

## The Generation and Organization of Chlamydomonas cDNA Information

Jeff Shrager\*, Charles Hauser, Chiung-Wen Chang, Elizabeth H Harris, John Davies, Jeff McDermott, Raquel Tamse, ZhaoDuo Zhang, and Arthur R Grossman<sup>1</sup>

The Carnegie Institution of Washington, 260 Panama Street, Department of Plant Biology, Stanford, CA 94305 (J.S., C.-W. C., A.R.G., Z.-D.Z.); Iowa State University, Dept. of Botany, 353 Bessey Hall, Ames, IA, 50011 (J.M., J.D. presently: Exelixis Plant Sciences, 16160 SW Upper Boones Ferry Road, Portland, OR 97224); Stanford Genome Technology Center, 855 California Avenue, Palo Alto, CA 94304 (R.T.); Duke University, Biology Department, DCMB Box 91000, Durham, NC 27708 (C.H., E.H.H.)

1. This work was supported by National Science Foundation, Molecular and Cellular Biosciences grant #9975765.

\* Corresponding author: jshrager@andrew2.Stanford.edu, (650) 325-1521x287.

### Introduction

A large volume of Chlamydomonas cDNA sequence information was generated, analyzed, and organized into a convenient public database. A brief description of the organization of the data and the potential to use of the data to identify genes predominantly expressed under specific environmental conditions was presented in a recently submitted manuscript (Shrager et al., 2003). The present document provides an expanded description of the cDNA production, sequencing, and assembly protocols used to produce the results in that paper, providing the reader with additional details, and a better understanding of the strengths and limitations of the techniques used to generate the substrates for analyses, and of the analytical tools applied to the sequence information.

### Materials and Methods

#### Library Construction:

**a. RNA isolation:** A 100 ml culture of cells was grown to a density of approximately  $5 \times 10^6$  cells/ml, pelleted at  $3,600 \times g$  for 5 min and resuspended in 2 ml of lysis buffer (2% SDS, 200 mM NaCl, 40 mM EDTA, 80 mM Tris-HCl, pH 8.0). In some cases, 5  $\mu$ l of a proteinase K stock solution (40 mg/ml) was added to the suspension, which was then incubated for 20 min at RT. The suspension was extracted twice with an equal vol of Tris-buffered phenol, pH 8.0, and once with an equal vol of chloroform:isolamyl alcohol (24:1). Nucleic acid was precipitated from the aqueous phase overnight at 4°C following the addition of 2 vol of 100% ethanol. The precipitate was pelleted by centrifugation for 20 min at  $10,000 \times g$  at 4°C, washed with 2 ml of 70% ethanol, resuspended in 600  $\mu$ l of DEPC-treated H<sub>2</sub>O, and then transferred to 1.5 ml Eppendorf tubes. The nucleic acid solution was mixed with an equal vol of 4 M LiCl and incubated overnight at 4°C. Precipitated nucleic acid was pelleted at full speed in an Eppendorf

microfuge at 4°C for 20 min, washed with 500 µl of 70% ethanol and then resuspended in 100 µl of DEPC-treated H<sub>2</sub>O. This procedure yields 0.5-2 µg nucleic acid/ml, most of which is RNA.

**b. Preparation of Single-Stranded DNA:** cDNA libraries were constructed using a λZAP cDNA synthesis kit (Stratagene, La Jolla, CA) according to the manufacturer's protocol, but was modified in order to accommodate the GC-rich content of *Chlamydomonas* nuclear genes. The reverse transcription reaction was performed using Superscript II reverse transcriptase (Life Technology, Gaithersburg, MD) at 50°C to eliminate most secondary structure in the RNA. Since first strand synthesis was performed at the higher temperature, a long primer GAGA-XhoI-(dT)30, which has a 3' end extended by 12 OligodT [from (dT)18 to (dT)30], was used to facilitate the annealing reaction.

pBlueScript phagemids containing cloned cDNA inserts were excised by co-infection with ExAssist helper phage (Stratagene, La Jolla, CA) according the manufacturer's mass excision protocol. After determining the titer of the phage solution, the excised phagemids were used to infect SOLR cells, and infected bacteria plated at 50,000 pfu/plate (150 x 15 mm) on LB/ampicillin (100 µg/ml) medium. Colonies were scraped from plates and each plate was washed with 1 ml of LB medium to recover residual cells. Phagemid DNA was extracted from the cells using the QIAquick MaxiPrep kit (Qiagen, Valencia, CA).

To generate and isolate single strand copies of the cDNA clones, phagemid DNA was electroporated into XL1-Blue MRF' electroporation competent cells, and transformants grown in 100 ml of LB/ampicillin (100 µg/ml) to an OD<sub>600</sub> of 0.2 at 37°C with shaking (250-300 rpm). Cells were then super-infected with VCSM13 helper phage by adding phage to an M.O.I. of 20, and the suspension incubated at 37°C with gentle agitation for 1 h. Next, kanamycin was added to the suspension to 75 µg/ml. Following a 3 h incubation at 37°C, the cells were pelleted by centrifugation at 16,000 x g, 4°C for 20 min and the supernatant incubated for 1 h with DNase I (final concentration 10 U/ml) at RT. 25 mL of 2.5M NaCl/20% PEG was then added to the phage suspension, the resulting mixture incubated for an additional 1 h on ice, centrifuged at 16,000 x g for 20 min at 4°C, and the pellet resuspended in TE. The suspension was made 100 µg/ml proteinase K and 0.1% SDS. After a 1 h incubation at 45°C, the suspension was extracted twice with phenol, once with chloroform, and nucleic acid precipitated overnight at -20°C in 2 vol ethanol. The pellet was resuspended in TE, digested with *Pvu*II (to eliminate DNA duplexes), and single-stranded DNA purified by hydroxylapatite (HAP) chromatography according to a modified procedure of Soares, et al. (Soares et al., 1994). Our procedure differed from that of Soares and colleagues in that after the samples were extracted with water-saturated ether, they were desalted by passage through QIAquick PCR spin column (Qiagen, Valencia, CA), followed by passage through a Chroma Spin-200 TE (pH 8.0) column (Clontech, Palo Alto, CA), and then ethanol precipitated.

**c. Normalization:** Libraries were normalized or subtracted to reduce the abundance of high copy number cDNAs using a modification of the re-association-based method developed by Bonaldo et al. (Bonaldo et al., 1996). A ten-fold excess of cDNA inserts ("Driver DNA") generated by PCR amplification of a small fraction of the single-stranded library was hybridized to the single-stranded cDNA clones of the library until a C<sub>0</sub>t value of 5 was reached. These conditions allow the highly and moderately abundant cDNAs to anneal to form double stranded DNA, and the remaining single-stranded phagemid DNA was purified by hydroxylapatite

column chromatography and converted to double-stranded cDNA, yielding the normalized library.

Driver DNA used for normalization was amplified using the Qiagen PCR system. The Q-solution was included in the procedure to facilitate amplification of GC rich *Chlamydomonas* sequences. 5-10 ng of single-stranded circular DNA was mixed with 2  $\mu$ l of 10 mM dNTP, 10  $\mu$ l of 20  $\mu$ M T7 and T3 primers, 10  $\mu$ l of 10x Qiagen PCR buffer, 20  $\mu$ l of 5x Q-solution, and 2.5 U of Qiagen Taq DNA polymerase (final vol 100  $\mu$ l). Single-stranded DNA was amplified by PCR as follows: An initial 3 min incubation at 94°C, followed by 30 cycles of 30 s at 94°C, 1 min at 46°C, and 4 min at 68°C. The final cycle ended with an incubation of 10 min at 68°C. Amplified inserts used as driver DNA were purified using the QIAquick PCR purification kit followed by ethanol precipitation. Driver DNA (10  $\mu$ g) was re-suspended in 20  $\mu$ l TE and mixed with 1  $\mu$ g of the single-stranded library, 200  $\mu$ g each of 5' blocking oligonucleotide (5'-GAAT TCCT GCAG CCCG GGGG ATCC ACTA GTTC TAGA) and 3' blocking oligonucleotide (5'-AATA CGAC TCAC TATA GGGC GAAT TGGG TACC GGGC CCCC CCTC GAG) and 100  $\mu$ l of deionized formamide (final volume 200  $\mu$ l). This mixture was overlaid with 100  $\mu$ l of mineral oil and heated at 80°C for 3 min prior to the addition of 20  $\mu$ l of 10x hybridization buffer [1.2 M NaCl, 0.1 M Tris (pH 8.0), 50 mM EDTA], and 20  $\mu$ l of 10% SDS. The resulting reaction mixture was incubated at 35°C for 20 h 30 min (calculated  $C_{ot} \sim 5$ ). Salmon sperm DNA (50  $\mu$ g denatured) was added to the hybridization reaction as a carrier, and the remaining single-stranded circles were purified by hydroxylapatite chromatography, precipitated with 2 vol of ethanol, and re-suspended in 20  $\mu$ l of H<sub>2</sub>O. Conversion from single- to double-stranded DNA was performed using the Roche Expand Long PCR system. Primers used for the conversion are four 20-mer oligonucleotides (JM1-JM4, see below) that anneal at positions approximately evenly spaced around the pBlueScript SK. 8  $\mu$ l of single-stranded DNA circles was mixed with 3  $\mu$ g of each of the primers JM1 (5'-GCTA TGTG GCGC GGTA TTAT), JM2 (5'-CTAC CAGC GGTG GTTT GTTT), JM3 (5'-CTGG CGTA ATAG CGAA GAGG), JM4 (5'-TGTG GAAT TGTG AGCG GATA), 5  $\mu$ l of 2 mM dNTP, 5  $\mu$ l of 10x Expand Long PCR buffer (number 2 buffer, 22.5 mM MgCl), 5 U of Expand Long DNA polymerase mix, and amplified for 5 min at 70°C, 1 min at 4°C, 5 min at 55°C, and 20 min at 68°C. Resulting double-stranded DNA circles were purified through a Chroma Spin-200 TE (pH 8.0) column (Clontech, Palo Alto, CA) and precipitated with 2 vol of ethanol. The number of clones represented by the normalized library was quantified by electrotransformation into DH10B *E. coli* cells.

The effectiveness of the normalization procedure was assessed by plating the un-normalized and normalized phage libraries at a density of approximately 5,000 pfu per plate (100 x 15 mm), performing plaque lifts onto nitrocellulose membranes and hybridizing the membranes to radiolabeled probes for the *RBCS2* gene (small subunit of RuBP carboxylase) and *ATSI* gene (ATP sulfurylase). Normalization resulted in a pronounced reduction in the frequency at which the clones encoding both of these polypeptides were represented in the library (Shrager et al., 2001); the representation of high abundance cDNAs decreased by over 100 fold. The subtraction procedure used to reduce sequencing redundancy was the same as for normalization, except that the starting material was the normalized library and the driver DNA was amplified inserts of all previously sequenced cDNA clones (see Shrager et al. 2003).

**Assembly of Contigs and Generation of Unique Gene Sets:** Sequences generated from the Core, Stress I, and Stress II Libraries were assembled based on sequence similarity using the Phrap assembly program ([www.phrap.org](http://www.phrap.org)). Details of the computational methods and parameters that we used for this assembly procedure are provided in **Appendix I**. Assembly was initiated specifically with 3' sequence reads from the cDNA sequence collection. These 3' reads were grouped into consensus sequences (contigs) in an iterative process that incorporates quality assessment of proposed contigs and the return of unassembled sequences to the sequence pool, followed by re-assembly of the reads in this pool until nearly all sequences in the collection have been either incorporated into contigs or determined to be unique. Each such contig is then disassembled, the corresponding 5' reads that are derived from the same clone as the 3' reads are gathered, and both 3' and 5' reads are re-assembled into one or more revised contigs that define a specific cDNA. A diagrammatic representation of this process is shown in **Figure 1** and discussed in more detail below. The final phase of assembly involves evaluation of the results and annotation of identifiable genes.

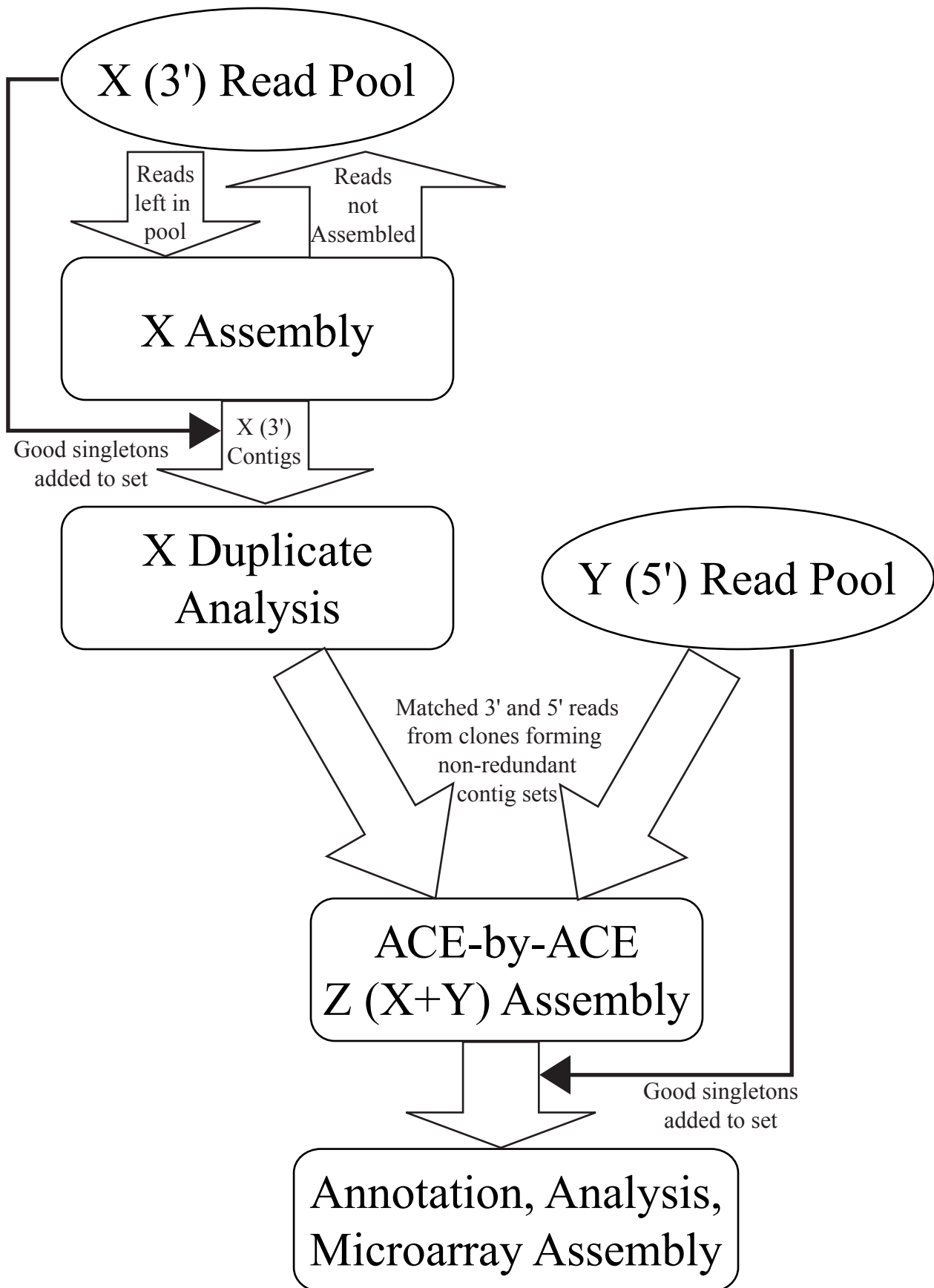


Figure 1. Diagrammatic representation of phases of the cDNA assembly process.

a. cDNA sequences: As described above, each cDNA clone was sequenced both from 3' (x) and 5' (y) ends. The x reads usually contain a 10-30 nucleotide poly-T tail, representing the product of reverse-transcription of poly-A mRNA. Each read is identified by a specific label that is given in the form: <project (library)><plate><well>.<x/y><read-number>. For example, the sequence designated 894030D05.x3 was from Project 894, plate 030, well D05, and was the third read obtained from this clone in the 3' (x) direction. (Often more than one x or y read was generated to improve sequence quality information.) Sequence 894030D05.y1 would be the first 5' (y) read obtained from the same clone. All reads used for subsequent assembly have been submitted to the EST section of GenBank. Prior to initiating sequence assembly, the reads are filtered for vector sequences (PhredPhrap script) and *E. coli* contamination (WU blastn, *E. coli* database), and then manually examined for primer-annealing artifacts (e.g. repeating runs of a single nucleotide). Sequences that pass these filters are placed into the read pool and used for the assembly process.

b. Assembly Process: The ultimate goal of the assembly process is to assign cDNA sequences to unique groups that are designated "ACEs", which theoretically correspond to a unique gene and contain one or more-consensus sequence segments generated from contigs. The contigs may represent full length or partial cDNA sequences and are constructed by aligning individual read sequences that have a high degree of sequence identity. ACEs and contigs are given integer numbers beginning at 1, and a specific contig is represented as <ACE-number>.<contig-number>. For example, 3790.2 is the second contig in ACE number 3790. Each ACE and contig is prefaced in the public database with the date of assembly, in this case June 30, 2002, making the complete designation for this specific contig 20020630.3790.2. The assembly process used to generate contigs and ACEs proceeds in three phases called 'X assembly', 'X duplicate analysis' and 'Z assembly'. These phases are depicted in **Figure 1**.

X Assembly Phase: To initiate sequence assembly, we iteratively assemble contigs (using Phrap, via the PhredPhrap script) from the read pool using only the 3' reads of each clone. Following each iteration, filtering constraints are applied to each read in an assembled contig (see **Appendix I** for definition of the constraints). Reads that pass this filter are eliminated from the read pool and re-assembled (using Phrap) to form a new contig from which a consensus sequence is generated. Reads that are not grouped within an acceptable contig during a particular round of assembly re-enter the read pool for consideration during the next round of assembly. Following each iteration, reads incorporated into acceptable contigs are removed from the read pool, leaving fewer reads to assemble. This cyclic assembly process is repeated until few new contigs are generated. Finally, the remaining high quality reads that do not assemble into contigs are declared to be "singletons", and are added to the contig set. Henceforth singletons are treated exactly as contigs, although they represent a single read.

X-Assembly Duplicate Analysis: In theory the Phrap program produces a unique set of contigs. However, in practice the statistical methods used by the Phrap algorithm may generate a significant number of duplicate contigs. The generation of duplicate contigs reflects the balance between combining reads into a single contig and separating them into different contigs;

therefore, the number of duplicates varies depending on the specific parameter settings of the program. As a result, a stage of duplicate analysis is required. To identify duplicate contigs, we compare the consensus sequence of each X-assembled contig to a dataset of all X-assembled contigs using blastn [WU-Blast ver 2.0], and then compute the maximal set of contigs with overlapping sequence (see **Appendix I** for details). Each such set produces one ACE (proposed unique gene) during the Z assembly phase

**Z Assembly (X + Y):** All of the 3' reads derived from a unique contig, or from an overlapping set of contigs defined by duplicate analysis as described above, are hypothesized to represent a unique gene. In the Z assembly phase, we combine these 3' (x) reads with their corresponding 5' (y) reads, and this combined pool of 3' and 5' reads is assembled by Phrap using permissive parameter settings (**Appendix I**). The resulting contig or set of contigs are placed in an ACE file, and each ACE contains the total of our sequence information for the proposed gene.

**Annotation:** Analysis of contigs and gene products encoded by an ACE involves three phases: (1) Placing the Z assembled contigs into a non-redundant uni-gene set; (2) finding contigs encoding previously identified Chlamydomonas genes by cross-matching sequences of the set of Z contigs to a non-EST Volvocales database with more than 2000 sequences; (3) identifying potential orthologs of the genes encoded by the remaining contigs. This process is described in Shrager et al. (2003).

**Uni-gene set:** Independent of Phrap-mediated assembly of contigs into ACE groups, described above, we generated a set of unique EST groups by cross-matching each cDNA sequence read against a dataset of all Chlamydomonas EST reads using WU-blastn ( $S > 1000$ , or HSP,  $> 95\%$  identity). This EST grouping contains sequences from all widely used strains, and includes the clones sequenced by the Kazusa group (Asamizu et al., 1999). See Shrager et al. (2003) for additional information.

**Web Access to Sequence and Annotation Data:** The sequence information was used to create a comprehensive relational database using PostgreSQL (<http://www.postgresql.org/>) which is maintained at the Chlamydomonas Genome Project web site ([http://www.biology.duke.edu/chlamy\\_genome/](http://www.biology.duke.edu/chlamy_genome/)). This database may be queried for annotation ('Gene search') or sequence information ('Contig / Clone search').

## Results

The assembly designated 20020630 (2002, June 30, the date on which this assembly was initiated) was generated from 3' (x) and 5' (y) reads (sequences) from individual cDNA clones present in three composite cDNA libraries designated Core, Stress I and Stress II Libraries, with other libraries including Stress III, Deflagellation, and Gametogenesis-Zygote Libraries being near sequence completion. Growth conditions for cells used in the construction of the different libraries, the project number of each library and the number of plates (96 well microtitre plates) sequenced are presented in the manuscript by Shrager et al. (2003; see **Table I** in this manuscript).

The total sequence read pool for the 20020630 assembly consists of 80,286 reads, distributed among the different libraries, as indicated in **Table 1**. This table also shows the total number of 3' (x), and 5' (y) reads that were initially pooled from each project, the number of reads from each project that was finally included in the assembly ('Reads Incorporated'), the % of reads from each project that was used ('% Incorporated'), and what % of the reads from each project contributed to the final read pool ('% Total').

<b>Project #</b>	<b>3' (x) Reads</b>	<b>5' (y) Reads</b>	<b>Total Reads</b>	<b>Reads Incorporated</b>	<b>% Incorporated</b>	<b>% Total</b>
874	760	692	1,452	1,061	73%	2
894	6,128	5,470	11,598	9,246	80%	15
963	9,010	10,554	19,564	12,585	64%	20
1024	11,372	10,591	21,963	17,172	78%	27
1031	14,451	11,258	25,709	22,716	88%	36

**Table 1.** Reads used for the 20020630 assembly and incorporated in ACEs.

Reads from the x direction were first assembled into sets of overlapping or contiguous sequences called **contigs**, each of which is represented by a consensus sequence; reads of high quality that do not overlap with any other read (referred to as 'singletons') are also designated contigs in the final assembly. Four iterations of the X assembly cycle, described in detail in the **Materials and Methods**, were used. The results from each iteration are shown in **Table 2**. A total of 33,040 reads were used for final contig generation. Of the 11,256 contigs generated during the X-assembly, 5,995 were 'true' contigs with more than one read (5,374 from the first iteration + 433 from the second iteration, 136 from the third iteration, and 52 from the fourth iteration) and 5,261 were singletons.

<b>Round</b>	<b>Reads Included</b>	<b>Final Contigs</b>
1	29,651	5,374
2	2,309	433
3	746	136
4	334	52

**Table 2.** x reads used and contigs generated during the X assembly.

After all of the x contigs (and singletons) were tested for duplications within the contig pool (see **Materials and Methods** for details), 8,629 maximal overlapping groups were generated. The 5' (y) reads for each of the 3' (x) reads were gathered and grouped with their corresponding x read; the 3' contig plus its corresponding 5' reads constitute an ACE. The 3' and 5' reads within each ACE was then re-assembled (Z assembly) into one or more revised contigs that define a specific cDNA. The Z assembly used a total of 62,780 combined x and y reads to generate 14,410 contigs and 8,628 ACEs (one failed to assemble properly and was dropped). Of the 8,628 ACEs generated during this assembly, 4,374 contained two contigs while 3,661 contained a single contig and 593 contained more than 4 contigs. Details of the actual X and Z assembly process are described in the **Materials and Methods** section and **Addendum I**.

**Size distribution of contigs and ACEs:** The longest contig generated by the assembly only used 5' sequences and was 3,321 nucleotides in length and the mean contig size was 791 nucleotides (see Shrager et al. 2003, **Figure 1**). ACEs generated from the assembly that consisted of a single contig in which 3' and 5' reads overlap may represent full-length clones or clones truncated at their 5' ends. For many of the longer cDNAs, the 3' and 5' contigs fail to overlap, and two or more contigs are present in a single ACE. Examples of this are provided in Shrager et al. (2003).

## **Conclusions**

The generation, sequencing, and assembly of cDNAs involves detailed procedures ranging from those required to construct a diverse set of recombinant libraries, to those required for assembling similar sequences and analyzing those sequences for the potential function of the gene product. Some of the procedures used were standard, while others required considerable modifications. The descriptions presented here provide the reader with the details of generation and analysis of cDNA information, and reveal both the advantages and limitations of the procedures that were used.

## **Appendix I: Computational Details**

The assembly process is a complex array of automatic and manual steps, most of which are conducted using Unix shell scripts, standard programs such as Phred and Phrap ([www.phrap.org](http://www.phrap.org)), and special-purpose heuristic analysis code, written in Lisp, a programming language typically used in artificial intelligence applications. The Lisp environment that we use is "ACL" (v. 6.0), distributed by Franz, Inc. ([www.franz.com](http://www.franz.com)). The Lisp programs are available from Jeff Shrager ([JShrager@Andrew2.stanford.edu](mailto:JShrager@Andrew2.stanford.edu)).

The parameters and methods described below were used in the 20020630 assembly. In our experience, each assembly requires slightly different, empirically-determined parameters to obtain the most useful results.

*Phrap parameters:* We used the default Phrap parameters for the X assembly phase, but added "-forcelevel 5" for the Z assembly phase, which makes the Z assembly more permissive.

*Standard read trimming:* In all assembly phases, reads are examined for vector, yeast, and *E. coli* sequence contamination by the PhredPhrap script. Runs of bases likely to be from these sources are replaced by "X"s, and are ignored in subsequent analyses.

Reads and contig consensus sequences are also trimmed according to the base call quality.

Base quality values are produced by the base calling program, Phred ([www.phrap.com](http://www.phrap.com)). A base with a quality of 30 has a 1 in 1000 chance of being wrong. (The quality values are represented as ten times the power of 10, so that a quality value of 10 = 1 chance of error in  $10^{1.0}$ , and a quality value of 33 = 1 chance of error in  $10^{3.3} = \sim 2000$ ). The quality of a read is generally poorest at both ends of the sequence (closest to and furthest away from the sequencing primers), and highest in the middle. Trimming proceeds by calculating a moving average of the quality along the sequence, and then trimming in from both ends of the sequence toward the middle, to the point at which the moving average quality of the read exceeds a value of 20 (that is, with the chance of a base call being wrong at less than 1 in 100 times). The moving average quality used to perform this "standard trim" was calculated using a window with a width of 10% of the length of the entire read, and whose step size is always 1 base. The 10% value was empirically determined. All analyses are performed with the non-X bases that remain after this standard trimming.

*Filtering parameters:* At all points in the assembly process, reads of less than 100 nucleotides in total length, or having an overall mean base quality of less than 10, were excluded from the analysis.

During X assembly, any read with more than 5% "mismatched" bases was removed from its contig and returned to the X-assembly pool for possible use in subsequent X cycles. Mismatched bases are those bases that do not match the consensus sequence of the contig in which they were placed, and whose base quality is greater than or equal to 35. For example, if the consensus sequence has an A, where a particular read has a G, but the G has a quality of 25, this is not counted as a mismatch. However, if the quality of the G is 35, it would be counted as a mismatch. The logic of this is that low quality base calls may be erroneous, and should not be used as evidence to eliminate a read from its contig.

At the end of the X and Z cycles, all remaining reads whose mean quality is greater than 15 (1 error in  $\sim 32$ ), whose length is greater than 200 bases, and with 80% or more "real" bases, are included as singletons. ("Real" bases are one of: G, A, C, or T, excluding Xs, which denote vector sequence, and Ns, indicating that a specific base could not be determined.)

*Computation of maximal overlapping sets (X duplicate analysis):* A maximal overlapping set of matching contigs is defined as the set containing all contigs that match one another, but that do not match any other contigs. For example, if Contig A (cA) matches cB and cC; and cC matches cA and cD; and cD only matches cA, then all of these: cA through cD, compose a single maximal overlapping set of contigs. This composition is required because the matches generated by blast are not associative. That is, if blast determines that cA matches cB, and that cB matches cC, it is not necessarily the case that blast will also report that cA matches cC.

## 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*. Generation of 3,433 non-redundant expressed sequence tags. *DNA Res* **6**: 369-373
- Bonaldo MF, Lennon G, Soares MB** (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* **6**: 791-806
- Shrager J, Chang C-W, Davies J, Harris EH, Hauser C, Tamse R, Surzycki R, Gurjal M, Zhang Z, Grossman AR** (2001) *Chlamydomonas* cDNAs; Assembly and potential role in understanding metabolic processes. In: Proceedings of the 12th International Congress on Photosynthesis. Brisbane, Australia
- Shrager J, Hauser C, Chang C-W, Harris EH, Davies J, McDermott J, Tamse R, Zhang Z-D, Grossman AR** (2003) *Chlamydomonas reinhardtii* Genome Project: A guide to the generation and use of the cDNA information. *Plant Physiol*, In press.
- Soares M, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A** (1994) Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci USA* **91**: 9228-9232