Curr Genet (2006) 49: 106–124
DOI 10.1007/s00294-005-0041-2

RESEARCH ARTICLE

Stephan Eberhard · Monica Jain · Chung Soon Im
Steve Pollock · Jeff Shrager · Yuan Lin
Andrew S. Peek · Arthur R. Grossman

Generation of an oligonucleotide array for analysis of gene expression in *Chlamydomonas reinhardtii*

Received: 3 October 2005 / Revised: 24 October 2005 / Accepted: 25 October 2005 / Published online: 7 December 2005
© Springer-Verlag 2005

Abstract The availability of genome sequences makes it possible to develop microarrays that can be used for profiling gene expression over developmental time, as organisms respond to environmental challenges, and for comparison between wild-type and mutant strains under various conditions. The desired characteristics of microarrays (intense signals, hybridization specificity and extensive coverage of the transcriptome) were not fully met by the previous *Chlamydomonas reinhardtii* microarray: probes derived from cDNA sequences (~300 bp) were prone to some nonspecific cross-hybridization and coverage of the transcriptome was only ~20%. The near completion of the *C. reinhardtii* nuclear genome sequence and the availability of extensive cDNA information have made it feasible to improve upon these aspects. After developing a protocol for selecting a high-quality unigene set representing all known expressed sequences, oligonucleotides were designed and a microarray with ~10,000 unique array elements (~70 bp) covering 87% of the known transcriptome was developed. This microarray will enable researchers to generate a global view of gene expression in *C. reinhardtii*. Furthermore, the detailed description of the protocol for selecting a unigene set and the design of oligonucleotides may be of interest for laboratories interested in developing microarrays for organisms whose genome sequences are not yet completed (but are nearing completion).

Keywords Chlamydomonas · Gene expression · Oligo-array · Genomics

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00294-005-0041-2 and is accessible for authorized users.

Communicated by D. Stern

S. Eberhard · M. Jain · C. S. Im · S. Pollock · J. Shrager
A. R. Grossman
Department of Plant Biology, The Carnegie Institution,
260 Panama Street, Stanford, CA 94305, USA

S. Eberhard (✉)
Laboratoire de Physiologie Moléculaire et Membranaire
du Chloroplaste, Institut de Biologie Physico-Chimique,
UMR7141 (CNRS – Université Paris VI),
13 Rue Pierre et Marie Curie, 75005 Paris, France
E-mail: Stephan.Eberhard@ibpc.fr
Tel.: +1-331-58415050

Y. Lin · A. S. Peek
Integrated DNA Technologies, 1710 Commercial Park,
Coralville, IA 52241, USA

Introduction

Chlamydomonas reinhardtii is a unicellular alga that has been used extensively as a model organism to study photosynthetic function and the biogenesis of the chloroplast (Harris 2001; Rochaix 2002, 2004; Wostrikoff et al. 2004), the structure and function of flagella and basal bodies (Silflow and Lefebvre 2001; Kamiya 2002; Dutcher 2003; Scholey 2003), nutrient deprivation and stress-related processes (Davies et al. 1996; Wykoff et al. 1999; Ledford et al. 2004; Miura et al. 2004; Zhang et al. 2004), photoperception (Huang et al. 2002; Sineshchekov et al. 2002; Nagel et al. 2003; Kateriya et al. 2004) and circadian control (Werner 2002; Mittag and Wagner 2003; Wagner et al. 2004; Mittag et al. 2005). Recently, extensive cDNA and genomic sequence information has become available, with approximately 90% of the nuclear genome being sequenced (JGI, unpublished data). This advance has made it possible to identify families of genes (Stauber et al. 2003; Elrad and Grossman 2004) and subsets of genes encoding proteins potentially involved in specific biological processes or that function in specific metabolic pathways (LaFontaine et al. 2002; Grossman et al. 2004; Lohr et al. 2005). A genomic analysis has identified many genes encoding polypeptide components of the basal body and the flagella and has demonstrated that some of these genes have counterparts in the human genome which, when mutated, result in human diseases such as obesity (Snell et al. 2004), Bardet-Biedl syndrome (Li et al. 2004), polycystic kidney

disease (Pazour et al. 2000; Qin et al. 2001; Pazour 2004), and primary cilia diskinesis (Omran et al. 2000). Furthermore, there are many molecular and genetic techniques that make *C. reinhardtii* an attractive organism for genomic analyses (Dutcher 2000; Grossman 2000; Dent et al. 2001; Grossman et al. 2003) and for the generation and use of high density microarrays. Recently, cDNA-based microarrays and macroarrays were generated and used to study changing gene expression that accompanies the transfer of cells from low to high light conditions (Im and Grossman 2002), high to low inorganic carbon concentrations (Miura et al. 2004), and during phosphorus (Moseley and Grossman 2005) and sulfur (Zhang et al. 2004) deprivation.

The limitations associated with the previous version of the *C. reinhardtii* array (v1.1) (Zhang et al. 2004) are low coverage of the nuclear transcriptome and the potential for nonspecific hybridization to the long cDNA sequences used for array construction. To provide the community of researchers working with *C. reinhardtii* with a powerful tool for analyses of gene expression, we have developed a new microarray (v2.0), with representation for approximately 10,000 genes, based on synthetic oligonucleotides. The use of oligonucleotide technology for microarray production has been widely used during the last several years for the study of gene expression in various model organisms (reviewed in Barrett and Kawasaki 2003; Stears et al. 2003; Park et al. 2004; He et al. 2005; Stoughton 2005). Long oligonucleotide arrays offer several advantages over cDNA-based arrays. First, probes can be designed entirely in silico, by taking advantage of available genome and EST sequence information, thus bypassing long and complex cloning, amplification and sequence verification procedures used when generating cDNA-based array elements. Second, because the probes are designed using accurate bioinformatic programs, they can be optimized for probe specificity, hybridization characteristics and homogeneity (e.g., T_M) with respect to the entire probe population. Tailored, long oligonucleotide-based arrays yield high specificity with respect to signal intensities and reproducibility, comparable to that obtained using PCR- or cDNA-based arrays (Kane et al. 2000; Stears et al. 2003; He et al. 2005). Many companies are now producing and distributing long-oligonucleotide arrays for analyses of gene expression in various model organisms. One drawback of commercial available platforms, discussed in several recent manuscripts (see Discussion), is the often poor correlation between results obtained using different microarray platforms and the fact that the information used to generate the arrays (e.g., sequences of oligonucleotide array elements, gene models used to develop the oligonucleotide sequences) are often not made publicly available.

In addition to the goal of providing the community of researchers working with *C. reinhardtii* with the best possible microarray, we also wanted to detail all steps involved in building the array, providing the end-user with the information that would enable them to critically design experiments and interpret their results. The protocols developed for microarray generation are based on exhaustive use of both cDNA and genomic information to define a unique set of *C. reinhardtii* genes; array generation was not solely based on gene model generation by ab initio protocols (frequently used for oligonucleotide design and microarray development). The protocols that we used to design the oligonucleotides emphasize the importance of oligonucleotide quality (analyzed by mass spectrometry), the specificity of the oligonucleotides for the target sequences, the intensity of the signals generated as a consequence of hybridization and the reproducibility of the results. Detailed information presented in this manuscript on the selection of the unigene set, probe design criteria and the characteristics of each of the designed oligonucleotides (the nucleotide sequence, match to gene models, annotation of each associated gene) provides researchers with (1) a thorough description of the new array (information not always available for commercially designed platforms), (2) up-to-date annotation of genes on the array, and (3) a discussion of protocols for oligonucleotide generation from genomic and cDNA information, as well as (4) the caveats associated with these protocols.

Results

Sequence resources used to generate a unigene set for oligonucleotide design

Before producing a high-density oligonucleotide-based microarray, it is critical to identify specific sequences that represent unique genes. Different sequence resources were combined to produce a high-quality unigene set, representing the known *C. reinhardtii* transcriptome.

Use of cDNA information

We used assembled cDNA sequences, supported by genomic information, as well as sequence information from previously characterized genes. Most of the cDNA sequences used for establishing the unique gene set are also represented by gene models generated by the Joint Genome Institute (JGI, Walnut Creek, CA, USA); these gene models were placed onto DNA scaffolds comprising the genome and are shown as a track on the *C. reinhardtii* genome browser (http://www.genome.jgi-psf.org/cgi-bin/browserLoad/41a8f1a03110b11773a7e-deb). The information used to establish a set of unique cDNAs for *C. reinhardtii* is given below:

1. Sequences present in *C. reinhardtii* recombinant libraries (Asamizu et al. 2000; Shrager et al. 2003) (http://www.kazusa.or.jp/en/plant/chlamy/EST/); these libraries were generated from cells grown under a broad range of environmental conditions.

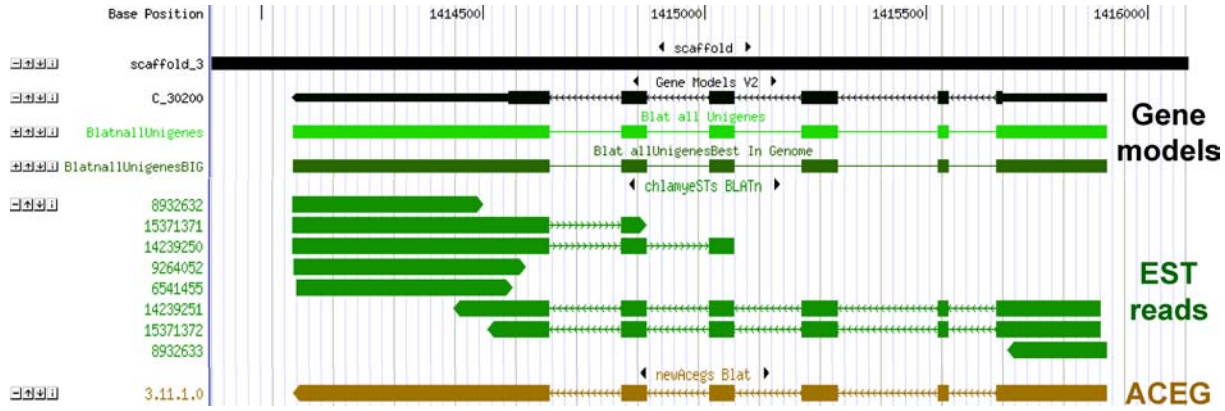

Fig. 1 *Chlamydomonas* genome browser developed by the Joint Genome Institutes. The example shown depicts a region with one predicted gene model with associated EST reads (ChlamyeSTs track) and a corresponding ACEG that is composed of a single contig (newACEGs track). The first number of the ACEG indicates the scaffold (3), the second designates the number of the ACEG on that scaffold (11), the third designates the number of contigs in the ACEG (1) and the fourth represents the origin of the sequence

2. Specific cDNA sequences generated by various researchers, although some were not present in the EST/cDNA libraries that were used for the generation of sequence information (e.g., low expression genes). Some of these sequences are in public databases (e.g., the EMBL database), while others may still not have been entered into the public databases.
3. The cDNA information was used to capture corresponding genomic information to improve the quality of the final assembled sequences.

Use of sequence information from the nuclear genome

Shotgun sequencing of the nuclear genome of *C. reinhardtii* was performed at approximate 10X coverage by JGI. The assembly was constructed with JAZZ, the JGI assembler, which exploited paired-end sequence information (5′ and 3′ sequences from each clone). After trimming vector and low quality sequences, 1.8 million reads were assembled into 3,211 scaffolds that together represent ~100 Mbp. Roughly half of the genome is contained on 72 scaffolds, all of which are at least 504 kbp. The genomic sequence information represents approximately 90% of the total *C. reinhardtii* nuclear genome.

Combining cDNA and nuclear information

There are currently approximately 180,000 cDNA sequence reads that have been generated by the *C. reinhardtii* genome consortium (Grossman et al. 2003; Shrager et al. 2003), as well as numerous other sequences either reported by the Kasuza sequence group (Asamizu et al. 1999, 2000), or by individual laboratories that have focused on specific genes. Shrager et al. (2003) placed the reads into distinct contigs (an assemblage of reads with overlapping nucleotide sequences), and contigs that group together as part of the same genes have been designated Assembly of Contigs generated from EST information (ACEs). Reads that do not assemble into contigs, but that are considered to come from the same gene (in the same ACE), represent sequences from the 3′ and 5′ ends of a given clone. All of the reads of a given ACE have also been mapped to the *C. reinhardtii* nuclear genome and the cDNA and its corresponding genomic sequence were reassembled; the resulting assemblage is called an ACE containing genomic Ghost sequences (ACEG). Building ACEGs provides a higher confidence level with respect to individual sequence calls within an ACE. Each ACEG has been placed onto the genome and has an identifier that starts with the scaffold number followed by the ACEG number; ACEG 5.8 is the eighth ACEG located on scaffold number five of the genomic sequence, as presented on the JGI browser (http://www.genome.jgi-psf.org/chlre2/chlre2.home.html). The assembly of ACEGs will be discussed in detail by J. Jain et al. (unpublished data). The JGI browser page displays different tracks, including the EST reads that associate with specific regions of the genome (ChlamySTs BLATn on the genome browser; see Fig. 1), and the gene models (finalModels V2) that predict the coding region of a gene (although they do not take into account cDNA and ACE sequence information). Figure 1 gives an example of a typical genome browser window on the JGI website. This example shows a region of scaffold 3 for which expression data is available (represented by the EST, ChlamyESTs BLATn tracks), and a corresponding gene model has been generated by the JGI using ab initio methods. Supplemental Table 1 lists all the ACEGs that have a corresponding gene model. In cases where an ACEG shares similarity with more than one gene model, all corresponding gene models are indicated. This might be the case for gene families, or when the gene model information is redundant.

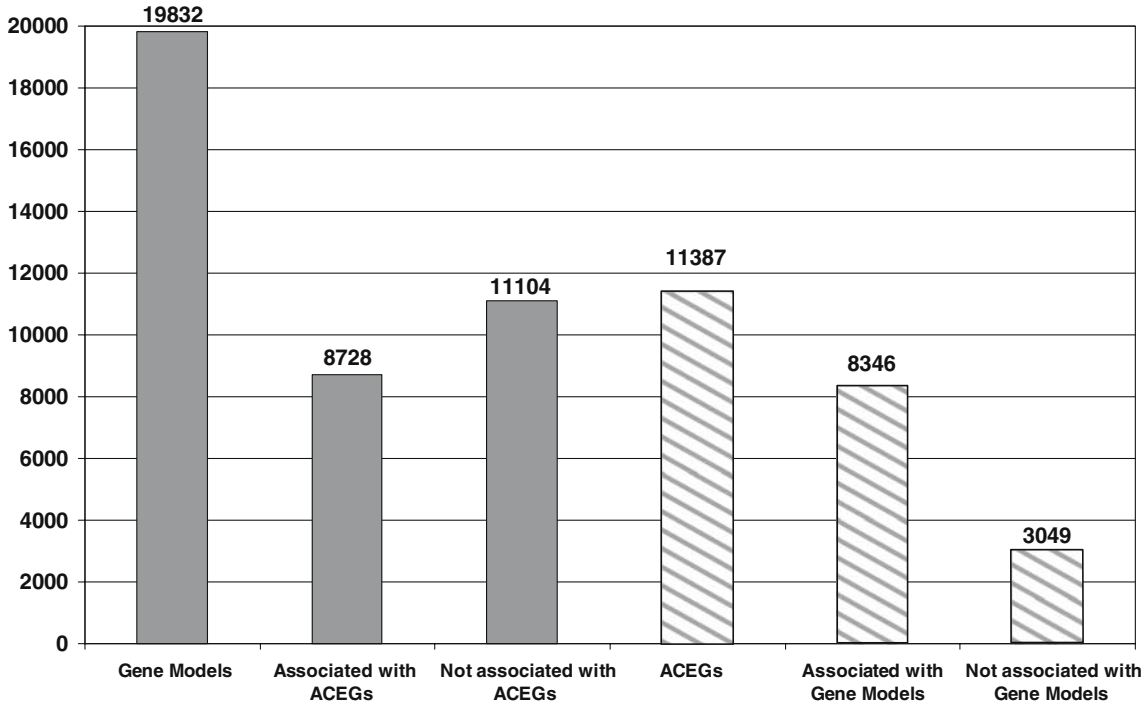

Fig. 2 Relationships between gene models and ACEGs. The *bar graph* depicts the total number of gene models predicted by the Joint Genome Institute, the number of gene models associated/not associated with ACEGs, the total number of ACEGs, and the number of ACEGs associated/not associated with gene models

Currently, there are nearly 20,000 gene models, of which 8,728 are associated with ACEGs, as shown in Fig. 2. The number of gene models is an overestimate of the number of *C. reinhardtii* genes since many of the smaller scaffolds are contained within the larger scaffolds, but were not assembled because of poor sequence information. There are also a number of cases for which expression data was deduced based on cDNA representation within the libraries, but the cDNA information was not associated with a gene model. As shown in Fig. 2, approximately 3,000 of the 11,387 ACEGs are not represented by a gene model. Furthermore, many of the gene models are inaccurate in their prediction of intron–exon boundaries. Lastly, gene models appear to be wrong in the 3′ region in a significant number of cases, based on comparisons with the cDNA information. This can be problematic because, as discussed later, the 3′ regions of transcripts are preferred for the generation of oligonucleotides used on the array. For these reasons, we decided to primarily exploit cDNA sequence information for designing the oligonucleotides used for array construction. It should be noted that several of the gene models predicted by JGI are not represented by cDNA sequence information. The gene models not associated with expression data may represent genes expressed at a low level or under specific conditions that are not represented by the suite of conditions to which the *C. reinhardtii* cells used for cDNA production were exposed. They may also represent silent pseudogenes originating from duplications of expressed genes. In generating the array probes, we did not use gene model information that lacked corresponding cDNA confirmation, but intend to exploit this information for future oligonucleotide array design.

ACEG composition

The number of contigs and corresponding EST sequences within the different ACEGs could be markedly different. In the current assembly, 56% of the ACEGs are composed of a single contig comprised of numerous sequences derived from both 5′ and 3′ cDNA reads (Fig. 3Ia). In 44% of the cases, an ACEG is composed of two or more contigs (Fig. 3Ib–f). ACEGs composed of multiple contigs fall into a number of different categories. An ACEG contains a nonoverlapping set of contigs when the sequences at the 3′ and 5′ ends of the clones have not been extended enough to generate a single overlapping sequence (Fig. 3Ib). For these ACEGs, the regions between contigs are not represented by cDNA sequence information, although they may be represented by genomic sequence information. Activating the "+" button near the 'newACEG' track on the JGI Browser for *C. reinhardtii* expands that track and shows the individual contigs that comprise a specific ACEG.

There are also many cases in which an ACEG may be composed of contigs that overlap each other, but the overlap was not detected by the Phrap assembly program either because the region of overlap was short

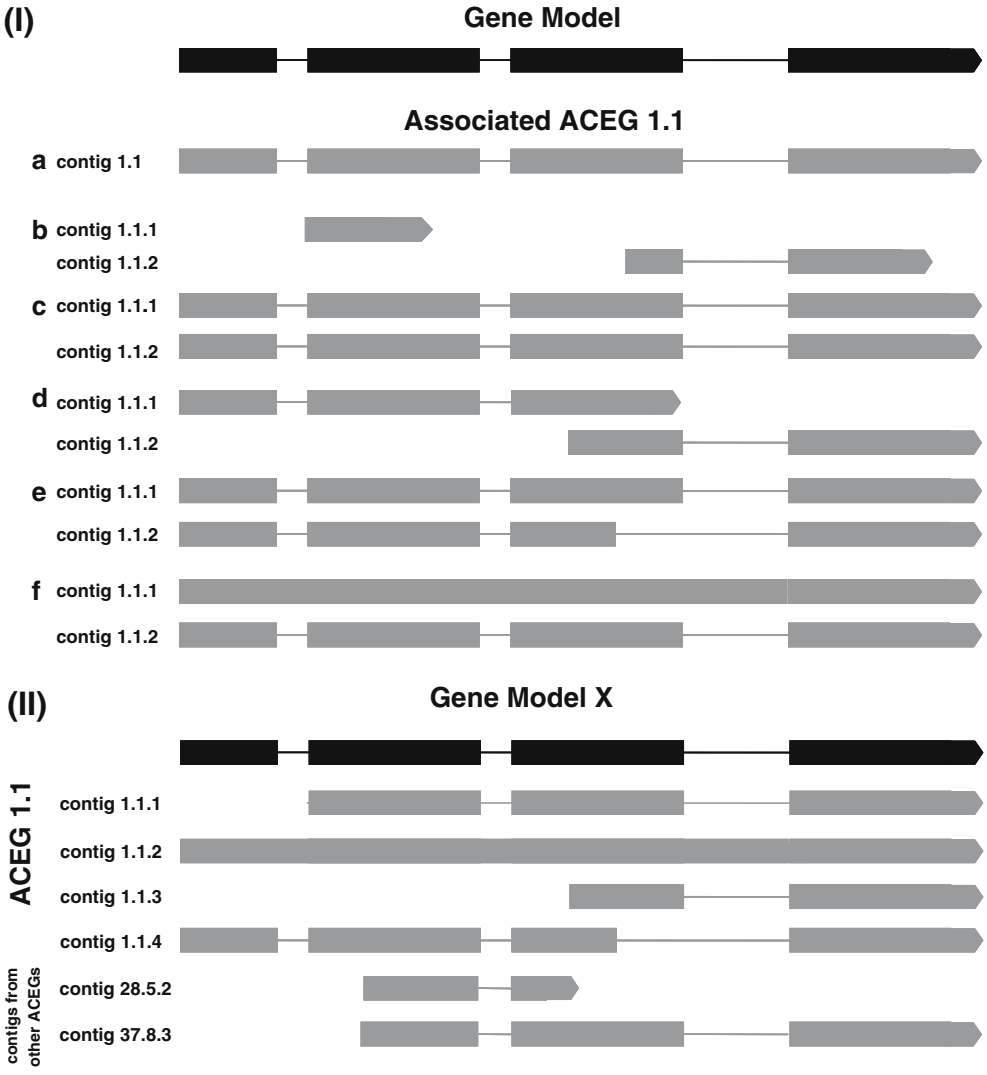

Fig. 3 Potential ACEG composition and similarities among ACEGs. *I* (*a*) An ACEG containing a single contig. (*b*) An ACEG composed of two nonoverlapping contigs. (*c*) An ACEG composed of two fully redundant contigs. (*d*) An ACEG composed of two contigs with a short overlap at their ends. (*e*) An ACEG composed of two contigs with alternative intron–exon junctions. (*f*) An ACEG containing a contig with all intronic sequences (probably representing genomic DNA contamination). *II* An example of an ACEG with complex structure. In this case, the theoretic ACEG 1.1 has four overlapping contigs and also shares sequence similarity with contigs 28.5.2 (from ACEG 28.5) and 37.8.3 (from ACEG 37.8). Such complex ACEG structures were indeed observed for a significant number of cases

(Fig. 3Id), or the overlapping sequences were extensive (Fig. 3Ic) but of low quality. We were able to detect the relatively short and/or low quality regions of overlap when individual reads from the ACEGs were blasted against the total read population.

Some ACEGs are composed of reads within contigs that differ in their intron–exon structure, as shown in Fig 3Ie. These reads may represent alternative splice variants of the same mRNA, or in some cases, they may result from contaminating nuclear DNA in the RNA samples that were used for construction of the cDNA library. The latter situation is often obvious since a read, represented by a separate contig, contains all of the intronic sequences (Fig. 3If)

Finally, some contigs may share sequence similarity with contigs belonging to different ACEGs. This can occur when the genes represented by these contigs are part of a gene family that share conserved domains. In other instances, the contigs may actually represent the same gene and its presence in more than one ACEG is a result of improper assembly of the reads. Figure 3II shows a theoretical example of complex ACEG structure. ACEG 1.1 is composed of two overlapping contigs that correspond with the predicted gene model (contigs 1.1.1 and 1.1.3), a third contig (contig 1.1.2) that contains intronic sequences, probably representing contaminating genomic DNA, and a fourth contig (contig 1.1.4) that reveals a possible splice variant. Finally, this group contains contigs that belong to different ACEGs (28.5 and 37.8); these contigs have sequences similar to contigs that comprise ACEG 1.1. Indeed, there are a number of cases in which ACEGs have contigs with different intron–exon information, and also cases of similarity among contigs within different ACEGs.

The breakdown of the number of contigs present in the total ACEG population for the most current cDNA assembly is presented in Table 1 and Supplemental Fig. 1. Only those ACEGs with four contigs or less were included in the subsequent generation of gene-specific oligonucleotides (~70 mers) used for construction of the version 2.0 microarray. In cases for which there were 2–4 contigs in a single ACEG, the contig with the highest EST support was chosen to represent this gene for the generation of a gene-specific 70 mer; the rationale for

Table 1 Composition and size distribution of ACEGs

Complete set		Contig composition of ACEGs					Size	
ACEGs	Contigs	1	2	3	4	**> 4 contigs**	> 200 bp	**< 200 bp**
11,387	19,038	6,334	3,453	1,038	378	**184**	11,192	**195**

Composition and size distribution of the complete set of ACEGs. Bold values show ACEGs that were discarded before analysis, i.e., those composed of more than 4 contigs (184), or for which each contig has a length < 200 bp (195). According to this filter, 11008 ACEGs were selected for further analysis

using this criterion was that those contigs with the highest EST support should contain the most accurate sequence data. Furthermore, for this version of the *C. reinhardtii* array, we did not attempt to generate oligonucleotides that would distinguish alternative splice variants, and again, when the EST information suggested the occurrence of splice variants, the sequence of contigs with the highest EST support were chosen for generating array elements. The contig composition of 100 ACEGs, based on manual examination, are given in Fig. 4.

Selecting unigenes for oligonucleotide generation

General considerations

A major reason for generating a set of unique ACEGs was to establish a unigene set to serve as the basis for synthesizing oligonucleotides for building a high-density microarray. This ACEG population would represent a nonredundant set of cDNAs, each ACEG representing a unique gene. Two issues must be considered when choosing a set of sequences for oligonucleotide design:

(a) For ACEGs composed of more than one contig, which contig should be used for the generation of the oligonucleotide?
(b) When a contig within an ACEG matches a contig from another ACEG, how do we differentiate between distinct genes with significant sequence similarity and redundant sequence information (contigs that represent the same gene but that assembled into two or more ACEGs)?

As mentioned previously, with respect to situation (a) we chose to use the contig with the highest EST support. Using the contig with the highest EST support eliminates a number of potential problems and abnormalities associated with particular contigs. These problems/abnormalities include the erroneous assembly of contigs into an ACEG (e.g., short sequences with significant similarity), and the presence of contaminating genomic DNA in the cDNA population (reflected by the presence of a few sequences within an ACEG that retain introns). Manual analysis of 100 ACEGs has shown that in all cases, the contig with the highest EST support does not retain intron sequences (generally there are a few intron-containing sequences and their presence is most apparent in ACEGs that are composed of numerous EST reads). In cases where two nonassembled contigs within an ACEG completely overlap (e.g., unspliced intron-containing DNA prevented assembly), or where they overlapped over a short sequence span (not high enough sequence quality and/or long enough sequence overlap to allow assembly), the contig with the highest EST support was used for designing the oligonucleotides. In cases where two or more contigs of an ACEG shared the same number of supporting EST reads, the most 3′ contig was

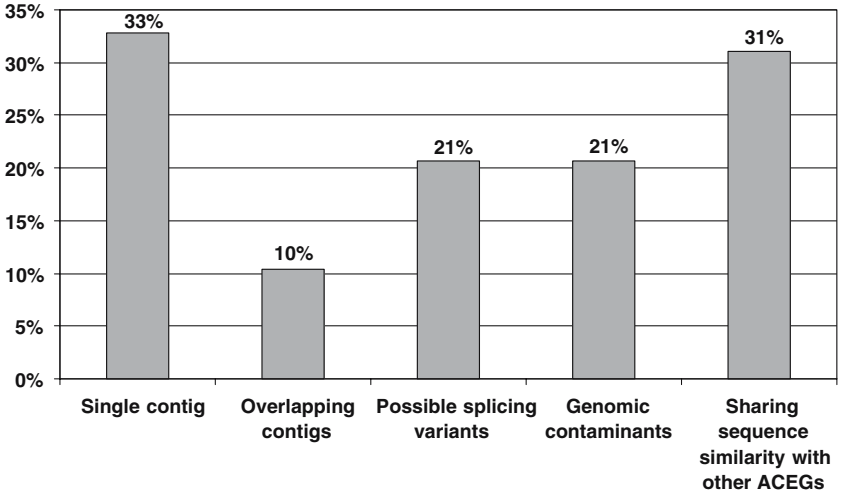

Fig. 4 Manual analysis of 100 ACEGs and their classification into different categories, as described in the text and on the *x*-axis of the *bar graph*. The sum is greater than 100% because some of the ACEGs are composed of multiple contigs falling under different categories

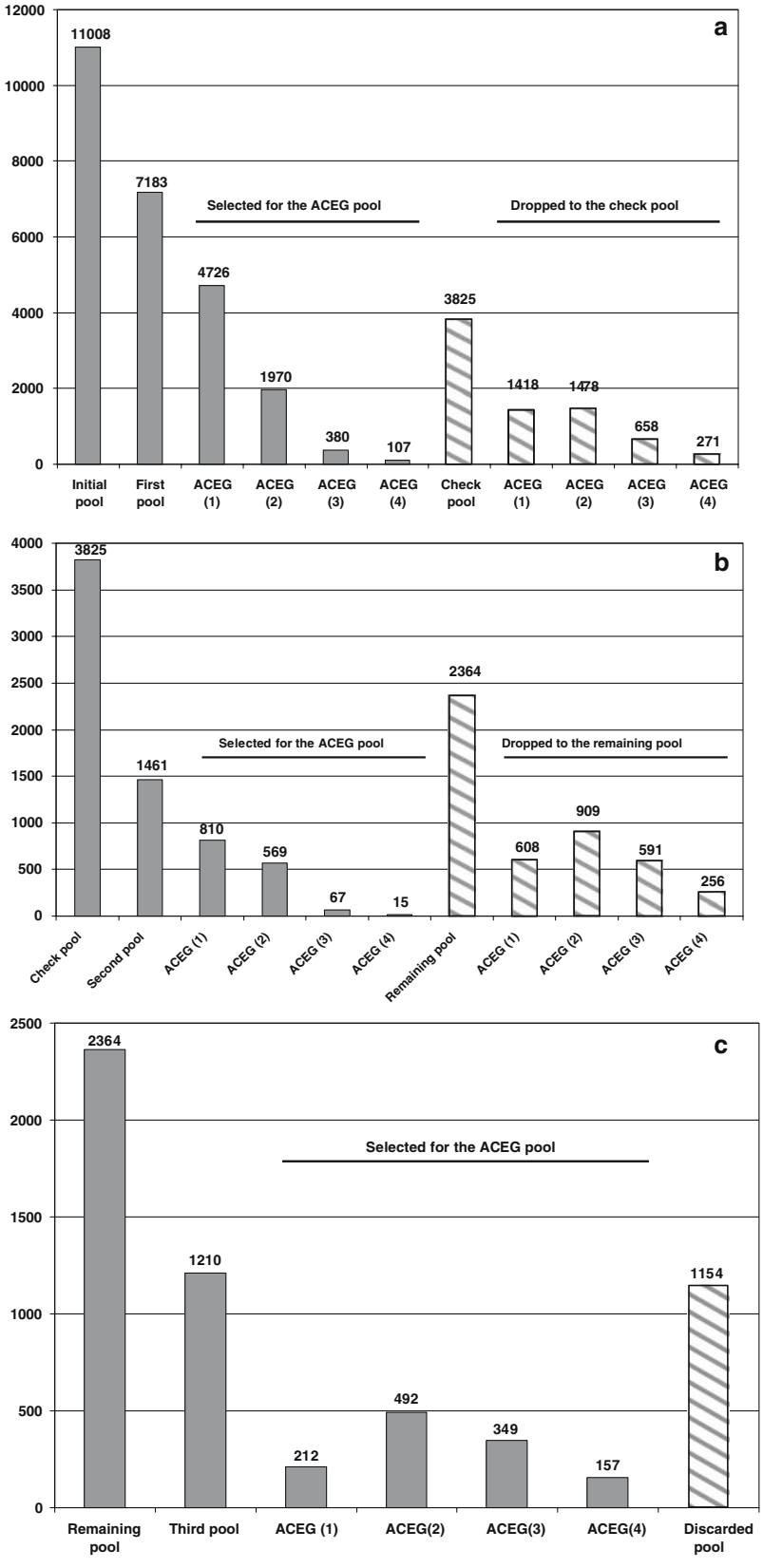

Fig. 5 Composition of the different ACEG pools that served as the database for the sequential generation of the unigene set used for the design of the oligonucleotides. The criteria used to generate the first, second and third ACEG pool (the *second bar* in each of the graphs **a**, **b** and **c**) and to place ACEGs into check-pools are described in the text. *Numbers in brackets* (on *x*-axis) indicate the number of contigs present in the ACEGs

selected for oligonucleotide generation. In some cases, overlapping contigs showed retention of introns that could represent splice variants. However, we did not attempt to generate 'variant-specific' oligonucleotides (again, we used the contig with the greatest EST support).

Table 2 Probe design criteria

	T_M criteria min, opt, max	Total target sequences	Probes satisfying all criteria	Probes with possible cross-hybridization	No probes designed
First round of design (T_M 75°C)	72, 75, 78	10,035	8,562	846	627
	78, 78, 84	1,473	725	711	37
	Combined T_M	10,035	9,287	711	37
Second round of design (T_M 78°C)	75, 78, 81	10,035	8,961	880	194
	81, 81, 87	1,074	230	782	62
	Combined T_M	10,035	9,191	782	62
Final	Combination of first and second rounds	10,035	9,368	630	37

Oligonucleotides designed during the first round (optimum T_M of 75°C) and the second round (optimum T_M of 78°C) were combined to yield the final set of oligonucleotides. Of the 9998 oligonucleotides selected for the generation of the array, 9,368 satisfied all of the criteria defined in the text. 630 oligonucleotides may show some cross-hybridization with nontarget sequences, and for 37 sequences no satisfying oligonucleotide could be designed

For those contigs in category (b), we developed a protocol to generate the most reliable set of unique sequences, reducing ACEG redundancy to a minimum. This protocol sacrificed some potentially unique sequence information, but yielded a higher degree of confidence in the unique nature of each of the selected ACEGs. The use of this nonredundant ACEG pool for oligonucleotide generation eliminates duplicate gene representation on the array, reducing the cost and increasing the overall quality of oligonucleotide design.

Protocol for ACEG selection

The protocol to select the set of ACEGs used for oligonucleotide design was divided into three phases, yielding three separate ACEG pools that are combined into a *final ACEG pool*. The database used for initiating the protocol consisted of the 11,387 ACEGs (Table 1) generated by the assembly protocol that will be described by M. Jain et al. (unpublished data) and which was briefly discussed above. The ACEGs (composed of 19,038 contigs) were placed into the 'complete ACEG pool'. As discussed above, each ACEG can be composed of one or more contig(s). All ACEGs containing more than four contigs (184 in total) were eliminated from the dataset since sequence information in the contigs that comprise these ACEGs is likely to be of low quality and hence they limit the assembly process. ACEGs composed of contigs shorter than 200 bp (195 in total) were also discarded. Eliminating ACEGs with >4 contigs and with all contigs shorter than 200 bp yielded a population of 11,008 ACEGs (Table 1 and Supplemental Table 2); 6,144 of these were *single* ACEGs composed of only one contig, while 4,864 were *multi* ACEGs composed of 2, 3 or 4 contigs.

To initiate the protocol of eliminating duplicate contigs, the consensus sequences of all contigs from the 11,008 ACEGs were blasted against themselves and a contig was not considered unique if its sequence matched another consensus contig sequence with an *E*-value of $1e^{-15}$ or lower. This enabled us to create three contig lists. One list contained the *unique* contigs, i.e., those contigs that did not match any other contig in the complete pool. A second list had *multi-own* contigs, i.e., contigs that match only contigs from the same ACEG. The latter represent overlapping contigs not assembled by Phrap assembly program. A third list contained *multi-other* contigs that were similar to contigs in other ACEGs.

Phase 1: filtering of the ACEGs In *phase 1*, contigs of *single ACEGs* that were present in the unique contig list were directly placed into the *final contig pool*. Of the 6,144 *single* ACEGs, the contigs from 4,726 were partitioned into the *final contig pool*. The remaining 1,418 *single* ACEGs were placed into the *check-pool*; the ACEGs in this pool were analyzed in *phase 2* of the protocol. Second, we selected *multi-ACEGs* (ACEGs with 2–4 contigs) for which *all* of their component contigs were present in the unique contig list (nonoverlapping contigs not represented by sequences in any other ACEG in the database). From these ACEGs, we chose the single contig from each with the highest EST sequence support and used that contig consensus sequence for the design of the oligonucleotide array elements. For cases in which the contigs had essentially identical EST support, we chose the contig that contained the polyA tail or that was closest to the 3′ end of the cDNA. Contigs representing these ACEGs were

Table 3 Origin and labeling scheme of the 9998 final pool sequences

Origin	Suffix	Sequences	Numbers
EMBL	.A	497	1.A to 497.A
Private	.B	38	498.B to 535.B
First pool	.C	6,974	536.C to 7509.C
Second pool	.D	1,411	7510.C to 8919.D
Third pool	.E	1,078	8920.E to 9998.E

Origin of the sequences used to generate the oligonucleotides in the final pool. Oligonucleotides were numbered from 1 to 9998 with a suffix (.A through .E) tracing their origin into the final pool of 9998 sequences

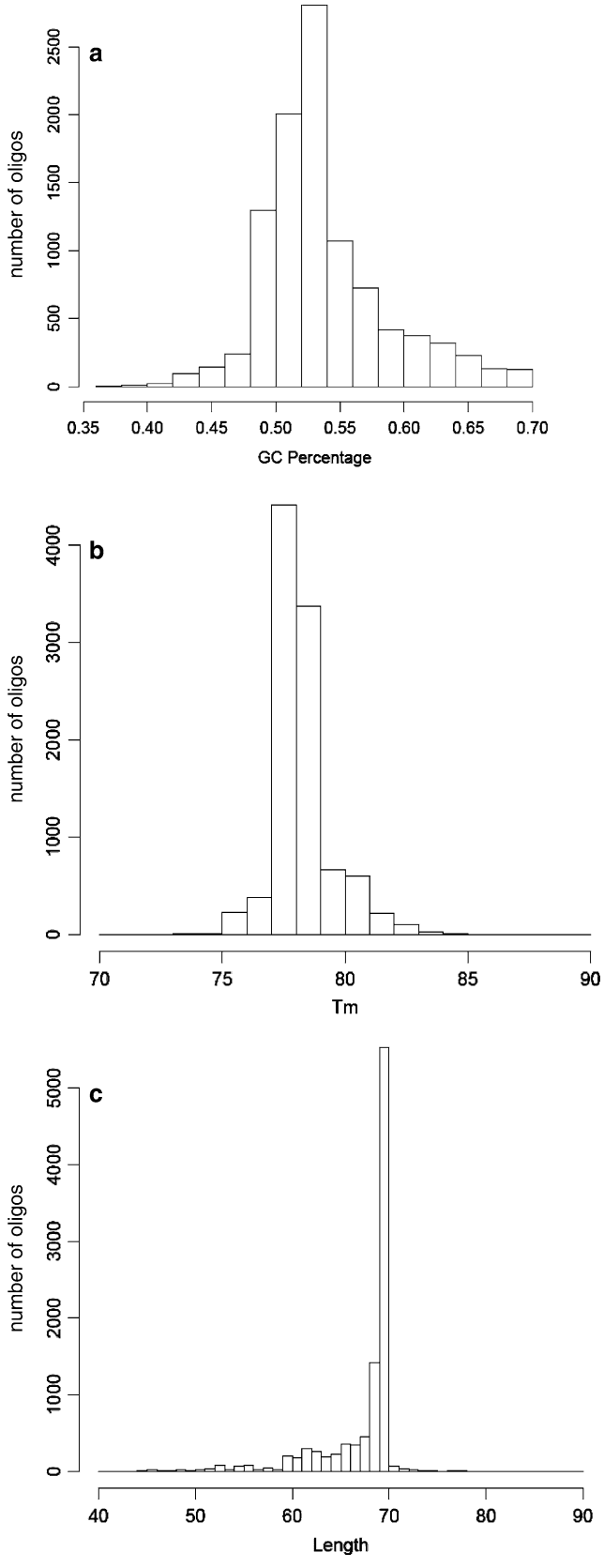

Fig. 6 The distribution of GC content (**a**), TM (**b**) and length (**c**) associated with the 9,998 specific oligonucleotides designed as elements for the microarray

added to the *final contig pool*. This analysis yielded an additional 1,640 unique contigs from the total of 4,864; 1,512 were from ACEGs composed of two contigs, 117 were from ACEGs composed of three contigs and 11 were from ACEGs composed of four contigs.

We also examined the *multi ACEGs* that had all their contigs in the multi-own contig list; these contigs matched other contigs, but only those present in the same ACEG. In all cases, the contigs within the single ACEG could be assembled (into a single contig), but they may have not been assembled by Phrap for several reasons; the overlap was too short, the sequence was not of a quality good enough, or the contigs had differences in their intron–exon structures. For these cases, we extracted the contig of over 200 bp with the highest EST support, and in cases where two contigs had the same EST support, we chose the one with the polyA tail or the one closest to the 3′ end, and placed it into the *final contig pool*. This analysis provided an additional 817 contigs (458 from ACEGs composed of two contigs, 263 from ACEGs composed of three contigs and 96 from ACEGs composed of four contigs). The remaining 2,407 *multi ACEGs* were placed into the check-pool.

At this point, the *final contig pool* had a total number of 4,726 + 1,640 + 817 = 7,183 contigs, or 4,726 contigs from ACEGs with 1 contig, 1,970 contigs from ACEGs with 2 contigs, 380 contigs from ACEGs with 3 contigs and 107 contigs from ACEGs with 4 contigs. Each of these selected contigs should represent a unique gene. The *check-pool* contained a total of 7,432 contigs derived from 3,825 ACEGs; the composition of the *check-pool* at this point was 1,418 single ACEGs, 1,478 ACEGs with two contigs, 658 ACEGs with three contigs and 271 ACEGs with four contigs (Fig. 5a and Supplemental Table 2).

Phase 2: analyzing the check-pool with new similarity criteria The *check-pool* contained all ACEGs composed of one or more contigs, of which at least one contig matched another contig from a different ACEG with an *E*-value of $1e^{-15}$ or less. The identification of similar contigs within different ACEGs might be a consequence of the existence of gene families and marked sequence similarities between the different members of those families, or result from the generation of duplicate ACEGs erroneously generated by the Phrap assembly program. In the latter case, poor sequence reads and/or alternative intron splicing may have resulted in the separation of sequences that really represent a single gene. In this second phase of ACEG selection, we attempted to eliminate duplicate ACEGs from the check-pool using relatively conservative criteria that would likely also eliminate some unique ACEGs. First, we established new similarity criteria and blasted all the contig sequences in the check-pool against each other. For two contigs to be considered duplicates, a contig of one ACEG must match a contig from at least one other ACEG with 90% identity over a span of a

nucleotide sequence consisting of at least 200 bp. Sequences meeting or exceeding these criteria thresholds were considered duplicates.

Within these new threshold limits, a number of the ACEGs in the check-pool were considered to be "unique" (e.g., none of the contigs within the ACEG matched a contig from another ACEG). This protocol generated an additional 810 contigs from *single ACEGs* and 651 contigs from the *multi-ACEGs*, or contigs representing 1,461 new ACEGs in all. Again, the contig with the highest EST count for each selected ACEG was included in the *final contig pool*. There were 2,364 ACEGs that did not pass this second filter (608 *single ACEGs* and 1,756 *multi ACEGs*) and were dropped to the *remaining pool*; all of these ACEGs had at least one contig that matched a contig from another ACEG, according to the new similarity criteria (Fig. 5b).

Phase 3: analyzing the remaining pool For the third part of the protocol, we blasted the contigs comprising the 2,364 ACEGs in the *remaining pool* against each other using a threshold value for 'match' of $1e^{-15}$. We then formed separate groups of ACEGs for which at least one contig of one ACEG matched another contig from a second ACEG with 90% or greater identity over a span of at least 200 bp. While some of these 'ACEG groups' contained two different ACEGs, others contained several ACEGs. Furthermore, a number of ACEGs were comprised of contigs that were present in more than one of the ACEG groups (contained more than one contig that matched contigs in other ACEGs; so a number of ACEGs were members of different ACEG groups). To select potentially unique, high quality ACEGs from this pool, we calculated the total EST support for each ACEG in the pool by adding the EST support for all of the contigs from which it was composed, sorted the ACEG groups according to EST support (those with highest support were placed at the top of the list and those with the lowest at the bottom) and then processed the ACEG groups from high to low support. The processing consisted of selecting the longest, single ACEG from its group of matching ACEGs while discarding the other ACEGs from that group (placing them into the *discarded pool*), and then identifying the contig within the selected ACEG with the highest EST support. When a specific ACEG was eliminated from one ACEG group, it was automatically eliminated from any of the other ACEG groups in which it was a member. This procedure yielded contigs representing another 1,210 ACEGs that were placed in the *final contig pool*; 1,154 ACEGs were discarded using this procedure (Fig. 5c). The final results of this selection protocol yielded 9,854 sequences in the *final contig pool* (Supplemental Table 2). To verify that all of the sequences in this pool were unique, we blasted them against each other. Based on the blast outputs, we eliminated 5 additional sequences because of insufficient sequence quality, leaving a *final contig pool* containing 9,849 unique contigs.

Sequences from the EMBL database

We also included in the population of sequences used for oligonucleotide generation those *C. reinhardtii* sequences present in the EMBL database. This dataset included 331 sequences based on genomic DNA information and 433 sequences based on cDNA information. We checked for duplicates among the two sets of sequence information and when duplicates were discovered, the longest sequence was retained. We considered two sequences duplicates if they were ≥90% identical over a region of at least 200 bp. This analysis yielded 505 EMBL sequences to be added to the *final contig pool*. If an EMBL sequence was not represented in the *final contig pool*, it was immediately added to that pool. If the sequence was already presented in the *final contig pool*, the contig generated from the cDNA assembly was removed, and it was replaced by the EMBL sequence. The rationale for performing this exchange was that in general, EMBL sequences were generated by extensive sequence analyses performed by individual investigators focused on a specific gene; therefore, the accuracy of the EMBL sequences would generally be better than consensus contig sequences generated by assembly of EST information. Also the lengths of the EMBL sequences were usually longer than the corresponding contig chosen from the ACEG pool. Of the 505 sequences, 195 were not already present in the *final contig pool* and 310 were used to replace 356 existing contigs. The replacement of the contigs with the EMBL sequences was not a one-for-one replacement because one EMBL sequence could replace one or more contigs if the similarities are high enough. After adding the 195 EMBL sequences and replacing the 356 contigs with the 310 EMBL sequences, the total number of sequences in the final contig pool was 9,998.

Contributed, unpublished sequencing efforts

Unpublished sequence information was sent by individual investigators and compared to the ACEG or EMBL databases for duplications, based on previously described criteria (≥90% identity over ≥200 bp). Sequences not present in either of the databases were added to the final pool. If the sequence was already represented in the final pool (either from ACEGs or EMBL sequences), the contig generated from the assembly was removed and was replaced by the contributed sequence. This process yielded another 38 sequences that were added to the final pool (only one of the sequences was represented by an EMBL sequence).

In the end, the final pool contained 9,493 ACEG contigs, 38 private sequences, 504 EMBL sequences or a total of 10,035 sequences representing unique genes. All these sequences were used to generate unique oligonucleotide array elements. A powerpoint file containing a more visual depiction of the sequence selection protocol used to generate the *final ACEG pool* can be downloaded from http://www.chlamy.org/micro.html.

Generating oligonucleotides from sequences in the final contig pool

The design of oligonucleotides from the sequences contained in the final sequence pool was performed using the array-oligonucleotide generation protocol developed by Integrated DNA Technologies, in which some parameters were customized. The design process is detailed below.

Introduction to probe design

Since microarray experiments involve numerous DNA/DNA or DNA/RNA hybridizations, the design criteria for all factors affecting hybridization, such as T_M (melting temperature), GC%, secondary structure, self-hybridization of the probe (self-dimerization) and sequence complexity (stretches of A, C, G or T) must be considered with respect to oligonucleotide design. Also, since thousands of different microarray hybridization events occur simultaneously on the same array platform, the T_M for the individual reactions must be approximately the same for all hybridization events; this helps equalize signals across the array. Furthermore, probe specificity must be considered in order to prevent cross-hybridization of the labeled cDNAs among different array elements. The labeled DNA used for hybridization with the array elements were generated from a reverse transcriptase reaction primed with oligo-dT (which binds to the poly-adenylated 3′ end of a mature mRNA). Usually 3′ sequences of transcripts are most effectively reverse transcribed in this reaction as a consequence of premature termination of the reverse-transcription reactions, leading to truncated cDNAs. Therefore, biasing the selection of array elements toward those localized to the 3′ region of each mRNA is likely to improve hybridization signals.

Probe design criteria for C. reinhardtii

Probe design based on the ACEG sequences used an iterative strategy because it was not possible to design a unique probe for each sequence in the 10 K sequence pool according to a single stringent set of criteria. Therefore, we adjusted or relaxed specific criteria to generate probes for sequences that did not contain suitable probe regions according to the original probe design criteria. Since the T_M value is important for maximizing both sensitivity and specificity of the binding of the probe to the cDNA, this criterion was fixed during probe selection and could only be adjusted by reinitiating the protocol for probe selection. For other criteria like GC%, free energy of probe self-dimer hybridization and probe length, the algorithm would automatically and gradually relax the criteria if no probe candidate met the initial criteria. We implemented the probe design protocol several times to generate a high quality probe set. Each time we changed the T_M criterion, the other criteria remained unchanged, or the algorithm performed parameter relaxation if necessary. The parameters used by our protocol are given below:

Minimum T_M The minimum T_M threshold temperature for each probe on the microarray. Only probe candidates that met or exceeded the T_M threshold value were considered.

Maximum T_M The maximum T_M threshold temperature for each probe on the microarray. Only probe candidates that met or fell below the T_M threshold value were considered.

Optimum T_M The optimum T_M for each probe on the microarray. The program selects probes with a T_M as close to this value as possible.

Maximum probe length Eighty-five nucleotides (first round) and 80 nucleotides (second round). This defined the maximum length of the probe selected by the algorithm. If no probe was identified that met the original criterion, the maximum probe length would be increased by as many as ten bases.

Minimum probe length Fifty-five nucleotides (first round) and 60 nucleotides (second round). This defined the minimum length of the probe selected by the algorithm. If no probe was identified that met the original criterion, the minimum probe length would be decreased by as many as ten bases.

Optimum probe length (70 nucleotides) This defined the optimum length of the probe. The algorithm selected probes as close to this length as possible by applying a linear penalty score to probe candidates that deviated from this length. The greater the deviation from optimal length, the greater the scoring penalty to candidate probes.

Distance from 3′ end (1,000 nucleotides) This set a limit to the search for suitable probes proximal to the 3′ end of the target sequence. The algorithm attempted to select probes from within the distance region defined by this variable.

GC percent minimum (30%) This set the minimum GC% threshold for each probe. If no probe was selected for a sequence using the original criteria, the algorithm may select a probe with a lower GC%, but in no case would it select a probe with lower than 20% GC content.

GC percent maximum (70%) This set the maximum GC% threshold for each probe. If no probe was selected for a sequence using the original criteria, the algorithm may select a probe with a higher GC%, but in no case would it select a probe with a higher than 80% GC content.

Number of probes for each target sequence (1) This parameter determined the number of probes selected for each target sequence.

Secondary structure minimum free energy The algorithm calculated the predicted secondary structure of each probe candidate, eliminating candidate probes with highly stable secondary structures, as evaluated by free energy calculations of Gibbs. The threshold value was set at -5 kcal mol^{-1} and candidate probes with a free energy for secondary structure of less than -5 kcal mol^{-1} were rejected.

Self-dimer free energy Some candidate probes were likely to form self-dimer structures. The algorithm rejected probes that form stable self-dimer structures, as measured by Gibbs free energy, with a threshold value of -12 kcal mol^{-1}. If no probe was selected using the original criteria, the algorithm may select probes that form more stable self-dimer structures than specified, although probes with a self-dimer free energy of less than -20 kcal mol^{-1} would always be rejected.

DNA concentration A concentration of 0.2 mM. The probe concentration used to calculate the T_M and the formation of potential secondary structures.

Sodium concentration A concentration of 0.1 M. The sodium ion concentration used to calculate the T_M and the formation of potential secondary structures.

Probe sequence complexity check All the probes were examined for complexity, and probes with regions of homopolymeric runs of single bases, or for which more than half consisted of a single base, were rejected.

Cross-hybridization check Checking for probe cross-hybridization to nontarget sequences was the most time-consuming part of the probe design. All probes were checked for cross-hybridization against a library composed of 19,580 sequences, which is a combination of the complete pool of 19,038 contigs and 542 sequences from EMBL or private sources. We used slightly more stringent criteria than those previously described (Kane et al. 2000); any probe candidates (1) with a region of 15 consecutive bases that exactly matched a region not belonging to itself or to another contig from the same ACEG; (2) or where the whole probe exhibited 70% or more sequence similarity with any other contig sequence, were considered as potentially cross-hybridizing probes and the algorithm avoided selection of such probes. If no noncross-hybridizing probe was identified, the probe with the least potential for cross-hybridization to nontarget sequences would be identified; the program also identifies those nontarget sequences that would potentially cross-hybridize.

Probe design process

We performed two rounds of probe design. The first round used the probe design criteria set to the *Minimum* T_M 72°C, *Maximum* T_M 78°C, and *Optimum* T_M 75°C, and the values for the remaining criteria were set as previously described. The first round of probe design resulted in the generation of unique target probes that met either the original or reduced stringency criteria for 8,562 sequences, with an additional 846 sequences for which designed probes were predicted to cross-hybridize with at least one nontarget sequence. No probes were selected for 627 sequences with the first round probe design criteria (Table 2, top).

Increasing the T_M criterion to a *Minimum* T_M of 78°C, *Maximum* T_M of 84°C, *Optimum* T_M of 78°C and rerunning the probe design program for 1,473 (846 + 627) target sequences for which we were not able to generate specific oligonucleotides that met the original or reduced stringency criteria, yielded an additional 725 suitable probes. Furthermore, the number of probes for which there was potential cross-hybridization was reduced to 711, and there were only 37 target sequences for which no satisfactory probe was identified.

The overall T_M criteria used for the first round of probe selection appeared to be a little low and the algorithm generated many shorter length oligonucleotide probes; most probes generated had lengths between 55 and 60 nucleotides. These short probes might result in reduced signal to noise output from the array relative to arrays constructed from elements of 70 nucleotides long (He et al. 2005). Therefore, we performed a second round of probe design for the entire set of target sequences in which we changed the T_M criteria to: *Minimum* T_M 75°C, *Maximum* T_M 81°C, *Optimum* T_M 78°C. Furthermore, we narrowed the probe length criteria: *Maximum probe length* 80, *Minimum probe length* 60, *Optimum probe length* 70. These new criteria prevented the large length variation observed in the first round of probe design.

The results of the second round of probe design are summarized in Table 2. There were 8,961 probes that met the original or reduced stringency criteria, 880 probes that did not meet cross-hybridization criteria, and 194 target sequences for which no satisfactory probes were designed. Just as in the first round of oligonucleotide design, we increased the T_M criteria to a *Minimum* T_M 81°C, *Maximum* T_M 87°C, *Optimum* T_M 81°C and reran the design protocol for the 1,074 target sequences that did not yield a satisfactory probe with the initial T_M criterion. These allowed us to generate an additional 230 probes that satisfied either the original or reduced stringency criteria. Furthermore, more than half of the probes generated by this protocol were 70 nucleotides in length.

Although the second round of probe design generated a greater number of probes that better satisfied the length requirement (lengths of close to 70 nucleotides),

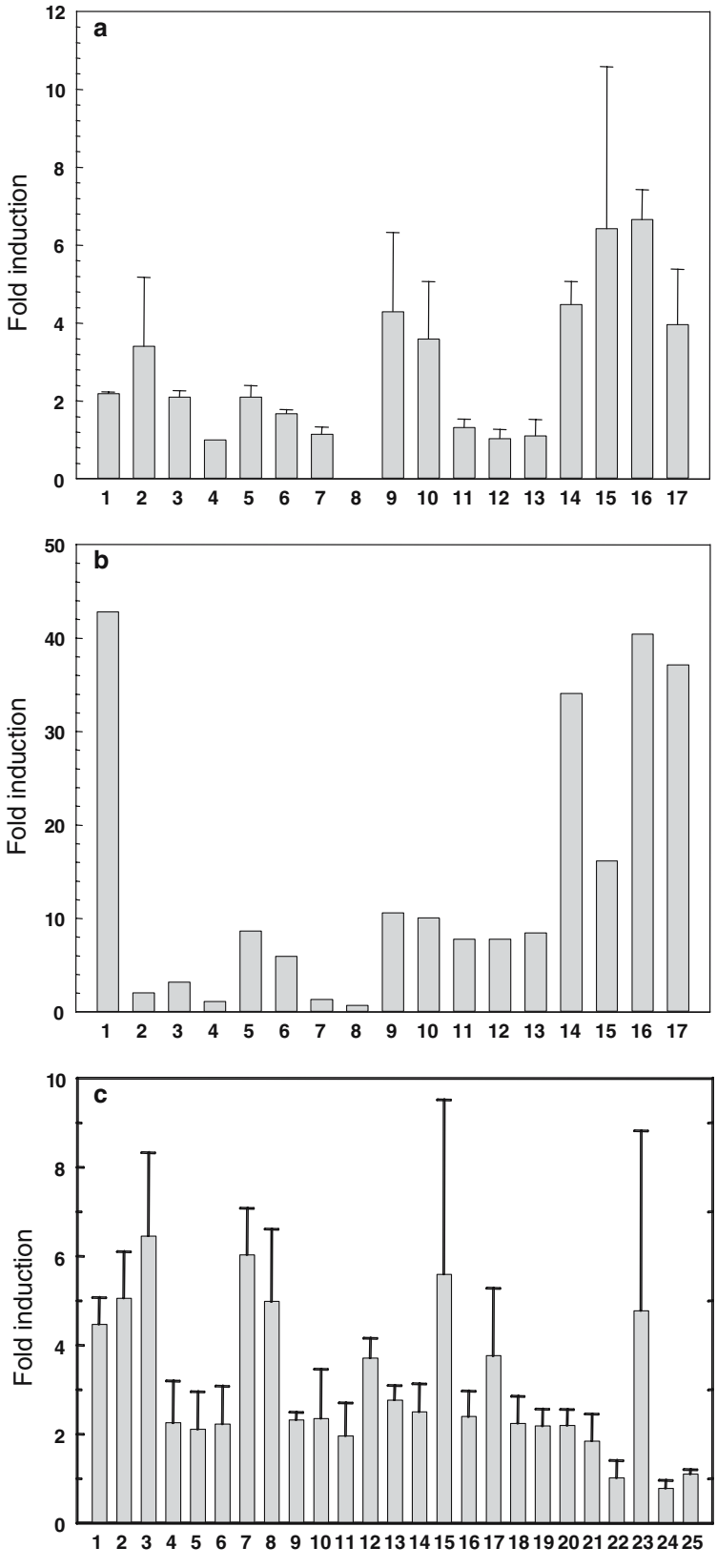

Fig. 7 A comparison of the fold increase in the levels of transcripts encoding proteins associated with sulfur acquisition and assimilation between the oligonucleotide array (**a**) and the cDNA array (**b**). The transcripts examined encode *1* arylsulfatase, ARS1, *2* sulfite reductase, SIR1, *3* sulfite reductase, SIR2, *4* sulfite reductase (bacterial type), SIR3, *5* ATP sulfurylase, ATS1, *6* ATP sulfurylase, ATS2, *7* adenylyl sulfate reductase, APR, *8* APS kinase, *9* serine acetyltransferase, SAT1, *10* O-acetylserine(thio)lyase, OSATL4, *11* cysteine dioxygenase, *12* UDP-SQ synthase, SQD, *13* UDP-SQ diacylglycerol, SQ, *14* ECP76, *15* selenium-binding protein, SBDP, *16* and *17* two SAC1-like proteins. The *error bars* represent the standard deviation. **c** Newly identified transcripts that increase in response to sulfur deprivation based on array analyses. These transcipts encode *1* ECP88, *2* serine/threonine kinase, *3* ARS2, *4* methionyl-tRNA transferase, *5* Tbc2 translation factor, *6* glycoside hydrolase, *7* hypothetic protein rich in serine, *8* similar to vanadium chloroperoxidase, *9* CyCAB1 A/B-type cyclin related protein, *10* serine/threonine kinase, *11* RNA-binding region RNP-1, *12* GAS28 hydroxyproline-rich glycoprotein, *13* unknown, *14* unknown, *15* unknown, *16* unknown, *17* unknown, *18* unknown, *19* unknown, *20* unknown, *21* unknown, *22* sulfate transporter SULTR1, *23* sulfate transporter SULTR2, *24* SbpA, sulfate transport system substrate-binding protein, *25* chloroplast sulfate transport system permease. The *error bars* represent the standard deviation

the first round yielded a greater number of satisfactory probes. Some sequences with satisfactory probes generated by the first set of criteria for probe design were not associated with satisfactory probes following application of the second set of criteria. To generate the largest population of target sequences having satisfactory

probes, satisfactory probes generated during the first round of probe design that did not yield a satisfactory probe during the second round were included in the final oligonucleotide probe population. These merged results (Table 2) show that of the total 9,998 probes generated, 9,368 satisfied all criteria (original or low stringency) while 630 were predicted to show nonspecific hybridization to some nontarget sequences; there were only 37 sequences for which no satisfactory probe could be designed. The distribution of the GC content, T_M, and length of the designed probes are shown in Fig. 6a–c. The probes were numbered from 1 to 9,998 with an added suffix that defines their origin. "A" for probes designed for EMBL sequences, "B" for probes designed for private sequences, "C" for probes designed for contigs from ACEGs in the first pool, "D" for probes designed for contigs from ACEGs in the second pool, "E" for probes designed for contigs from ACEGs in the third pool (Table 3). Finally an "*" was added to the 630 probes that are predicted to show some level of cross-hybridization. Supplemental Table 3 gives the complete list of array element IDs from 1.A to 9998.E, their sequences, length, T_M and GC content. Also indicated is the sequence from which the oligo was derived, i.e., EMBL sequences, private sequences, or the corresponding ACEG contig. Correspondence to gene models and annotation data is also provided, when available. Finally, the 37 sequences that did not yield usable oligonucleotides are indicated at the end of the file. Some oligonucleotides did not produce a significant hit on the nuclear genome sequence of *C. reinhardtii* and are labeled as 'NO HITS' in this table. This reflects the fact that oligonucleotides were mainly derived from ACEG contig sequences and not genome sequences, and that genome sequences are still of insufficient quality in some regions of the nuclear genome. Also, oligonucleotides that span an intron–exon junction were not recognized by the blast algorithm. This matter is currently being investigated and appropriate corrections will be made. The '.gal' file to be used with this microarray can be downloaded from http://www.chlamy.org/micro.html.

Influence of Sulfur deprivation on transcript abundance

To examine the consistency of the new oligonucleotide array with respect to the previously generated cDNA array for *C. reinhardtii*, we compared expression profiles of sulfur-starved *C. reinhardtii* cells generated using the oligonucleotide array with previous results reported for the cDNA array (Zhang et al. 2004).

The results of an oligonucleotide array analysis for a subset of genes previously associated with sulfur deprivation (mostly for sulfate uptake and assimilation) are given in Fig. 7a. Figure 7b presents previous results in which changes in the levels of transcripts encoding the same proteins as represented in Fig. 7a, following sulfur deprivation, were monitored using a cDNA microarray (Zhang et al. 2004). While the changes in transcript levels were not the same for the two experiments (compare the upper and lower bar graphs), the changes were always in the same direction; all transcripts except the one encoding the bacterial type sulfite reductase (SIR3) increased. The difference in the fold increase could be explained in a number of ways. One possibility is that the oligonucleotide probes may be less sensitive, yielding somewhat lower signals than the cDNA probes, while another possibility is that the oligonucleotide probes may have greater specificity. The latter case is suggested by the finding that while the *ARS1* transcript shows a twofold increase during sulfur deprivation, which is 20 times less than the value reported previously (Zhang et al. 2004), the oligonucleotide array detected a 6.5-fold increase in *ARS2* transcripts, as shown in Fig. 7c. The cDNA probe would not have distinguished the two transcripts (and potential transcripts from other members of the *ARS* gene family; there are potentially nine *ARS* genes), which are distinguished by the oligonucleotide probes.

This new array also identified transcripts that increased following the imposition of sulfur deprivation that were not present on the v1.1 array (Fig. 7c). In some cases, we expected an increase in the level of these transcripts; an example of this is the transcript for ECP88 (Takahashi et al. 2001). In other cases, the transcripts encode proteins that might serve a regulatory function, including two transcripts for serine threonine kinases, the Tbc2 translation factor, the CYCAB1 AB-type cyclin related protein and a putative mRNA binding protein. Furthermore, transcripts for some of the putative sulfate transporters, such as SULTR2, increase slightly, which has been observed by qPCR analyses (Pootakham and A.R. Grossman, unpublished). Finally, there are a number of transcripts encoding proteins of unknown function that are controlled by sulfur deprivation; it will be a challenge to uncover the functions of the polypeptides encoded by these transcripts in the acclimation response.

Discussion

The availability of complete genome sequences has made it feasible to develop microarrays that represent the entire transcriptome of an organism. These arrays could be used for gene discovery and studying global expression of genes, diagnoses of disease states, acclimation of cells/organisms to environmental change, as well as an analysis of DNA copy number (Choudhuri 2004; Dharmadi and Gonzalez 2004; Mantripragada et al. 2004). Oligonucleotide probes representing individual genes are designed from predicted gene models that are bioinformatically derived from the genome sequence and/or the use of the available cDNA information. The first generation of the *C. reinhardtii* microarrays consisted mainly of a nonredundant set of cDNA sequences of various lengths. The varying lengths of these cDNA fragments and the fact that they were not exhaustively

checked for specificity probably resulted in nonoptimal signal specificity and differences in the sensitivity of the individual probes. To improve upon specificity of probe hybridization, many groups have shifted to the use of synthetic oligonucleotides of similar size and T_M value, and that have been optimized for specificity. However, while some studies indicate a reasonable correlation between the results obtained using different array formats (Kim 2003; Wang et al. 2005), others have found marked inconsistencies among the different formats (Kothapalli et al. 2002; Kuo et al. 2002; Tan et al. 2003; Hollingshead et al. 2005). These differences highlight the difficulties in directly evaluating the output quality of microarray experiments, and care should be taken when using any of the commercially developed platforms. The major issues to be considered when working with an oligonucleotide microarray are: (1) the quality of transcriptome/genomic sequences employed to develop the set of oligonucleotides, (2) the length and fidelity of the individual oligonucleotides, and (3) the specificity with which the oligonucleotides hybridize with their target sequence.

The transcriptome sequences used as the database to design oligonucleotide sequences that are specific for individual genes must be of a high quality. Ideally, only sequences demonstrated to represent mature transcripts should be used for designing the oligonucleotides; gene models predicted solely from genomic sequences might not represent a gene and/or might have errors in the predicted intron–exon junctions. The distribution of sizes of each synthesized oligonucleotide should also be carefully evaluated (by mass spectrometry). The end product of each synthesis should ideally contain over 50% full-length product. Furthermore, each oligonucleotide should have a high specificity for its target genes, and the population of all oligonucleotides used on the array should have similar T_M values. Oligonucleotides that are truncated or contain sequence errors will cause a significant loss of signal (poor hybridization to target gene) and/or nonspecific hybridization, and based on our experience, the frequency with which the individual oligonucleotides are truncated can be relatively high (often the level of full-length product is significantly below 50% and sometimes below 10%). We characterized synthetic oligonucleotides purchased from three different vendors, with respect to distribution of lengths in the population of each of the oligonucleotide sequences by mass spectroscopy. The proportion of accurate, full-length oligonucleotides in a population representing an individual oligonucleotide varied significantly; the variation was substantially dependent on the company that performed the synthesis. In some cases, the full-length oligonucleotide product represented significantly less than 20% of the total product. The product for nearly all the oligonucleotides synthesized by Integrated DNA Technologies were better than 50% full-length. Therefore, the set of array elements developed have high, uniform specificity and sensitivity.

To generate the *C. reinhardtii* oligonucleotide-based array, we developed a sequential protocol that allowed us to select a high-quality unigene set of sequences representing all known expressed sequences of the *C. reinhardtii* transcriptome. This procedure was particularly important since the available cDNA and ACEG information exhibited large variations in sequence quality; some of the sequences were redundant, some of the ACEGs contained multiple contigs with different intron–exon content, and some ACEGs showed significant sequence similarity with other predicted genes of the nuclear genome or with other ACEGs. The last situation might result from the generation of duplicate ACEGs that were not assembled by the Phrap assembly program, the occurrence of gene families or the occurrence of highly similar repeated domains. A thorough bioinformatics analysis led to the generation of a final unigene set that we used to develop the oligonucleotide array elements; the design of these elements was based on the most accurate sequence information, and the final set showed a high degree of target gene specificity. It should be noted that the selection protocol for identifying unique sequences (ACEG contig, EMBL sequence or private sequence) used for oligonucleotide design could result in the loss of gene representation on the array since those sequences that exceed the set similarity threshold would be eliminated (e.g., members of gene families that have similar DNA sequences). However, problems associated with analyzing large genome and EST databases make it difficult (and impractical) to scrutinize individual, ambiguous situations that might allow us to capture additional genomic information. Furthermore, when analyzing gene expression for individual members of gene families, array data is generally not reliable; a quantitative analysis of expression of individual members of a gene family is better achieved using qPCR. For example, members of the *LHC*, and particularly members of the *LHCB* gene family, did not pass our uniqueness filter, and thus many were not included in the current printing of the microarray. We are performing oligonucleotide design and manual sequence optimization for individual members of gene families; these oligonucleotides will be included in the next version of the array. A similar step will be taken in the future to generate specific oligonucleotides for all genes not currently represented in the array. Finally, while our conservative approach resulted in the loss of some ACEG/gene information, it also makes it more likely that essentially all the array elements represent unique genes. The population of ACEGs placed in the checkpool and the remaining pool represents a fertile database for discovering new members of gene families.

To examine the consistency of the new oligonucleotide array with the previously generated cDNA array for *C. reinhardtii*, we compared previous expression profiles obtained with the cDNA-based array for sulfur-starved cells with results using the new array. All the transcripts previously associated with sulfate acquisition and assimilation that were determined to increase during

sulfur deprivation using the v1.0 cDNA array (Zhang et al. 2004) were also shown to increase based on an analysis with the v2.0 oligonucleotide array These transcripts include those encoding arylsulfatase, ATP sulfurylase, serine acetyltransferase, the selenium binding protein and the SAC1-like proteins, which probably represent Na^+/sulfate co-transporters (Pootakham and A.R. Grossman, unpublished data). In most cases, the increase in the transcript level calculated from an analysis of the oligonucleotide array was not as high as that of the cDNA array. However, this observation may reflect an increased specificity of the new array. This is almost surely true for the *ARS* transcript. There are at least two *ARS* genes (possibly more), and the cDNA probe likely hybridized to transcripts from both genes; the transcript levels for both genes increased during sulfur deprivation based on analyses using the specific oligonucleotide probes (Fig. 7). Also, a recent comparison of short oligonucleotide arrays (25–30 mers), long oligonucleotide arrays (50–80 mers) and cDNA arrays has highlighted the fact that while the direction of change of transcript abundance could be reproduced on those different platforms using the same RNA preparation, the magnitude of change varied significantly between platforms (Petersen et al. 2005).

Finally, the new array has enabled us to identify a number of novel transcripts that increased in response to sulfur deprivation (Fig. 7c). Some of these transcripts encode putative regulatory elements that could potentially function at the level of transcription, RNA stability and translation. These results clearly demonstrate that this array represents an important new tool that can be profitably used by the community of researchers working on *C. reinhardtii* to generate a more global view of gene expression in this organism, and will foster comparisons of this model alga with other model photosynthetic systems such as *Arabidopsis thaliana*.

Methods

Microarray fabrication

Microarrays were fabricated in the Stanford Functional Genomics Facility (SFGF) at Stanford University (http://www.microarray.org/). Printing material was suspended in 8 µl 3XSSC on a Beckman Coulter BioMek FX liquid handling robot to a concentration of 50 µM. This print material was then deposited onto Corning GAPS II or UltraGAPS slides using a custom-built microarray robot equipped with Majer Precision MicroQuill 2000 Array pins. Replicate spots were created by printing the entire plate set twice in succession. The ".gal" file, which locates specific sequences on the array, and also reports genes present on the array, their gene models (when available), and available gene annotation information, is available at the website http://www.chlamy.org/micro.html. Printed slides were maintained in a desiccator until they were used. Prior to use, the slides were hydrated in humidity chamber (100% humidity) for 5 min. The slides were then snap dried on a 100°C hot plate (~3 s, array side up) and the array elements UV cross-linked to the aminosilane surface of the slide at 600 mJ (=6,000×100 µJ) using the Stratalinker. The arrays were then stored until prehybridization.

RNA isolation

Total RNA was isolated as previously described (Zhang et al. 2004). The RNA used for preparing cDNA probes was isolated from both nutrient-replete and sulfur-starved (10 h) wild-type cells. 10 µl of 10x DNase I buffer and 2.5 µl of DNase I were added to approximately 40 µg of total RNA; the total volume of the reaction was made 100 µl with RNase-free water. The reaction was incubated for 10 min at room temperature and the RNA purified from the reaction components using the *Qiagen RNeasy MinElute Kit* (Qiagen, Valencia, CA, USA, Cat# 74204), which eliminates degraded DNA, tRNA, 5.5 rRNA, DNAse and other proteins and potential inhibitors of the reverse transcriptase reaction. The volume of the eluted RNA was made 100 µl with sterile water and thoroughly mixed with 350 µl of buffer RLT followed by the addition of 250 µl of 100% EtOH. The solution was then immediately loaded (700 µl) onto an RNeasy MinElute column in a 2 ml collection tube. The column assembly was centrifuged at maximum speed (14,000 rpm) for 15 s, the flow through was discarded, the column transferred to a new 2 ml collection tube, 500 µl of buffer RPE was added to the top of the column, which was then centrifuged at max speed for 15 s. The flow-through was discarded and 500 µl of 80% EtOH was added to the column, and the column was centrifuged at maximum speed for 3 min. Again, the flow-through was discarded, and the column was placed in a new collection tube and centrifuged at full speed for 5 min to eliminate the remaining ethanol. To elute the RNA, the spin column was placed in a new 1.5 ml Eppendorf tube, 20 µl of milliQ water preheated to 40°C was placed at the center of the matrix of the column, and the column was centrifuged at maximum speed for 4 min. The OD_{260} of the eluted RNA was measured and 4 µg of purified RNA reserved for the labeling reactions.

Labeling and purification of reverse-transcribed cDNAs

Four microgram of purified RNA was adjusted to 4 µl with sterile milliQ treated water. One microliter of oligo-dT-(V) (2 µg/µl), consisting of 23 consecutive T residues followed at the 3′ end by an A, T, G or C, was added to the solution prior to heating the reaction mixture for 10 min at 70°C and then quickly chilling it on ice. The following was then added to the reaction mixture: 2 µl 5X superscript buffer; 1 µl 0.1 M DTT, 0.2 µl 50x dNTPs (5 mM dATP, dCTP, dGTP, 10 mM dTTP), 1 µl Cy3- or Cy5-dUTP, 0.8 µl Superscript RT (200 U/

µl), with a final reaction volume of 10 µl. The reaction was incubated at 42°C for 2 h followed by the addition of 0.5 µl Superscript RT and a further 30 min incubation at 50°C. The reaction was stopped by the addition of 0.5 µl of 500 mM EDTA and 0.5 µl of 500 mM NaOH and the solution incubated at 70°C for 10 min to degrade RNA. Neutralization of the reaction mixture was achieved by the addition of 0.5 µl of 500 mM HCl. The Qiagen Qiaquick PCR purification kit (Qiagen Inc., Cat #28106) was used to purify the labeled cDNA. The Cy3- and Cy5-labeled cDNAs were mixed with 90 µl of DNase-free water and 500 µl of Buffer PB. The solution was thoroughly mixed and immediately placed onto a QiaQuick column. The column was washed with 750 µl of Buffer PE by centrifugation at maximum speed for 1 min and the flow-through was discarded. The wash procedure was repeated and the column was centrifuged at maximum speed in an Eppendorf microcentrifuge for 2.5 min to remove the remaining Buffer PE. The column was then transferred to a new 1.5 ml Eppendorf tube and 50 µl of milliQ water preheated to 40°C was applied to the resin bed. After 1 min incubation, the column was centrifuged at maximum speed for 4 min to elute the labeled cDNAs. The eluate was dried in the dark in a SpeedVac; keeping the probe in the dark prevents photobleaching.

Hybridization of the oligonucleotide arrays

All solutions used in the prehybridization and hybridization protocols were filtered through 0.2 µm Nalgene Bottle–Top filters and when possible autoclaved. Prehybridization was performed *immediately* before starting the hybridization. The arrays were incubated for 1 h in the pre-warmed prehybridization solution (5X SSC, 25% formamide, 0.1% SDS, 0.1 mg/ml BSA) at 42°C. Following this incubation, the slides were transferred to 0.1X SSC and gently agitated at RT for 5 min. The 0.1X SSC wash was repeated and the arrays were then transferred to ddH_2O for 30 s and dried by centrifugation at 1,000 rpm for 10 min (Eppendorf Centrifuge 5810R). The dried cDNA was resuspended in 25 µl of milliQ water followed by the addition of 25 µl of 2X hybridization buffer (6X SSC, 0.2% SDS, 0.4 µg/µl poly(A), 0.4 µg/µl yeast tRNA, 40% formamide), both preheated to 40°C (preventing precipitation of SDS). The resuspended samples were boiled for 3 min, centrifuged for 2 min in an Eppendorf microfuge at full speed to remove debris, and then 50 µl of the probe solution was placed in the middle of the prehybridized, dried array. The solution spreads over the entire surface of the array when a Fisher brand large coverslip (Fisherbrand Microscope Cover Glass 12-544-G22X60-1.5) is carefully placed over the surface of the array. Three drops of 10 µl 3X SSC were placed on the surface of the array (not too close to the position of the coverslip) and then the array was sealed in an air-tight chamber and incubated in a 42°C water bath for 24 h. For washing the slides after hybridization, containers with 350 ml 2XSSC/0.1% SDS were preheated to 42°C, and 350 µl of freshly prepared 0.1M DTT was added to each just before use. The hybridization chambers were removed from the water bath, and the individual slides immersed in 2XSSC/0.1% SDS (in one of the containers) until the coverslip moved away from the slides. The slide was then transferred to fresh, preheated 2XSSC/0.2% SDS and gently dipped up and down for 5 min, followed by 5 min washes in 0.1XSSC/0.1% SDS and 0.1XSSC, both at room temperature. The slides were rinsed in 0.01XSSC for 10 s and immediately dried by centrifugation. Detailed and updated versions of the protocols used for RNA-labeling, slide prehybridization, hybridization and washing can be downloaded at http://www.chlamy.org/micro.html

Scanning, quantification and analysis of the slides

The slides were scanned using a GenePix 4000B scanner (Molecular Devices, CA, USA) and GenePix pro 4.0 (Molecular Devices). The images were quantified as previously described (Zhang et al. 2004), and analyses of the data were performed using GeneSpring 6.1 (Agilent Technologies, CA, USA). Two replicate slides, including a dye-swap experiment with two sets of array elements each, were used for the analysis. The signal-to-noise ratio (SNR) defined as (signal mean − background mean)/(background standard deviation) combined for the Cy3 and Cy5 channels for the spots were as follows: 2.9% of the spots had an SNR > 100; 22.2% > 10; 57% > 3 and 69% > 2, while 43% of the spots had an SNR < 3 and 31% < 2. Considering a value of SNR > 3 as representing exploitable spots (Tiquia et al. 2004; see also http://www.biocompare.com/techart.asp?id = 909), a little more than half of the spots printed on the array gave meaningful data under these experimental conditions. Lowering the threshold to that of an SNR > 2 yields meaningful signals for about two-thirds of the spots; the validity of many of these signals would require extensive post-array investigation, including RT-qPCR or RNA blot hybridizations.

Acknowledgments We would like to thank Michael Fero, Elena Seraia and John Coller of the SFGF Laboratory at Stanford for technical advice, printing the arrays and preparing the .gal file in a timely and expert manner. Furthermore, the development of the microarray presented in this manuscript could not have been accomplished without the major sequencing effort of the *C. reinhardtii* genome that was performed by the Joint Genome Institute (Walnut Creek, CA, USA). This work was supported by NSF Grant MCB 0235878 awarded to ARG. The authors would also like to thank Wirulda Pootakham and Olivier Vallon for helpful discussions and Erika Schraner for help with Supplemental Table 3.

References

Asamizu E, Nakamura Y, Sato S, Fukuzawa H, Tabata S (1999) A large scale structural analysis of cDNAs in a unicellular green alga, *Chlamydomonas reinhardtii* I Generation of 3433 non-redundant expressed sequence tags. DNA Res 6:369–373

Asamizu E, Miura K, Kucho K, Inoue Y, Fukuzawa H, Ohyama K, Nakamura Y, Tabata S (2000) Generation of expressed sequence tags from low-CO_2 and high-CO_2 adapted cells of *Chlamydomonas reinhardtii*. DNA Res 7:305–307

Barrett JC, Kawasaki ES (2003) Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. Drug Discov Today 8(3):134–141

Choudhuri S (2004) Microarrays in biology and medicine. J Biochem Mol Toxicol 18:171–179

Davies J, Yildiz F, Grossman AR (1996) Sac1, a putative regulator that is critical for survival of *Chlamydomonas reinhardtii* during sulfur deprivation. EMBO J 15:2150–2159

Dent RM, Han M, Niyogi KK (2001) Functional genomics of plant photosynthesis in the fast lane using *Chlamydomonas reinhardtii*. Trends Plant Sci 6:364–371

Dharmadi Y, Gonzalez R (2004) DNA microarrays: experimental issues, data analysis, and application to bacterial systems. Biotechnol Prog 20:1309–1324

Dutcher SK (2000) *Chlamydomonas reinhardtii*: biological rationale for genomics. J Eukaryot Microbiol 47:340–349

Dutcher SK (2003) Elucidation of basal body and centriole functions in *Chlamydomonas reinhardtii*. Traffic 4:443–451

Elrad D, Grossman AR (2004) A genome's-eye view of the light-harvesting polypeptides of *Chlamydomonas reinhardtii*. Curr Genet 45:61–75

Grossman AR (2000) *Chlamydomonas reinhardtii* and photosynthesis: genetics to genomics. Curr Opin Plant Biol 3:132–137

Grossman AR, Harris EE, Hauser C, Lefebvre PA, Martinez D, Rokhsar D, Shrager J, Silflow CD, Stern D, Vallon O, Zhang Z (2003) *Chlamydomonas reinhardtii* at the crossroads of genomics. Eukaryot Cell 2:1137–1150

Grossman AR, Lohr M, Im CS (2004) *Chlamydomonas reinhardtii* in the landscape of pigments. Annu Rev Genet 38:119–173

Harris EH (2001) *Chlamydomonas* as a model organism. Annu Rev Plant Physiol Plant Mol Biol 52:363–406

He Z, Wu L, Fields MW, Zhou J (2005) Use of microarrays with different probe sizes for monitoring gene expression. Appl Environ Microbiol 71(9):5154–5162

Hollingshead D, Lewis DA, Mirnics K (2005) Platform influence on DNA microarray data in postmortem brain research. Neurobiol Dis 18(3):649–655

Huang K, Merkle T, Beck CF (2002) Isolation and characterization of a *Chlamydomonas* gene that encodes a putative blue-light photoreceptor of the phototropin family. Physiol Plant 115:613–622

Im CS, Grossman AR (2002) Identification and regulation of high light-induced genes in *Chlamydomonas reinhardtii*. Plant J 30:301–313

Kamiya R (2002) Functional diversity of axonemal dyneins as studied in *Chlamydomonas* mutants. Int Rev Cytol 219:115–155

Kane MD, jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res 28(22):4552–4557

Kateriya S, Nagel G, Bamberg E, Hegemann P (2004) "Vision" in single-celled algae. News Physiol Sci 19:133–137

Kim HL (2003) Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells. Exp Mol Med 35:460–466

Kothapalli R, Yoder SJ, Mane S, Loughran TP Jr (2002) Microarray results: how accurate are they? BMC Bioinform 3:22

Kuo WP, Jenssen T-K, Butte AJ, Ohno-Machado L, Kohane IS (2002) Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18:405–412

LaFontaine S, Quinn JM, Nakamoto SS, Page MD, Gohre V, Moseley JL, Kropat J, Merchant S (2002) Copper-dependent iron assimilation pathway in the model photosynthetic eukaryote *Chlamydomonas reinhardtii*. Eukaryot Cell 1:736–757

Ledford HK, Baroli I, Shin JW, Fischer BB, Eggen RI, Niyogi KK (2004) Comparative profiling of lipid-soluble antioxidants and transcripts reveals two phases of photo-oxidative stress in a xanthophyll-deficient mutant of *Chlamydomonas reinhardtii*. Mol Genet Genomics 272:470–479

Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, May-Simera H, Li H, Blacque OE, Li L, Leitch CC, Lewis RA, Green JS, Parfrey PS, Leroux MR, Davidson WS, Beales PL, Guay-Woodford LM, Yoder BK, Stormo GD, Katsanis N, Dutcher SK (2004) Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell 117:541–552

Lohr M, Im CS, Grossman AR (2005) Genome-based examination of chlorophyll and carotenoid biosynthesis in *Chlamydomonas reinhardtii*. Plant Physiol 138:490–515

Mantripragada KK, Buckley PG, de Stahl TD, Dumanski JP (2004) Genomic microarrays in the spotlight. Trends Genet 20:87–94

Mittag M, Wagner V (2003) The circadian clock of the unicellular eukaryotic model organism *Chlamydomonas reinhardti*i. Biol Chem 384:689–695

Mittag M, Kiaulehn S, Johnson CH (2005) The circadian clock in *Chlamydomonas reinhardtii*. What is it for? What is it similar to? Plant Physiol 137:399–409

Miura K, Yamano T, Yoshioka S, Kohinata T, Inoue Y, Taniguchi F, Asamizu E, Nakamura Y, Tabata S, Yamato KT, Ohyama K, Fukuzawa H (2004) Expression profiling-based identification of CO_2-responsive genes regulated by CCM1 controlling a carbon-concentrating mechanism in *Chlamydomonas reinhardtii*. Plant Physiol 135:1595–1607

Moseley JL, Chang, C-W, Grossman AR (2005) Genome-based approaches to understanding phosphorus deprivation responses and PSR1 Control in *Chlamydomonas reinhardtii*. Eukaryot Cell (in press)

Nagel G, Szellas T, Huhn W, Kateriya S, Adeishvili N, Berthold P, Ollig D, Hegemann P, Bamberg E (2003) Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. Proc Natl Acad Sci USA 100:13940–13945

Omran H, Haffner K, Volkel A, Kuehr J, Ketelsen UP, Ross UH, Konietzko N, Wienker T, Brandis M, Hildebrandt F (2000) Homozygosity mapping of a gene locus for primary ciliary dyskinesia on chromosome 5p and identification of the heavy dynein chain DNAH5 as a candidate gene. Am J Respir Cell Mol Biol 23:696–702

Park PJ, Cao YA, Lee SY, Kim JW, Chang MS, Hart R, Choi S (2004) Current issues for DNA microarrays: platform comparison, double linear amplification and universal RNA reference. J Biotech 112:225–245

Pazour GJ (2004) Intraflagellar transport and cilia-dependent renal disease: the ciliary hypothesis of polycystic kidney disease. J Am Soc Nephrol 15:2528–2536

Pazour GJ, Dickert BL, Vucica Y, Seeley ES, Rosenbaum JL, Witman GB, Cole DG (2000) *Chlamydomonas* IFT88 and its mouse homologue, polycystic kidney disease gene tg737, are required for assembly of cilia and flagella. J Cell Biol 151:709–718

Petersen D, Chandramouli GVR, Geoghegan J, Hilburn J, Paarlberg J, Kim CH, Munroe D, Gangi L, Han J, Puri R, Staudt L, Weinstein J, Barret JC, Green J, Kawasaki ES (2005) Three microarray platforms: an analysis of their concordance in profiling gene expression. BMC Genomics 6(1):63

Qin H, Rosenbaum JL, Barr MM (2001) An autosomal recessive polycystic kidney disease gene homolog is involved in intraflagellar transport in *C. elegans* ciliated sensory neurons. Curr Biol 11:457–461

Rochaix JD (2002) *Chlamydomonas*, a model system for studying the assembly and dynamics of photosynthetic complexes. FEBS Lett 529:34–38

Rochaix JD (2004) Genetics of the biogenesis and dynamics of the photosynthetic machinery in eukaryotes. Plant Cell 16:1650–1660

Scholey JM (2003) Intraflagellar transport. Annu Rev Cell Dev Biol 19:423–443

Shrager J, Hauser C, Chang CW, Harris EH, Davies J, McDermott J, Tamse R, Zhang Z, Grossman AR (2003) *Chlamydomonas reinhardtii* genome project. A guide to the generation and use of the cDNA information. Plant Physiol 131:401–408

Silflow CD, Lefebvre PA (2001) Assembly and motility of eukaryotic cilia and flagella. Lessons from *Chlamydomonas reinhardtii*. Plant Physiol 127:1500–1507

Sineshchekov OA, Jung KH, Spudich JL (2002) The rhodopsins mediate phototaxis to low- and high-intensity light in *Chlamydomonas reinhardtii*. Proc Natl Acad Sci USA 99:225–230

Snell WJ, Pan J, Wang Q (2004) Cilia and flagella revealed: from flagellar assembly in *Chlamydomonas* to human obesity disorders. Cell 117:693–697

Stauber EJ, Fink A, Markert C, Kruse O, Johanningmeier U, Hippler M (2003) Proteomics of *Chlamydomonas reinhardtii* light-harvesting proteins. Eukaryot Cell 2:978–994

Stears RL, Martinsky T, Schena M (2003) Trends in microarray analysis. Nature Med 9(1):140–145

Stoughton RB (2005) Applications of DNA microarrays in biology. Annu Rev Biochem 74:53–82

Takahashi H, Braby CE, Grossman AR (2001) Sulfur economy and cell wall biosynthesis during sulfur limitation of *Chlamydomonas reinhardtii*. Plant Physiol 127:665–673

Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC (2003) Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 31:5676–5684

Tiquia SM, Wu L, Chong SC, Passovets S, Xu D, Xu Y, Zhou J (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. Bio Tech 36(4):664–675

Wagner V, Fiedler M, Markert C, Hippler M, Mittag M (2004) Functional proteomics of circadian expressed proteins from *Chlamydomonas reinhardtii*. FEBS Lett 559:129–135

Wang H, He X, Band M, Wilson C, Liu L (2005) A study of inter-lab and inter-platform agreement of DNA microarray data. BMC Genomics 6(1):71

Werner R (2002) *Chlamydomonas reinhardtii* as a unicellular model for circadian rhythm analysis. Chronobiol Int 19:325–343

Wostrikoff K, Girard-Bascou J, Wollman FA, Choquet Y (2004) Biogenesis of PSI involves a cascade of translational autoregulation in the chloroplast of *Chlamydomonas*. EMBO J 23:2696–2705

Wykoff D, Grossman A, Weeks DP, Usuda H, Shimogawara K (1999) Psr1, a nuclear localized protein that regulates phosphorus metabolism in *Chlamydomonas*. Proc Natl Acad Sci USA 96:15336–15341

Zhang Z, Shrager J, Jain M, Chang CW, Vallon O, Grossman AR (2004) Insights into the survival of *Chlamydomonas reinhardtii* during sulfur starvation based on microarray analysis of gene expression. Eukaryot Cell 3:1331–1348