known as GSN. The detailed approach is described in Supplementary information, Data S1.

Functional annotation

The Mapman functional annotation system of *Arabidopsis thaliana* from the PPDB database (http://ppdb.tc.cornel.edu) was used to annotate previously uncharacterized proteins with GOM described in Vazquez *et al.* [53].

Yeast two-hybrid analysis

Twelve pairs of protein interactions were validated by yeast two-hybrid experiments (Supplementary information, Table S6). The full-length cDNA of these proteins was obtained from RIKEN (Japan). The cDNAs were amplified using primers listed in Supplementary information, Table S9, and the PCR products were cloned into the pMD18-T vector (Takara, Japan). The full-length cDNAs in the T-vectors were then subcloned into the *Eco*RI and *Xho*I/*Sal*I sites of the pJG4-5 vector to construct baits. Also, full-length cDNAs in the T-vectors were subcloned into *Eco*RI and *Xho*I sites of pB42AD to construct preys. Combinations of prey and bait and the reporter vector pSH18-34 were co-transformed into the yeast strain EGY48 according to previously described procedures [65]. The selection of transformants and the analysis of the galactosidase were performed as described by McNellis *et al.* [66].

The subcellular localization of AT1G52220

The construct of the green fluorescence protein (GFP) fusion was followed as in Wang *et al*. [67]. The full-length coding sequence of AT1G52220 was fused in-frame with the pEGFP (Clon Tech, Palo Alto, CA, USA), with expression driven by the cauliflower mosaic virus 35S promoter using the pMON530 plasmid. The forward primer (5'ggt acc ATG GCT TCA ATT TCT GCA3'),

which has a *Kpn*I site, and the reverse primer (5'cca tgg GGC CAA GTA TAT CCG CTA C3'), which has an *Nco*I site, were used for PCR amplification and then the PCR products were inserted into the pEGFP plasmid. Next, the *BamH*I/*Eco*RI fragment from the inserted plasmid was inserted into the pMON530 to obtain p530-EGFP-52220. The *Eco*RI/*BamH*I double-digested pEGFP plasmid fragment and the *Eco*RI/*Bgl*II double-digested pMON530 plasmid vector were ligated to obtain 35S-EGFP. The p530-EGFP-52220 and 35S-EGFP (control) were introduced into the wild-type of Columbia plant using the floral dip method [68]. Stable transgenic plants were then obtained. Young leaves of transgenic plants were examined using a Zeiss LSM5 PASCAL confocal laser scanning microscope equipped with an Axiovert 100 inverted microscope. The filtersets used were BP505-545 (excitation, 488 nm; emission, 505-545 nm) and LP585 (excitation, 488 nm; emission, 585 nm) to detect GFP and the chlorophyll autofluorescence.

Acknowledgements

We thank the RIKEN BRC in Japan for provision of all full-length cDNA in this study. National Natural Science Foundation of China (grants numbers 30530100 and 90408010), the State Key Program of Basic Research of China (grant numbers 2007CB947600 and 2007CB108800), and Hi-Tech Research and Development Program of China (grant number 2006AA02Z313) supported this project.

References

1 Westerlund IvHG, Emanuelsson O. Lump – a neural network predictor for protein localization in the thylakoid lumen. *Protein Sci* 2003; **12**:2360-2366.
2 Friso G, Giacomelli L, Ytterberg AJ, *et al.* In-depth analysis of the thylakoid membrane proteome of *Arabidopsis thaliana* chloroplasts: new proteins, new functions, and a plastid proteome database. *Plant Cell* 2004; **16**:478-499.
3 van Wijk KJ. Plastid proteomics. *Plant Physiol Biochem* 2004; **42**:963-977.
4 Leister D. Chloroplast research in the genomic age. *Trends Genet* 2003; **19**:47-56.
5 Baginsky S, Gruissem W. Chloroplast proteomics: potentials and challenges. *J Exp Bot* 2004; **55**:1213-1220.
6 Ferro M, Salvi D, Brugiere S, *et al*. Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*. *Mol Cell Proteomics* 2003; **2**:325-345.
7 Ferro M, Salvi D, Riviere-Rolland H, *et al.* Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. *Proc Natl Acad Sci USA* 2002; **99**:11487-11492.
8 Froehlich JE, Wilkerson CG, Ray WK, *et al.* Proteomic study of the *Arabidopsis thaliana* chloroplastic envelope membrane utilizing alternatives to traditional two-dimensional electrophoresis. *J Proteome Res* 2003; **2**:413-425.
9 Giacomelli L, Rudella A, van Wijk KJ. High light response of the thylakoid proteome in *Arabidopsis* wild type and the ascorbate-deficient mutant vtc2-2. A comparative proteomics study. *Plant Physiol* 2006; **141**:685-701.
10 Kleffmann T, Russenberger D, von Zychlinski A, *et al.* The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. *Curr Biol* 2004; **14**:354-362.
11 Peltier JB, Cai Y, Sun Q, *et al.* The oligomeric stromal proteome of *Arabidopsis thaliana* chloroplasts. *Mol Cell Proteomics* 2006; **5**:114-133.
12 Peltier JB, Emanuelsson O, Kalume DE, *et al.* Central functions of the lumenal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction. *Plant Cell* 2002; **14**:211-236.
13 Peltier JB, Friso G, Kalume DE, *et al.* Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell* 2000; **12**:319-341.
14 Peltier JB, Ytterberg AJ, Sun Q, van Wijk KJ. New functions of the thylakoid membrane proteome of *Arabidopsis thaliana* revealed by a simple, fast, and versatile fractionation strategy. *J Biol Chem* 2004; **279**:49367-49383.
15 Schubert M, Petersson UA, Haas BJ, Funk C, Schroder WP, Kieselbach T. Proteome map of the chloroplast lumen of *Arabidopsis thaliana*. *J Biol Chem* 2002; **277**:8354-8365.
16 Ytterberg AJ, Peltier JB, van Wijk KJ. Protein profiling of plastoglobules in chloroplasts and chromoplasts. A surprising site for differential accumulation of metabolic enzymes. *Plant Physiol* 2006; **140**:984-997.
17 Kleffmann T, Hirsch-Hoffmann M, Gruissem W, Baginsky S. plprot: a comprehensive proteome database for different plastid types. *Plant Cell Physiol* 2006; **47**:432-436.
18 Yellaboina S, Goyal K, Mande SC. Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: comparison with high-throughput experimental data. *Genome Res* 2007; **17**:527-535.
19 Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 2001; **98**:4569-4574.
20 Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000; **403**:623-627.
21 Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* 2003; **302**:1727-1736.
22 Li S, Armstrong CM, Bertin N, *et al.* A Map of the Interactome Network of the Metazoan *C. elegans*. *Science* 2004; **303**:540-543.
23 Stelzl U, Worm U, Lalowski M, *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005; **122**:957-968.
24 Uhrig JF. Protein interaction networks in plants. *Planta* 2006; **224**:771-781.
25 de Folter S, Immink RG, Kieffer M, *et al.* Comprehensive interaction map of the *Arabidopsis* MADS Box transcription factors. *Plant Cell* 2005; **17**:1424-1433.
26 Hackbusch J, Richter K, Muller J, Salamini F, Uhrig JF. A central role of *Arabidopsis thaliana* ovate family proteins in networking and subcellular localization of 3-aa loop extension homeodomain proteins. *Proc Natl Acad Sci USA* 2005;

102:4908-4912.
27 Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M. A predicted interactome for *Arabidopsis*. *Plant Physiol* 2007; **145**:317-329.
28 Walhout AJ, Vidal M. Protein interaction maps for model organisms. *Nat Rev Mol Cell Biol* 2001; **2**:55-62.
29 Buckingham S. Bioinformatics: programmed for success. *Nature* 2003; **425**:209-215.
30 Matthews LR, Vaglio P, Reboul J, *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* 2001; **11**:2120-2126.
31 Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* 2001; **29**:482-486.
32 Grigoriev A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res* 2001; **29**:3513-3519.
33 Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998; **23**:324-328.
34 Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 1999; **96**:2896-2901.
35 Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999; **402**:86-90.
36 Date SV, Marcotte EM. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* 2003; 21:1055-1062.
37 Huynen M, Snel B, Lathe W, 3rd, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 2000; **10**:1204-1210.
38 Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999; **96**:4285-4288.
39 Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes. *Nucleic Acids Res* 2001; **29**:1216-1221.
40 Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D. Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. *Genome Biol* 2003; **4**:R59.
41 Sun J, Sun Y, Ding G, *et al*. InPrePPI: an integrated evaluation method based on genomic context for predicting protein-protein interactions in prokaryotic genomes. BMC Bioinform 2007; **8**:414.
42 Rhodes DR, Tomlins SA, Varambally S, *et al*. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol* 2005; **23**:951-959.
43 Jansen R, Yu H, Greenbaum D, *et al*. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003; **302**:449-453.
44 Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. *Science* 2004; **306**:1555-1558.
45 Troyanskaya O, Cantor M, Sherlock G, *et al*. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; **17**:520-525.
46 Small I, Peeters N, Legeai F, Lurin C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 2004; **4**:1581-1590.
47 Calvo S, Jain M, Xie X, *et al*. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* 2006; **38**:576-582.
48 Cui J, Li P, Li G, *et al*. AtPID: *Arabidopsis thaliana* protein interactome database an integrative platform for plant systems biology. *Nucleic Acids Res* 2008; **36**:D999-D1008.
49 Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nat Biotechnol* 2000; **18**:1257-1261.
50 Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004; **5**:101-113.
51 Strogatz SH. Exploring complex networks. *Nature* 2001; **410**:268-276.
52 Yook SH, Oltvai ZN, Barabasi AL. Functional and topological characterization of protein interaction networks. *Proteomics* 2004; **4**:928-942.
53 Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* 2003; **21**:697-700.
54 Khrouchtchova A, Hansson M, Paakkarinen V, *et al*. A previously found thylakoid membrane protein of 14 kDa (TMP14) is a novel subunit of plant photosystem I and is designated PSI-P. *FEBS Lett* 2005; **579**:4808-4812.
55 Petsalaki EI, Bagos PG, Litou ZI, Hamodrakas SJ. PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genom Proteom Bioinform* 2006; **4**:48-55.
56 Yu H, Luscombe NM, Lu HX, *et al*. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* 2004; **14**:1107-1118.
57 The *Arabidopsis* Genomics Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000; **408**:796-815.
58 Frishman D, Albermann K, Hani J, *et al*. Functional and structural genomics using PEDANT. *Bioinformatics* 2001; **17**:44-57.
59 Berardini TZ, Mundodi S, Reiser L, *et al*. Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol* 2004; **135**:745-755.
60 Karaoz U, Murali TM, Letovsky S, *et al*. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci USA* 2004; **101**:2888-2893.
61 Bonaventura CJM. Fluorescence and oxygen evolution form Chlorella pyrenoidosa. *Biochim Biophys Acta* 1969; **189**:366-383.
62 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **21**:403-410.
63 Sun J, Xu J, Liu Z, *et al*. Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics* 2005; **21**:3409-3415.
64 Batagelj V, Mrvar A. Pajek – Analysis and visualization of large networks. In: Junger M, Mutzel P, eds. Graph drawing software. Berlin: Springer Press, 2003:77-103.
65 Chen DC, Yang BC, Kuo TT. One-step transformation of yeast in stationary phase. *Curr Genet* 1992; **21**:83-84.
66 McNellis TW, Torii KU, Deng XW. Expression of an N-

terminal fragment of COP1 confers a dominant-negative effect on light-regulated seedling development in *Arabidopsis*. *Plant Cell* 1996; **8**:1491-1503.

67 Wang PC, Wang C, Mi HL, Zhou GY, Yang ZN. Study of subcellular localization of the expressed protein in *Arabidopsis thaliana*. *Chin Bull Bot* 2006; **23**:249-254.

68 Clough SJ, Bent AF. Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J* 1998; **16**:735-743.

(**Supplementary Information** is linked to the online version of the paper on the Cell Research website.)