APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Sept. 2004, p. 5298–5304 Vol. 70, No. 9
0099-2240/04/$08.00+0 DOI: 10.1128/AEM.70.9.5298–5304.2004

Modified Serial Analysis of Gene Expression Method for Construction of Gene Expression Profiles of Microbial Eukaryotic Species†

Kathryn J. Coyne,[1] JoAnn M. Burkholder,[2] Robert A. Feldman,[3]‡ David A. Hutchins,[1] and S. Craig Cary[1]*

Graduate College of Marine Studies, University of Delaware, Lewes, Delaware[1]; Center for Applied Aquatic Ecology, North Carolina State University, Raleigh, North Carolina[2]; and Amersham Biosciences, Inc., Sunnyvale, California[3]

Received 16 February 2004/Accepted 13 May 2004

Serial analysis of gene expression (SAGE) is a powerful approach for the identification of differentially expressed genes, providing comprehensive and quantitative gene expression profiles in the form of short tag sequences. Each tag represents a unique transcript, and the relative frequencies of tags in the SAGE library are equal to the relative proportions of the transcripts they represent. One of the major obstacles in the preparation of SAGE libraries from microorganisms is the requirement for large amounts of starting material (i.e., mRNA). Here, we present a novel approach for the construction of SAGE libraries from small quantities of total RNA by using Y linkers to selectively amplify 3′ cDNA fragments. To validate this method, we constructed comprehensive gene expression profiles of the toxic dinoflagellate *Pfiesteria shumwayae*. SAGE libraries were constructed from an actively toxic fish-fed culture of *P. shumwayae* and from a recently toxic alga-fed culture. *P. shumwayae*-specific gene transcripts were identified by comparison of tag sequences in the two libraries. Representative tags with frequencies ranging from 0.026 to 3.3% of the total number of tags in the libraries were chosen for further analysis. Expression of each transcript was confirmed in separate control cultures of toxic *P. shumwayae*. The modified SAGE method described here produces gene expression profiles that appear to be both comprehensive and quantitative, and it is directly applicable to the study of gene expression in other environmentally relevant microbial species.

Serial analysis of gene expression (SAGE) is a powerful and efficient method for compiling quantitative gene expression profiles of specific tissue or cell types (26–28). In contrast to conventional cDNA libraries, this technique allows the simultaneous analysis of gene expression for thousands of transcripts by cataloging short (10-base) diagnostic sequence tags located at a specific site near the 3′ end of each gene transcript. The SAGE method has several advantages over other methods for analysis of differential gene expression (see, for example, references 6, 12, and 13). No prior knowledge of a gene sequence is required, and the sequence data generated can be archived for comparison to future libraries. In addition, the serial and parallel analysis of sequence tags increases data output by several orders of magnitude, generating libraries that are both quantitative and comprehensive. Although SAGE was developed initially for medical research, it has been successfully extended to the analysis of gene expression in a diverse number of species, such as the yeast *Saccharomyces cerevisiae* (28), rice seedlings (17), and the malarial parasite *Plasmodium falciparum* (18).

In spite of the value of SAGE libraries, their construction and analysis can be challenging. A major drawback to the original SAGE protocol is the initial requirement for microgram quantities of mRNA. For biological materials available only in limited quantities, PCR-based methods that allow construction of SAGE libraries from small amounts of total RNA have been developed (10, 14, 19, 22). In particular, preamplification of full-length reverse-transcribed cDNA (19, 22) or restricted, linker-ligated cDNA 3′-end fragments (14) has been shown to provide a suitable substrate for SAGE and results in representative libraries of transcripts.

Here, we describe a modification of the SAGE method that facilitates construction of SAGE libraries from small quantities (1 μg or less) of total RNA, making it directly applicable to the study of gene expression in environmentally relevant microbial species. This protocol differs from previously described SAGE methods in that removal of excess linkers or other (non-3′-end) cDNA fragments is not required for specific amplification of cDNA 3′-end fragments by PCR, thereby reducing the loss of rare transcripts. At the same time, a modification in linker design significantly reduces the formation and amplification of contaminating linker-linker by-products, resulting in an increase in the size and comprehensiveness of the library. Construction of SAGE libraries by use of the modified protocol also provides a pool of linker-ligated cDNA 3′-end fragments that may be stored for amplification by PCR under stringent conditions. Using this method, we have demonstrated that even rare tags can be extended to provide cDNA sequence information for gene identification or development of probes.

We validated this method by preparing SAGE libraries of the toxic life stage of the marine dinoflagellate *Pfiesteria shumwayae*. Current methods for establishing the presence of toxic *Pfiesteria* strains in environmental samples require both examination of the plate structures of isolated cells by scanning

* Corresponding author. Mailing address: University of Delaware Graduate College of Marine Studies, 700 Pilottown Rd., Lewes, DE 19958. Phone: (302) 645-4288. Fax: (302) 645-4007. E-mail: caryc@udel.edu.

† ECOHAB contribution number 102.

‡ Present address: Sym-Bio Corporation, Menlo Park, CA 94025.

electron microscopy and the performance of fish bioassays, which must be carried out in biohazard facilities (4). These assays may take days to several weeks to complete. A more timely and economical prescreening approach is necessary for implementation of rapid-response measures following a fish kill. Recently, unique sites in 18S rRNA gene sequences have been targeted for identification of *Pfiesteria* spp. in environmental samples by PCR (3, 9, 20). Although these techniques are both accurate and sensitive, they cannot be used to distinguish between toxic and nontoxic stages of the *Pfiesteria* life cycle.

Our intent was to develop a method to identify genes that are expressed by *P. shumwayae* during toxic zoospore stages (4). These genes may then be used to design life stage-specific molecular probes for detection of toxic *Pfiesteria* dinoflagellates in environmental samples. Because of the hazards of and inherent difficulties in maintaining large cultures of toxic *P. shumwayae*, we were limited to methods that do not require large amounts of starting material. It was also essential that the results of our analysis provide data that could be archived for future comparisons to other life stages of *P. shumwayae*. We concluded that open-platform, PCR-based methods such as SAGE would be ideal for analysis of gene expression by *P. shumwayae* during toxic life stages. To validate the modifications to the SAGE method, we constructed gene expression profiles of toxic fish-fed and alga-fed cultures of *P. shumwayae* (designated ToxF and ToxA, respectively). Representative transcripts were chosen for further analysis, and their expression was confirmed in separate control cultures of toxic *P. shumwayae*.

MATERIALS AND METHODS

***P. shumwayae* culture.** *P. shumwayae* (isolate CAAE 416T) was maintained in a 9-liter culture vessel filled with 0.2-μm-filtered synthetic seawater (deionized water with Coralife scientific-grade marine salts [Energy Savers, Carson, Calif.]; salinity, 15) at 23°C with a 12 h:12 h light:dark cycle. The culture designated ToxF was fed two or three juvenile tilapia (*Oreochromis mossambicus*; 5 to 7 cm long) daily. Isolate CAAE 416 is highly toxic, causing fish death in from <1 to 2 h in standardized fish bioassays (5). For the culture designated ToxA, zoospores from the toxic culture (ToxF) were gently cleaned by flow cytometry (11), to prevent carryover of contaminating species from the fish culture, and then supplemented with *Rhodomonas* (CCMP 757; Provasoli-Guillard Culture Center, Booth Bay Harbor, Maine) as prey for 48 h prior to extraction. At the time of extraction, the *Pfiesteria*/prey ratio in the ToxA culture was about 10:1. *Pfiesteria* dinoflagellates from each culture were gravity filtered through a 20-μm-pore-size polycarbonate filter to remove debris and then harvested by filtration onto an 8-μm-pore-size polycarbonate membrane under a vacuum with a maximum pressure of 3×10^4 Pa. The filters were immediately immersed in 0.6 ml of solution D [4 M guanidine thiocyanate, 0.5% *N*-lauryl sarcosine, 25 mM sodium citrate (pH 7)] (8) and stored at −80°C until RNA extraction was performed.

RNA extraction. Filtered *P. shumwayae* dinoflagellates in solution D were incubated at 60°C for 5 min and then transferred to a new tube; 0.1 volume of 2 M sodium acetate (pH 4.3) was added, and the mixture was extracted with 1 volume of water-saturated phenol and 0.75 volume of chloroform-isoamyl alcohol (24:1). The supernatant was then extracted with 1 volume of chloroform. Total RNA was precipitated from the supernatant by addition of 0.1 volume of 3 M sodium acetate (pH 7.2) and 1 volume of isopropanol. Total RNA was resuspended in RNase-free water and evaluated by gel electrophoresis.

Y linker preparation. Oligomers L1AY (5′-TCCCTATTAAGCCTAGTGAG CTGATCTTCA-3′) and L2AY (5′-TCCCCGTACATCGTTCTCCTAGGTCC GGAA-3′) were treated with kinase in separate 20-μl reaction mixtures as described in the SAGE protocol (Genzyme [Cambridge, Mass.] Molecular Oncology SAGE protocol 010300). L1AY and L2AY were annealed to oligomers L1A (5′-TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATAGGGAC ATG-3′) and L2A (5′-TTTCTGCTCGAATTCAAGCTTCTAACGATGTACG GGGACATG-3′), respectively, also as described in the SAGE protocol, to form two Y linkers, L1Y and L2Y.

cDNA synthesis. One microgram of total RNA was reverse transcribed in a 20-μl reaction mixture containing 0.5 μM oligo(dT)-Heel primer (5′-CCAGTG TCTTGAGCAGTGACT$_{18}$VN-3′, where V is A, C, or G and N is any nucleotide), 1× SuperScript buffer (Invitrogen, Carlsbad, Calif.), 10 mM dithiothreitol, 0.5 mM each deoxynucleoside triphosphate (dNTP), and 200 U of SuperScript II reverse transcriptase (Invitrogen). The reaction was allowed to proceed for 50 min at 42°C. Second-strand synthesis was carried out in the same tube by addition of 40 μl of 2.5× second-strand buffer (Promega, Madison, Wis.) 9 U of DNA polymerase I (Promega), 0.8 U of RNase H (Promega), and enough water to bring the total volume to 100 μl. The mixture was incubated at 14°C for 2 h. The reaction was stopped by extraction with 1 volume of PC8 (phenol [pH 8.0]-chloroform, 1:1 [vol/vol]), and the cDNA was precipitated at −80°C overnight after addition of 3 μl of glycogen, 100 μl of 10 M ammonium acetate, and 700 μl of ethanol.

Modified SAGE protocol. cDNA was resuspended in 43 μl of LoTE (3 mM Tris-HCl [pH 7.5], 0.2 mM EDTA) and restricted with 10 U of NlaIII restriction enzyme (New England Biolabs [NEB], Beverly, Mass.) in 1× NEB Buffer 4 for 1 h at 37°C. The mixture was diluted to 200 μl with LoTE and extracted with an equal volume of PC8. The restricted cDNA was precipitated as described above, resuspended in 15 μl of LoTE, and divided into three 5-μl pools.

Two pools of cDNA were ligated to Y linkers L1Y and L2Y in separate 10-μl reaction mixtures containing 1× ligase buffer (Invitrogen), 400 ng of annealed Y linker, and 5 U of ligase (Invitrogen) for 2 h at 16°C. A no-ligase control reaction was carried out using the third pool of cDNA containing 1× ligase buffer and 400 ng of L1Y. The ligated products were diluted to 20 μl with LoTE, and a 1:50 dilution was made for PCR amplification. The remaining linker-ligated cDNA was stored at −80°C for future analysis.

The 3′ ends of the linker-cDNA constructs and no-ligase control were then amplified in four 50-μl reaction volumes with 1 μl of diluted template, 1× *Taq* polymerase buffer (Promega), 0.2 mM each dNTP, 1.25 mM $MgCl_2$, 0.5 μM biot-Heel primer (5′-biotin-CCAGTGTCTTGAGCAGTGAC-3′), and 0.5 μM P1-N (5′-amino-GGATTTGCTGGTGCAGTACA-3′) or P2-N (5′-amino-CT-GCTCGAATTCAAGCTTCT-3′). The reaction mixture was incubated at 94°C for 2 min before addition of 2.5 U of *Taq* polymerase (Promega) followed by 22 cycles of PCR consisting of 30 s at 94°C, 30 s at 56°C, and 1.5 min at 72°C. The four replicate reaction mixtures were pooled and the cDNA was extracted with an equal volume of PC8, precipitated as described above, and resuspended in LoTE.

The precipitated amplification products were digested with the tagging enzyme BsmFI in a 50-μl reaction mixture containing 0.2 mM each dNTP, 1× NEB Buffer 4, 1× bovine serum albumin (BSA), and 2 U of BsmFI (NEB) for 1 h at 65°C. Restriction products were filled in by the addition of 3 U of T4 polymerase (NEB) to the reaction mixture and incubated for 20 min at 12°C. The reaction mixture was diluted to a volume of 200 μl with LoTE, and the products were extracted with PC8 and precipitated as described above.

The restricted and filled-in DNA fragments of linker-plus-tag sequence were isolated from a 3% NuSieve agarose gel (FMC Bioproducts, Rockland, Maine) and eluted from the gel by using Gel Elute agarose spin columns (Supelco Inc., Rockford, Ill.). An equal volume of 2× B+W solution (10 mM Tris-HCl [pH 7.5], 1 mM EDTA, 2 M NaCl) was added to the recovered tags, and the mixture was incubated with streptavidin-coated Dyna Beads (M280; Dynal, Oslo, Norway) to remove contaminating biot-Heel primers. The DNA was diluted to a volume of 200 μl with LoTE, ethanol precipitated, and resuspended in 6 μl of LoTE.

L1Y- and L2Y-cDNA restriction fragments were ligated and amplified, as described in the original SAGE protocol, using biotinylated primers P1 (5′biotin-GGATTTGCTGGTGCAGTACA-3′) and P2 (5′biotin-CTGCTCGAATTCA AGCTTCT-3′). The PCR products were precipitated and digested with NlaIII for 1 h at 37°C in a 150-μl reaction mixture containing 1× NEB Buffer 4, 1× BSA, and 120 U of NlaIII. The restriction products were extracted with PC8 and ethanol precipitated.

The 26-bp ditag product of the NlaIII restriction was isolated as described in the original SAGE protocol. An equal volume of 2× B+W solution was added to the recovered ditags, and the mixture was incubated with streptavidin-coated Dyna Beads M280 to remove contaminating linkers and primers (23). The ditags were ethanol precipitated and resuspended in 7 μl of LoTE.

Concatemers were formed by ligation of ditags in a 10-μl reaction volume containing 1× ligase buffer and 5 U of ligase (Invitrogen). The concatemers were size fractionated using Sephacryl-S400 spin columns (Promega), ethanol precipitated, and resuspended in 10 μl of LoTE. Six microliters (about 50 ng) of a 1:5

TABLE 1. PCR primers used in this study

Tag primer	Sequence[a]
343Fwd	GAATGCTTCTTGGAAGATTAG
343Rev	TAGATGAAGATGATAACCGTCT
277Fwd	AATAGGGACATGCATTGAGTCC
277Rev	ACGCCTGGTCTGGATAACA
033Fwd	ATGGAGGTGCCGTCAGAAT
033Rev	GGATAACAGGTTCTTGAATCCTA
012Fwd[b]	AGCTCGTGTTCCGTACG
016Fwd	CATGGTCGCAGCCGTAG
016Rev	ATCACACGGAGGATACGTCG
017Fwd	CATGTATAGCCGCGTTTG
017ARev	CACAAGGCGCCACTCTA
009Fwd	CATGCATTCCGTCTGAGCTT
009Rev	CGTTCGAGATCGACGGA
004Fwd[b]	GGACATGATGGTTGGCG

[a] Linker-specific sequences are underlined.
[b] The Heel primer was used with 012Fwd and 004Fwd.

dilution of the concatemer solution was ligated into SphI-cut pZero plasmid (Invitrogen) and cloned into Top 10 competent cells (Invitrogen).

Glycerol stocks of individual clones were sequenced at the Amersham Biosciences (Sunnyvale, Calif.) production sequencing laboratory. By use of the TempliPhi protocol (Amersham Biosciences), the clones were prepared for DNA sequencing. The TempliPhi method uses bacteriophage ϕ29 DNA polymerase to exponentially amplify circular DNA templates by rolling-circle amplification (16). The products of the TempliPhi reaction were sequenced in both directions with the M13 forward (5′-GTTTTCCCAGTCACGACGTTGTA-3′) and M13 reverse (5′-TGAGCGGATAACAATTTCACAGGA-3′) primers, using DYEnamic ET Terminator Cycle Sequencing kits (Amersham Biosciences). Reaction products were resuspended in formamide and sequenced using a MegaBACE 1000 (Amersham Biosciences) capillary array DNA sequencing instrument. Data were collected and analyzed with the Molecular Dynamics (La Jolla, Calif.) Basecaller version 3.12 software.

SAGE library evaluation. Eight tags, ranging in frequency from 3 tags (0.026% of total tags) to 366 tags (3.3% of total tags) in the libraries, were chosen for further analysis. Six to 10 bases of the L1Y linker sequence were combined with the 10-base tag sequence to produce 16- to 20-base PCR primers. L1Y-ligated cDNA fragments (prepared as described above) were diluted 1:50 and used as templates in PCR. The 3′ ends of the cDNAs of interest were each amplified in 50-μl reaction volumes containing 1 μl of diluted template, 1× polymerase buffer (Sigma Chemical Co., Inc., St. Louis, Mo.), 0.2 mM each dNTP, 1.25 mM $MgCl_2$, 0.5 μM Heel primer (5′-CCAGTGTCTTGAGCAGTGAC-3′), a 0.5 μM concentration of the appropriate linker-tag primer, and 2.5 U of Jump Start *Taq* polymerase (Sigma Chemical Co.). The reaction mixture was incubated at 94°C for 2 min and then subjected to 20 touchdown cycles of 30 s at 94°C, 30 s at 65°C (decreasing by 0.5°C each cycle), and 1.5 min at 72°C. The amplification reaction continued for an additional 15 cycles of 30 s at 94°C, 30 s at 56°C, and 1.5 min at 72°C.

PCR products were cloned into the pCR4-TOPO plasmid (Invitrogen). Several clones from each reaction were sequenced using a Big Dye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems, La Jolla, Calif.) and an ABI PRISM 310 genetic analyzer (Applied Biosystems). Gene-specific primers were designed for each of the eight transcripts and are listed in Table 1. These gene-specific primers were then used to verify expression in ToxF and ToxA *P. shumwayae* strains by amplification of L1Y-ligated cDNA 3′-end fragments, using the PCR protocol described above.

Verification of gene transcripts in toxic cultures. Three control cultures of toxic fish-fed *P. shumwayae*, designated CAAE 103046, 102988, and 102994, were used to verify transcription. *P. shumwayae* cells were isolated from each of the three cultures by differential filtration as described above. RNA was extracted and reverse transcribed using the oligo(dT)-Heel primer as described above.

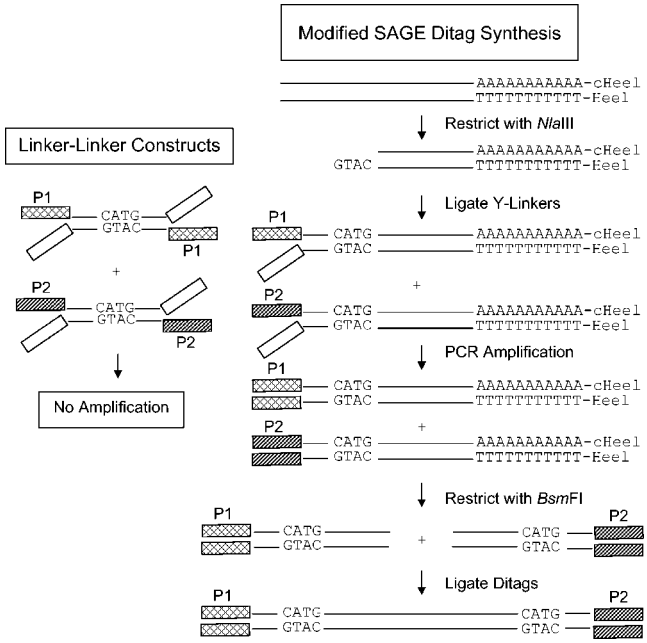

FIG. 1. Schematic diagram of the modified SAGE protocol. P1, P2, and Heel represent PCR primer sites (see text). NlaIII recognition site, CATG.

Transcription of each of the eight genes of interest was verified for the toxic cultures by performing PCR with gene-specific primers as described above.

Nucleotide sequence accession numbers. Nucleotide sequences have been deposited in GenBank under accession numbers CN498777 to CN498784.

RESULTS

Modified SAGE protocol. A schematic diagram of the modified SAGE method is presented in Fig. 1. The first-strand cDNA was primed using a 2-base-anchored oligo(dT) with a 20-base extension at the 5′ end (Heel) that served as a priming site in subsequent PCRs. Second-strand synthesis by DNA polymerase I incorporated a sequence complementary to the 20-base Heel at the newly synthesized 3′ end of each cDNA.

The major modification of the SAGE protocol was the use of Y linkers for selective amplification of the 3′-end fragments of the cDNA library. Linkers were prepared from oligonucleotides with noncomplementary 5′ ends. The two linkers were ligated to separate pools of the NlaIII-restricted cDNA. The cDNA 3′-end fragments were then amplified using primer P1-N or P2-N and the biotinylated Heel primer. Linker-linker ligation products as well as linker-cDNA-linker ligation products were not amplified because they lack sequences complementary to the P1-N and P2-N primers. These complementary sequences were present only after extension of the Heel-primed templates during the first round of PCR. Since the priming site for the Heel primer is located only at the 3′ ends of the cDNAs, only the 3′-most restriction fragments were amplified (24). The appropriate number of cycles for PCR amplification of the cDNA ends was determined empirically by comparing products on a 3% agarose gel. Because PCR bias becomes more pronounced with increased numbers of cycles (25), the smallest number of cycles that generated a visible

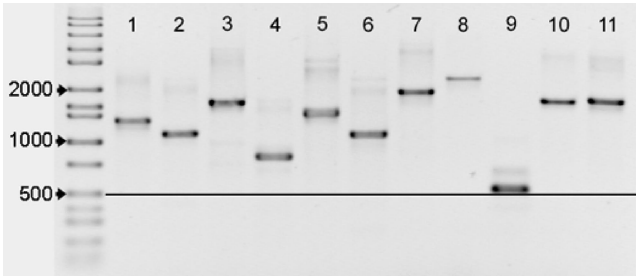

FIG. 2. PCR amplification of random clones of SAGE ditag concatemers. Clones were amplified with (vector-specific) M13 forward and M13 reverse primers.

PCR product after electrophoresis of 5 μl of a 50-μl reaction volume was chosen for amplification of the cDNA library.

Pooled PCR products from four PCRs produced about 1.5 to 2.0 μg of DNA. The PCR products were then cut with the tagging enzyme BsmFI in the presence of a 0.2 mM concentration of each dNTP, and the 3′ recessed ends were extended by addition of T4 polymerase to produce blunt ends. The 51-bp linker-tag constructs were isolated from a 3% agarose gel and ligated together to produce linker-ditag-linker products. Removal of contaminating biotinylated Heel primers from BsmFI-restricted cDNA 3′-end fragments was found to be essential to the preparation of a high-quality SAGE library. Incorporation of amino linkers at the 5′ ends of P1 and P2 primers ensures proper orientation during ligation of the blunt-end fragments.

Ditag amplification, digestion, ligation, and cloning were carried out as described in the original SAGE protocol, with two modifications. Amplification of the ditags was carried out with biotinylated P1 and P2 primers to facilitate removal of linkers, as described by Powell (23). Concatemers were size selected for fragments larger than 400 bp, using Sephacryl S-400 Size Select columns. Several clones were picked and amplified to determine insert size (Fig. 2). Approximately 10,000 tags from each of the SAGE libraries were sequenced. Library statistics are presented in Table 2.

Evaluation of SAGE library. Reverse transcription (RT)-PCR amplification of gene transcripts represented by the eight randomly selected tags is shown in Fig. 3. Expression of each of the eight genes was also verified by RT-PCR using total RNA extracted from three control cultures of toxic fish-fed *P. shumwayae* (Fig. 3). Relative expression levels of the eight genes were not determined for these cultures.

DISCUSSION

Our objective was to develop a method to identify life stage-specific gene transcripts of *P. shumwayae*. Analysis of *Pfiesteria*

TABLE 2. SAGE library statistics

Library	No. of tags		% Diversity	No. of linkers (% of total)
	Total	Unique		
ToxA	9,140	5,272	57	35 (0.38)
ToxF	11,168	6,597	59	139 (1.2)

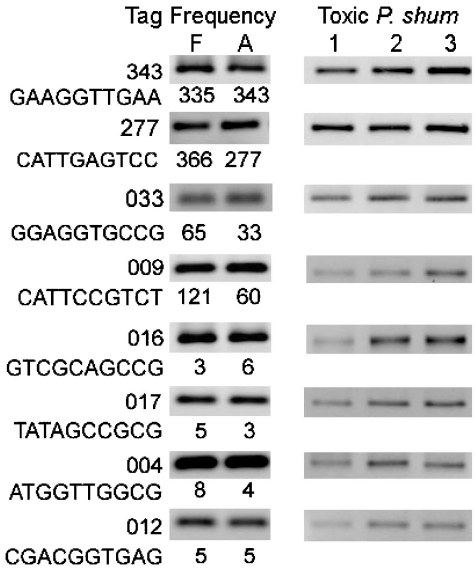

FIG. 3. Tag sequences, frequencies, and PCR amplification products from ToxA (A) and ToxF (F) cDNA. PCR products for lanes A and F were amplified from linker-ligated cDNA 3′-end fragments by using gene-specific primers. Lanes 1 to 3 are PCR products of amplification of cDNA from fish-fed toxic *P. shumwayae* (*P. shum*) control cultures 103046 (lane 1), 102988 (lane 2), and 102994 (lane 3), obtained using gene-specific primers.

gene expression is not a straightforward process. *Pfiesteria* cultures do not grow to high densities under laboratory conditions, and hazards of exposure to toxic life stages prohibit mass production of large cultures. In addition, *Pfiesteria* cannot be grown in monoculture but rather must be maintained on fish or algal prey, so that cDNA libraries are likely to include transcripts of other contaminating eukaryotes. Our approach was to construct SAGE libraries from both fish-fed and alga-fed toxic cultures. By comparing the two libraries, we could eliminate tags that are not shared and might be attributed to contaminating species and could focus on tags that appear in both libraries.

In addition to difficulties in working with toxic *Pfiesteria* cultures, the preparation of SAGE libraries presents a separate set of challenges. Evaluation of the original SAGE protocol revealed three problem areas: (i) large amounts (2 μg) of mRNA are required, (ii) the presence of contaminating linkers in the SAGE libraries results in increased sequencing costs and decreases the comprehensiveness of the library, and (iii) identification of gene transcripts from 14-bp tags is technically challenging. Although each of these obstacles has been addressed in separate publications (6, 7, 10, 14, 22), we found that the major problems in preparation of a high-quality SAGE library from small amounts of material could be resolved with a change in linker design. Use of Y linkers along with the PCR preamplification step described here produced SAGE libraries that appeared to be both comprehensive and quantitative while reducing the number of manipulations and significantly reducing the presence of contaminating linker sequences in the library.

Using the modified SAGE method presented here, we were able to generate a library from as little as 1 μg of total RNA.

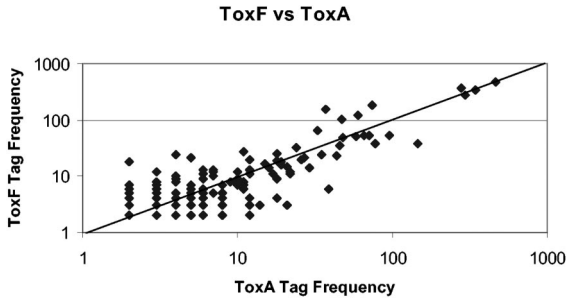

FIG. 4. Dot plot of tag frequencies for tags represented in both ToxA and ToxF SAGE libraries. Tag frequencies for each library were normalized to 10,000 tags to reduce bias attributable to library size differences. Several tags had the same frequencies; hence, each point in the plot may represent more than one tag sequence. Points above the line bisecting the libraries represent tags of greater frequency in the ToxF library. Points below the line represent tags of greater frequency in the ToxA library.

Successful preparation of SAGE libraries from a limited amount of material has previously been achieved by preamplification of full-length cDNA (19, 22) or isolated cDNA 3′-end fragments after ligation to the linkers (14). Although there is some concern about loss of representation because of PCR bias, especially for transcripts of low abundance, it was demonstrated that preamplification at low cycle numbers does not result in significant distortion of transcript levels (14, 19). A scatter plot (Fig. 4) of frequencies for tags that appear in both ToxF and ToxA libraries suggests that SAGE libraries produced using our modified protocol are also quantitative (R^2 = 0.89, slope = 1.0) and that these results would be comparable to those of other preamplification strategies. The amount of scatter, especially at low frequencies, may be in response to different food sources or physiological status and may be problematic in identification of rare tags that are specific to toxic stages of *P. shumwayae*. The increased scatter for low-frequency tags, however, would most likely be reduced by increasing the size of the SAGE library (i.e., sequencing more tags).

The comprehensiveness or depth of SAGE libraries may also be compromised by preamplification. This quality can be judged by the proportion of unique tags in the library. For the library generated using our modified SAGE approach, 57 and 59% (ToxA and ToxF, respectively) of the tags were unique (Table 2). These values are comparable to tag diversities reported for profiles generated using other SAGE protocols (23 to 68% diversity; see ftp://ftp.ncbi.nlm.nih.gov/pub/sage/OLD/extr/stats.txt), suggesting that preamplification of the cDNA did not decrease the comprehensiveness of the libraries.

In the original SAGE protocol, cDNA 3′-end fragments must be isolated from excess linkers and other cDNA fragments by performance of several successive washing steps prior to amplification. For libraries constructed from limited amounts of material, one would assume that each manipulation results in some loss of cDNA, especially for transcripts that occur in low abundance. A major advantage to the modified SAGE method is the ability to amplify cDNA 3′-end fragments without separating them from either excess linkers or 5′ cDNA fragments. This modification reduces the number of manipulations early in the protocol, increasing the probability of detecting rare transcripts. The specificity of the preamplification reaction is illustrated in Fig. 5A. No products were visible after 35 cycles of PCR amplification for reactions in which no primer (Fig. 5A, lane 7) or only a single primer (Fig. 5A, lanes 1, 2, 4, and 5) was added.

We also considered the possibility that several cDNA fragments may have ligated to the 3′ end before ligation of the Y linker. This would result in amplification of chimeric cDNA made up of fragments from several unrelated reverse-transcribed mRNA transcripts. If this were the case, restriction of the preamplification products with NlaIII would produce a shift in product size that would be visible after gel electrophoresis. Figure 5B illustrates that restriction with BsmFI followed by NlaIII resulted in an expected decrease in size of the linker-tag product but with no visible reduction in size or intensity of the higher-molecular-weight 3′-end fragments.

In the original SAGE protocol, amplified contaminating linker-linker ligation products must be separated from ditag PCR products by isolating the appropriate 102-bp band from

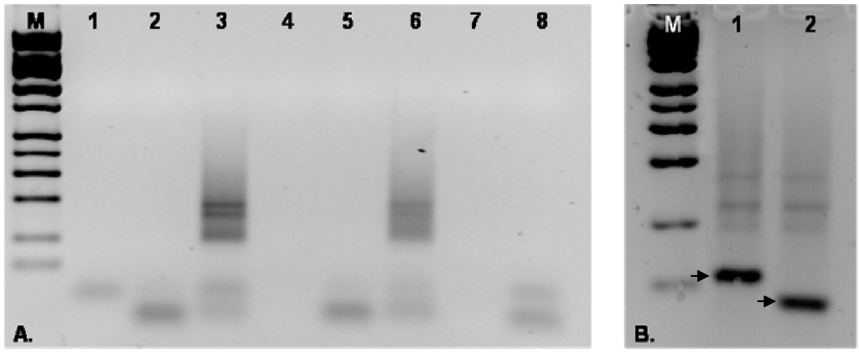

FIG. 5. (A) PCR amplification of linker-ligated cDNA, performed with one, two, or no primers. Lanes 1 to 3, PCR amplification of L1Y-ligated cDNA 3′-end fragments with P1-N primer only, Heel primer only, or both P1-N and Heel primers, respectively; lanes 4 to 6, PCR amplification of L2Y-ligated cDNA fragments with P2-N primer only, Heel primer only, or both P2-N and Heel primers, respectively; lane 7, PCR amplification of L1Y-ligated cDNA fragments with no primers added; lane 8, no-template negative control reaction. (B) PCR amplification and restriction of linker-ligated cDNA 3′-fragments. Lane 1, cDNA amplified with P1-N and Heel primers and restricted with BsmFI to yield a 55-bp linker-tag product (arrow); lane 2, same PCR product restricted first with BsmFI and then with NlaIII to yield a 41-bp linker product (arrow). M, molecular size markers.

polyacrylamide gels. Isolated ditags may then require further purification before restriction with NlaIII (2). PCR-amplified ditags generated by the modified SAGE method were present as a single 102-bp band when examined by gel electrophoresis and could be precipitated and restricted by NlaIII without gel purification. Even after purification, SAGE libraries prepared using the original protocol still include linker-linker ditag products that are indistinguishable from true (cDNA-generated) ditags until they are identified during sequencing. These contaminating ditags can represent a large proportion of the SAGE library (10), resulting in increased sequencing costs and a reduction in the size and comprehensiveness of the library. A survey of SAGE libraries deposited in GenBank revealed that the proportion of unusable tags (listed as linkers plus questionable sequences) can reach as high as 13% of the library (ftp://ftp.ncbi.nlm.nih.gov/pub/sage/OLD/extr/stats.txt). The use of Y linkers in the construction of SAGE libraries significantly reduces the fraction of contaminating linker sequences. The proportions of linker sequences in the SAGE libraries prepared here are 1.2 and 0.38% of the total tags in the ToxF and ToxA libraries, respectively (Table 2). We recently prepared a third library, for a different algal species, with only one contaminating linker sequence in over 11,000 tags (data not shown). Overall, contaminating linker sequences represent 0.56% of the total tags (over 31,000) generated in our laboratory by using the protocol presented here.

Although SAGE libraries are valuable for identifying unique or differentially expressed transcripts, additional sequence information is essential to confirm identification and to fully characterize these genes. Identification of cDNAs from the nucleotide tag sequences generated by SAGE, however, can present a problem. The short sequences may match several unrelated genes (7), or, more frequently, they may have no matches in the database (28). This is especially true for eukaryotic species with low representation in public databases, such as *Pfiesteria* dinoflagellates. Using the modified SAGE method, the 3′-end fragments of genes corresponding to the tags of interest can be easily amplified from the pool of linker-ligated fragments with primers designed from the 10-base tag–CATG linker sequences. Only a fraction of the linker-ligated cDNA pool is used in preparation of the library, leaving enough material for further analysis of at least several hundred tag sequences.

To validate the modified SAGE method, eight tags ranging in frequency from 0.026 to 3.3% of total tags in the *P. shumwayae* libraries were randomly selected for analysis. The 3′ cDNA end fragment of each transcript was amplified and sequenced using primers designed from the linker and 10-base tag sequences. Transcript-specific primers were designed, and the expression of each of the eight transcripts was further verified by RT-PCR analysis of RNA extracted from three independent cultures of toxic *P. shumwayae*. Interestingly, sequence analysis revealed some variability in length and sequence. Heterogeneity in mRNA cleavage site selection is common among plants (29) and *S. cerevisiae* (30) and was recently described for mammalian transcripts (21). In addition, Zhang and Lin (32) characterized several cotranscribed ribulose bisphosphate carboxylase/oxygenase (RUBISCO) cDNA sequences from the dinoflagellate *Prorocentrum minimum* which differed in both length and sequence of the 3′ untranslated regions (UTRs). The variability in dinoflagellate 3′ UTR sequence and polyadenylation cleavage site selection may have important implications for the study of transcript regulation in other protists.

Putative identities of the eight cDNA 3′-end fragments were obtained through BLAST (1) comparison to sequences deposited in GenBank. Tag 343 matches a sequence located upstream from and in the direction opposite of the *P. piscicida* mitochondrial gene encoding cytochrome *b* (accession no. AF357520) (15, 31). The SAGE tag for this gene was present in nearly equal frequencies in both libraries (3.00 and 3.75% for ToxF and ToxA, respectively), suggestive of constitutive expression in toxic *P. shumwayae*. A search of expressed sequence tag (EST) libraries revealed similarities to two other sequences: the sequence for tag 33 is similar to an EST library sequence for the dinoflagellate *Lingulodinium polyedrum* (GenBank accession no. CD810329), and tag 277 is similar to sequences in EST libraries prepared from the dinoflagellate species *L. polyedrum* (GenBank accession no. CD809649) and *Alexandrium tamarense* (GenBank accession no. CF751911). The presence of these transcripts in libraries from two orders of dinoflagellate taxa suggests that they may be conserved among dinoflagellate species. Although the frequencies of tags 277 and 33 in the *P. shumwayae* SAGE library indicate that they are highly expressed, the functions of these two genes are not known, and BLAST analysis of sequences similar to tags 277 and 33 in *L. polyedrum* and *A. tamarense* suggests that these sequences may be unique to dinoflagellates.

Application of the SAGE method to microbial eukaryotes can generate an enormous wealth of information. In our hands, the changes to the SAGE method described here greatly facilitated the process and resulted in gene expression profiles for *P. shumwayae* that appear to be both quantitative and comprehensive. The use of this technique for other microbial eukaryotic species is certain to enhance our understanding of cellular response to environmental stimuli and, ultimately, provide a means to assess ecosystem health.

ACKNOWLEDGMENTS

We are grateful to Howard Glasgow, Jr., of North Carolina State University's Center of Applied Aquatic Ecology, Raleigh, for valuable advice and guidance during this investigation. We thank Yaohong Zhang (University of Delaware College of Marine Studies, Lewes), Matthew Parrow, Nora Deamer-Melia, and Cheng Zhang (North Carolina State University Center of Applied Aquatic Ecology, Raleigh) for assistance with culturing and collecting material used for this project. We are also grateful to Antares Phram and Shellie Bench at Amersham Biosciences, Sunnyvale, Calif., for assistance with DNA sequencing.

This work was supported by a grant from NOAA/ECOHAB (NA 860P0495).

REFERENCES

1. **Altschul, S., T. Madden, A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Angelastro, J. M., L. P. Klimaschewski, and O. V. Vitolo.** 2000. Improved *Nla*III digestion of PAGE-purified 102 bp ditags by addition of a single purification step in both the SAGE and microSAGE protocols. Nucleic Acids Res. **28:**E62.
3. **Bowers, H. A., T. Tengs, H. B. J. Glasgow, J. M. Burkholder, P. A. Rublee, and D. W. Oldach.** 2000. Development of real-time PCR assays for rapid detection of *Pfiesteria piscicida* and related dinoflagellates. Appl. Environ. Microbiol. **66:**4641–4648.
4. **Burkholder, J. M., H. B. Glasgow, and N. Deamer-Melia.** 2001. Overview

and present status of the toxic *Pfiesteria* complex (Dinophyceae). Phycologia **40:**186–214.
5. **Burkholder, J. M., H. G. Marshall, H. B. Glasgow, D. W. Seaborn, and N. J. Deamer-Melia.** 2001. The standardized fish bioassay process for detecting and culturing actively toxic *Pfiesteria*, used by two reference laboratories for Atlantic and southeastern states. Environ. Health Perspect. **109**(Suppl. 5)**:** 745–756.
6. **Carulli, J. P., M. Artinger, P. M. Swain, C. D. Root, L. Chee, C. Tulig, J. Guerin, M. Osborne, G. Stein, J. Lian, and P. T. Lomedico.** 1998. High throughput analysis of differential gene expression. J. Cell. Biochem. Suppl. **30–31:**286–296.
7. **Chen, J. J., J. D. Rowley, and S. M. Wang.** 2000. Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. Proc. Natl. Acad. Sci. USA **97:**349–353.
8. **Chomszynski, P., and N. Sacchi.** 1987. Single step method of RNA isolation by guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. **162:**156–159.
9. **Coyne, K. J., D. A. Hutchins, C. E. Hare, and S. C. Cary.** 2001. Assessing temporal and spatial variability in *Pfiesteria piscicida* distributions using molecular probing techniques. Aquatic Microb. Ecol. **24:**275–285.
10. **Datson, N. A., J. van der Perk-de Jong, M. P. van den Berg, E. R. de Kloet, and E. Vreugdenhil.** 1999. MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acids Res. **27:**1300–1307.
11. **Glasgow, H. B., J. M. Burkholder, S. L. Morton, and J. Springer.** 2001. A second species of ichthyotoxic *Pfiesteria* (Dinamoebales, Dinophyceae). Phycologia **40:**234–245.
12. **Green, C. D., J. F. Simons, B. E. Taillon, and D. A. Lewin.** 2001. Open systems: panoramic views of gene expression. J. Immunol. Methods **250:**67–79.
13. **Kuhn, E.** 2001. From library screening to microarray technology: strategies to determine gene expression profiles and to identify differentially regulated genes in plants. Ann. Bot. **87:**139–155.
14. **Lee, S., J. J. Chen, G. L. Zhou, and S. M. Wang.** 2001. Generation of high-quantity and quality tag/ditag cDNAs for SAGE analysis. BioTechniques **31:**348–354.
15. **Lin, S., H. Zhang, D. F. Spencer, J. E. Norman, and M. W. Gray.** 2002. Widespread and extensive editing of mitochondrial mRNAs in dinoflagellates. J. Mol. Biol. **320:**727–739.
16. **Lizardi, P. M., X. Huang, Z. Zhu, P. Bray-Ward, D. C. Thomas, and D. C. Ward.** 1998. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat. Genet. **19:**225–232.
17. **Matsumura, S., S. Nirasawa, and R. Terauchi.** 1999. Technical advance: transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression. Plant J. **20:**719–726.
18. **Munasinghe, A., S. Patankar, B. P. Cook, S. L. Madden, R. K. Martin, D. E. Kyle, A. Shoaibi, L. M. Cummings, and D. J. Wirth.** 2001. Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A-T rich genomes. Mol. Biochem. Parasitol. **113:**23–34.
19. **Neilson, L., A. Andalibi, D. Kang, C. Coutifaris, J. F. Strauss, J. A. L. Stanton, and D. P. L. Green.** 2000. Molecular phenotype of the human oocyte by PCR-SAGE. Genomics **63:**13–24.
20. **Oldach, D. W., C. F. Delwiche, K. S. Jakobsen, T. Tengs, E. G. Brown, J. W. Kempton, E. F. Schaefer, H. A. Bowers, H. B. Glasgow, Jr., J. M. Burkholder, K. A. Steidinger, and P. A. Rublee.** 2000. Heteroduplex mobility assay-guided sequence discovery: elucidation of the small subunit (18S) rDNA sequences of *Pfiesteria piscicida* and related dinoflagellates from complex algal culture and environmental sample DNA pools. Proc. Natl. Acad. Sci. USA **97:**4303–4308.
21. **Pauws, E., A. H. C. van Kampen, S. A. R. van de Graaf, J. J. M. de Vijlder, and C. Ris-Stalpers.** 2001. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. Nucleic Acids Res. **29:**1690–1694.
22. **Peters, D. G., A. B. Kassam, E. Heidrich-O'Hare, H. Yonas, R. E. Ferrell, and A. Brufsky.** 1999. Comprehensive transcript analysis in small quantities of mRNA by SAGE-Lite. Nucleic Acids Res. **27:**E39.
23. **Powell, J.** 1998. Enhanced concatemer cloning—a modification to the SAGE (serial analysis of gene expression) technique. Nucleic Acids Res. **26:**3445–3446.
24. **Prashar, Y., and S. M. Weissman.** 1996. Analysis of differential gene expression by display of 3′ end restriction fragments of cDNAs. Proc. Natl. Acad. Sci. USA **93:**659–663.
25. **Suzuki, M., M. S. Rappe, and S. J. Giovannoni.** 1998. Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. Appl. Environ. Microbiol. **64:**4522–4529.
26. **Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler.** 1995. Serial analysis of gene expression. Science **270:**484–487.
27. **Velculescu, V. E., B. Vogelstein, and K. W. Kinzler.** 2000. Analysing uncharted transcriptomes with SAGE. Trends Genet. **16:**423–425.
28. **Velculescu, V. E., L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, P. Hieter, B. Vogelstein, and K. W. Kinzler.** 1997. Characterization of the yeast transcriptome. Cell **88:**243–251.
29. **Wahle, E., and W. Keller.** 1992. The biochemistry of 3′ end cleavage and polyadenylation of messenger RNA precursors. Annu. Rev. Biochem. **61:** 419–440.
30. **Wahle, E., and U. Ruegsegger.** 1999. 3′-End processing of pre-mRNA in eukaryotes. FEMS Microbiol. Rev. **23:**277–295.
31. **Zhang, H., and S. Lin.** 2002. Detection and quantification of *Pfiesteria piscicida* by using the mitochondrial cytochrome *b* gene. Appl. Environ. Microbiol. **68:**989–994.
32. **Zhang, H., and S. Lin.** 2003. Complex gene structure of the form II RUBISCO in the dinoflagellate *Prorocentrum minimum* (Dinophyceae). J. Phycol. **39:**1160–1171.