TOGA: An automated parsing technology for analyzing expression of nearly all genes

J. Gregor Sutcliffe*†, Pamela E. Foye*, Mark G. Erlander*‡, Brian S. Hilbush§, Leon J. Bodzin§, Jayson T. Durham§, and Karl W. Hasel*§

*Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037; and §Digital Gene Technologies, 11149 North Torrey Pines Road, La Jolla, CA 92037

Communicated by Floyd E. Bloom, The Scripps Research Institute, La Jolla, CA, December 10, 1999 (received for review September 10, 1999)

We have developed an automated, high-throughput, systematic cDNA display method called TOGA, an acronym for total gene expression analysis. TOGA utilizes 8-nt sequences, comprised of a 4-nt restriction endonuclease cleavage site and adjacent 4-nt parsing sequences, and their distances from the 3′ ends of mRNA molecules to give each mRNA species in an organism a single identity. The parsing sequences are used as parts of primer-binding sites in 256 PCR-based assays performed robotically on tissue extracts to determine simultaneously the presence and relative concentration of nearly every mRNA in the extracts, regardless of whether the mRNA has been discovered previously. Visualization of the electrophoretically separated fluorescent assay products from different extracts displayed via a Netscape browser-based graphical user interface allows the status of each mRNA to be compared among samples and its identity to be matched with sequences of known mRNAs compiled in databases.

Genome project-oriented studies have concentrated on the production of physical markers for refinement of genetic maps, determination of the linear sequences of chromosomes, and identification of short portions of cDNA clones of mRNAs [expressed sequence tags (ESTs)]. Ultimately, completion of these tasks may lead to identification of all of the genes and their protein products. What is needed additionally to advance biochemical understanding of physiological function is knowledge of which genes are active and to what extent in each cell type. This information could be obtained retrospectively by array-based (chip), closed-system technologies (1, 2) once a complete set of genes was identified, but such endeavors would be possible only for model organisms for which results from a completed genome sequence were available and whose genes could be identified from the genomic sequence or from massive EST programs.

Here we describe TOGA (an acronym for total gene expression analysis), an approach that utilizes sequences near the 3′ ends of mRNA molecules to give each mRNA species in the organism a single identity. The identity features are used in PCR-based assays performed on tissue extracts to determine the presence and relative concentrations of nearly every mRNA species in the extracts, regardless of whether the mRNA has been discovered previously: it is an open-system approach. Comparisons of the assay results from different extracts visualized electronically on a graphical user interface allow recognition of mRNAs whose concentrations vary depending on the source of the extract as well as those mRNAs whose concentrations do not change.

Part of the TOGA approach is conceptually related to previously described differential display and RNA arbitrarily primed (RAP)-PCR methods, which utilize mismatch pairing of oligonucleotide primers of arbitrarily chosen sequence to cDNA templates and PCR to visualize cDNA subsets (3, 4). However, the TOGA method utilizes exact primer matching and, therefore, systematically attributes each cDNA with a single, defined identity feature based on a combination of nucleotide sequence and nucleotide length, two parameters derived from the primary sequence of the mRNA. Consequently, each mRNA detected by TOGA can be easily linked electronically with other information accumulating in genome databases. In contrast to PCR-based methods that utilize a series of restriction endonuclease digestions to generate products (5, 6), the resolution provided by the TOGA parsing and visualization formats and its single-product-per-mRNA feature relieve much of the dependence on prevalence for detection of individual mRNAs, thereby allowing closure (the identification of all mRNA species) to be approached.

Methods

TOGA Method. The TOGA method (Fig. 1) is based on the observation that virtually all eukaryotic mRNAs conclude with a 3′ poly(A) tail. We prepared double-stranded cDNA (7) from poly(A)-enriched cytoplasmic RNA extracted (8) from the tissue samples of interest by using an equimolar mixture of all 48 5′ biotinylated anchor primers of the set GAATTCAACTGGAAGCGGCCGCAGGAAT$_{18}$VNN (V = A, C, or G; N = A, C, G, or T) to initiate reverse transcription. One member of this mixture of 48 primers initiates synthesis at a fixed position at the 3′ end of all copies of each mRNA species in the sample, thereby defining a 3′ endpoint for each species. We cleaved each cDNA sample with the restriction endonuclease *Msp*I, which recognizes the sequence CCGG, isolated the 3′ fragments by streptavidin Dynabead capture (Dynal, Great Neck, NY), and, after washing, released the product by digestion with *Not*I, which cleaves at an 8-nt sequence within the anchor primers but rarely within the mRNA-derived portion of the cDNAs. The 3′ *Msp*I-*Not*I fragments, which are of uniform length for each mRNA species, were directionally ligated into *Cla*I-, *Not*I-cleaved plasmid pBC SK$^+$ (Stratagene) in an antisense orientation with respect to the vector's T3 promoter, and the product was used to transform *Escherichia coli* SURE cells (Stratagene). Each library contained in excess of 1×10^6 recombinants to ensure a high likelihood that the 3′ ends of all mRNAs with concentrations of 0.001% or greater were represented multiply. Plasmid preparations (Qiagen, Chatsworth, CA) were made from the cDNA library of each sample.

An aliquot of each library was digested with *Msp*I, which effects linearization by cleavage at several sites within the parent vector while leaving the cDNA inserts and their flanking sequences, including the T3 promoter, intact. The product was

Abbreviations: TOGA, total gene expression analysis; DST, digital sequence tag; RSF, rich sequence file.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AF178682).

†To whom reprint requests should be addressed. E-mail: gregor@scripps.edu.

‡Present address: R. W. Johnson Pharmaceutical Research Institute, 3210 Merryfield Row, San Diego, CA 92121.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.040537997. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.040537997

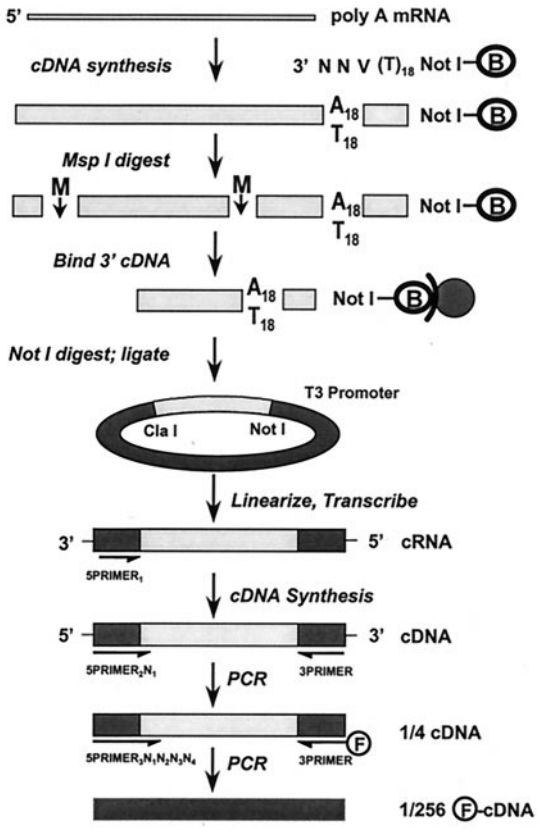

Fig. 1. Schematic view of TOGA. Poly(A)-selected RNA serves as template for double-strand cDNA synthesis by using a pool of *Not*I-containing biotinylated (B) primers degenerate in their 3′ ultimate three positions that phase the 3′ end of the cDNA at the poly(A) tail. After cleavage with *Msp*I (M), the 3′ biotinylated fragment is captured on streptavidin magnetic beads and released from the beads by digestion with *Not*I, and the 3′ *Msp*I–*Not*I fragments are cloned into an RNA expression vector in an orientation antisense to its T3 promoter. After cleavage with *Msp*I to linearize insert-containing plasmids and inactivate insertless plasmids, antisense transcripts are produced with T3 RNA polymerase. These serve, after removal of DNA template, as substrates for reverse transcriptase by using a primer that anneals to vector sequences. A PCR step with a primer that extends across the nonreconstituted *Msp*I/*Cla*I site by one of the four possible nucleotides and a universal 3′ primer subdivides the cDNA species into four pools. A subsequent PCR in the presence of a fluorescent 3′ primer and each (in separate reactions) of the 256 possible 5′ primers that extends 4 nt into the inserts subdivides the input species into 256 subpools for electrophoretic resolution.

incubated with T3 RNA polymerase (MEGAscript transcription kit; Ambion, Austin, TX) to generate antisense cRNA transcripts of the cloned inserts containing known vector sequences (tags) abutting the *Msp*I and *Not*I sites from the original cDNAs. Plasmid DNA was removed by incubation with RNase-free DNase. Each cRNA sample would be contaminated to a different extent with transcripts from insertless plasmids, which could lead to variability in the efficiency of the later PCRs for different samples because of differential competition for primers. However, the polylinker region of the parent vector contains a site for *Msp*I between its *Cla*I and *Not*I sites, and, therefore, the *Msp*I digestion step eliminated the 5′ tag from cRNAs transcribed from insertless plasmids, rendering them inert in the product-amplification steps described below.

At this stage, each of the cRNA preparations was processed in a three-step fashion. In step one, the 250 ng of cRNA was converted to first-strand cDNA by using the primer $5PRIMER_1$ (AGGTCGACGGTATCGG). In step two, 400 pg of cDNA product was used as PCR template with each of the four primers of the form $5PRIMER_2N_1$ ($5PRIMER_2$ = GGTCGACGGTATCGG; N = A, C, G, or T) paired with a universal 3′ primer 3PRIMER = GAGCTCCACCGCGGT, utilizing the program 94°C for 15 sec, 65°C for 15 sec, and 72°C for 60 sec, for 20 cycles. In step three, which was performed by a robot designed especially for the task (see below), the product of each subpool was divided further into 64 sub-subpools (2 ng in 20 μl) for a second PCR with 100 ng each of the fluorescently tagged universal 3′ primer 3PRIMER = 6-FAM-GAGCTCCACCGCGGT and the appropriate primer of the form $5PRIMER_3N_1N_2N_3N_4$ ($5PRIMER_3$ = CGACGGTATCGG; N = A, C, G, or T) by using a program (94°C for 15 sec, (X)°C for 15 sec, 72°C for 30 sec, for 30 cycles) that included an annealing step slightly above the T_m (X) of each $5PRIMER_3N_1N_2N_3N_4$ to minimize artifactual mispriming and promote high-fidelity copying. Both PCRs were performed in the presence of TaqStart antibody (CLONTECH). The products from each of the tissue samples were resolved on a series of denaturing DNA sequencing gels by using the automated ABI Prism 377 DNA sequencer. Data were collected by using the GENESCAN software package (Perkin–Elmer), fluorescence noise was removed by filtering the raw signals and applying Parzen smoothing (9), lane migrations were calibrated to the migrations of intralane standards by a two-step interpolation protocol, and amplitudes were normalized by using parameters determined on a panel-to-panel basis by interlane and intralane fluorescence-ratio statistics. Complete execution of this series of reactions generated 256 product pools for the entire $5PRIMER_3N_1N_2N_3N_4$ set for each of the cRNA samples.

Robotics. Because of quality control and throughput issues raised by the large number of PCRs used by TOGA for each mRNA sample, the potentially different temperature optima for its 256 primers, and the desirability of using the method to examine differences in mRNA expression patterns in many paradigms, we designed an automated processing station (not shown). An Orca arm (Sagian, Indianapolis, IN) controlled from a computer terminal transports disposable pipette tips and a bar-coded, 96-well reaction tray from a carousel to a Biomek liquid handling platform (Beckman), whose deck is maintained at 0°C so that reactions remain inert during their assembly. The arm collects bar-coded plates containing substrates, primers, and PCR reagents from a computer-controlled refrigerator, verifies their bar codes, and transfers the plates to the liquid handling platform, where the PCRs are assembled in batches grouped according to their thermocycling programs. The arm collects the reaction tray, places it in a sealing device to insulate each well against evaporation and contamination, identifies the bar code, and transports the tray to a thermocycler (PTC-200; MJ Research, Cambridge, MA) in which the PCR is executed under control of the computer. Upon completion of the reactions, the arm transfers the tray to the refrigerator. The arm services four thermocyclers, giving the station the capacity of producing 256 TOGA product pools from 6,000 mRNA samples per year.

Rationale. According to the TOGA scheme (Fig. 1), the cDNA libraries produced from each of the mRNA samples contain copies of the extreme 3′ ends, from the most distal site for *Msp*I to the beginning of the poly(A) tail, of nearly all poly(A)$^+$ mRNAs in the starting RNA sample approximately according to the initial relative concentrations of the mRNAs. Because both ends of the inserts for each species are defined exactly by the sequence of the mRNAs themselves, the fragment lengths are uniform for each species, allowing their later visualization as discrete bands on gels. These lengths are constant regardless of the tissue source of the mRNA. mRNAs containing no *Msp*I-recognition sequences are not represented in the product set, but

are captured by reiterating the process with additional 4-nt-recognizing restriction endonucleases (see below).

A fundamental aspect of TOGA is the use of sequences adjacent to the 3′ *Msp*I site to sort the cDNAs in successive PCR steps. The first PCR step utilizes a primer that anneals with sequences derived from pBC SK$^+$, but extends across the CGG of the nonregenerated *Msp*I site to include the first adjacent nucleotide (N_1) of the insert. This step segregates the starting population of mRNAs into four subpools. In the subsequent PCR step, in which a label is incorporated into the products for their detection by laser-induced fluorescence, each of the four pools is segregated further by division into 64 for a total of 256 sub-subpools by using more insert-invasive primers ($N_1N_2N_3N_4$). Electrophoresis resolves the molecules into distinct bands of measurable lengths. Thus, each final PCR product is assigned an identity or address based on an 8-nt sequence ($CCGGN_1N_2N_3N_4$) and the distance of that sequence to the 3′ end of the mRNA plus a known length of vector-derived sequence added during the TOGA processing. When the nucleotide sequence of a TOGA fragment, either real or generated conceptually from a database sequence, is known, the fragment is referred to as a digital sequence tag (DST), that is, a 3′ end EST derived by the TOGA process.

Graphical User Interface. To accommodate data viewing, analysis, warehousing, and integration with external databases, we designed a graphical user interface, or browser, based on the easily recognizable Netscape format. The normalized TOGA profiles for a desired comparison are assembled in graphics interchange format (GIF) files as a vertically stacked display and are accessed either from a scrollable, alphabetically arranged list of the 256 primers or by typing the desired primer sequence into a text-box entry field located between forward–backward selector buttons (Fig. 2). Profiles in the display can be set to any of four scales to emphasize fragment peaks of different amplitude ranges. A guideline placed by mouse click on the vertically aligned TOGA panels activates nucleotide length and fragment peak amplitude measurements, which appear instantaneously below the panels. A save button can be clicked to record a selected peak and its corresponding length and amplitude measurements to a user log file, creating hyperlinks to the original data. Standard nucleotide sequence manipulation tools in browser-compatible format can be accessed from scrollable lists above the panels.

Once a TOGA fragment peak is observed, the user may inquire whether it might correspond to the DST for any known mRNA species. Integrated into the browser display are entries in public databases that have been sorted by species (i.e., human, mouse, etc.), processed to identify mRNA sequences that have a high probability of extending to their 3′ ends, scanned to locate their most 3′ *Msp*I sites, and parsed according to the adjacent 4 nt to generate a DST for each mRNA. The resulting virtual DSTs, updated regularly, are mapped by length (including the vector-derived sequences added during TOGA processing) above the appropriate TOGA panels as small hash marks and also appear in a scrollable list at the base of the display, arranged by length, accession number, and entry name, and hyperlinked to the GenBank rich sequence file (RSF). Thus, a TOGA product that migrates under a hash mark has a candidate identity; those under no hash mark represent probable novel mRNAs. Further information about candidate DSTs can be obtained by following hyperlinks in the RSF to citations and related publications or by BLAST searches initiated by clicking a hyperlink embedded within the RSF.

Results

Method Validation: Reproducibility and Specificity. The details of the present TOGA protocol were developed from a preliminary systematic display methodology (10, 11) by execution of a series of pilot studies in which primer lengths, template dilutions, number of PCR cycles, annealing conditions, and other parameters were varied and automation was implemented. To demonstrate the efficacy of the method, we isolated polyadenylated RNA from a human osteosarcoma cell line grown in culture and either deprived of serum for 24 hr to cause mitotic arrest or supplemented after the 24-hr arrest period for 4 hr with serum

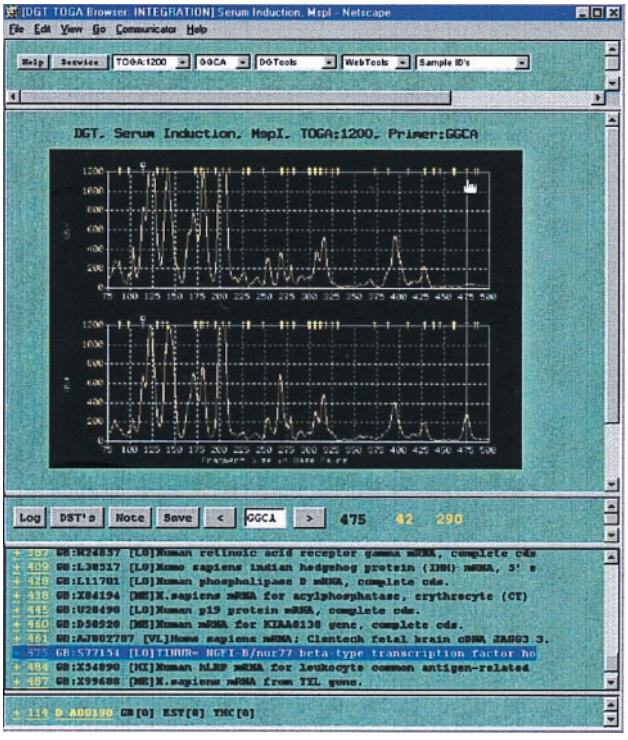

Fig. 2. TOGA data are viewed through a Netscape-based browser. The normalized TOGA profiles appear below five pull-down menus, from left to right: (*i*) Display Type lists four selections for the *y*-axis amplitude scale, with ranges from 0 to 600, 1,200, 2,000, or 8,000 normalized fluorescence units (the 1,200 scale is that selected); (*ii*) Primer lists the 256 permutations of the $N_1N_2N_3N_4$ nucleotides (GGCA is that selected; this selection is also indicated below the display flanked by forward–backward selector buttons for navigating through the alphabetical list and again above the trace display in a title region that also specifies collaborator, experiment name, enzyme identity, and the selected *y*-axis scale); (*iii*) DGTools contains selections for searching public and private sequence databases by using BLAST, GAP-BLAST, or the Keyword search algorithm, for retrieval using the GCG FETCH program, for comparing a selected group of sequences using PILEUP, and for translation of a nucleotide sequence into the putative proteins encoded by its different reading frames (all of these operations are carried out behind a security firewall); (*iv*) Web Tools gives access to programs and databases on the Internet that require navigation beyond the firewall, including PubMed for medline literature searches and the National Center for Biotechnology Information for searching gene databases; (*v*) Sample ID's lists the displayed sample identities (here, human osteosarcoma cells deprived of or replenished with serum). A guideline has been placed by mouse on a peak. The length (475 nt) and fragment peak amplitudes (42 and 290 fluorescence units) measurements then automatically appear below the panels. Selection of the save button records the selected peak and its corresponding length and amplitude measurements to a user log file, creating hyperlinks to the original data. Access to that log file is via the button ''log.'' Below the trace display are the virtual DSTs derived by processing human entries in public databases, arranged by length, and including accession number and entry name and hyperlinked to the GenBank RSF. These virtual DSTs are also mapped above the TOGA traces as yellow hash marks. The 475-nt differentially represented peak selected by the guideline falls under such a hash mark. Consultation of the virtual DST list suggests a candidate identity with TINUR, an NGFI-B/nur77 homologue, which was validated by extended primer PCR and direct nucleotide sequence analysis of the excised fragment.

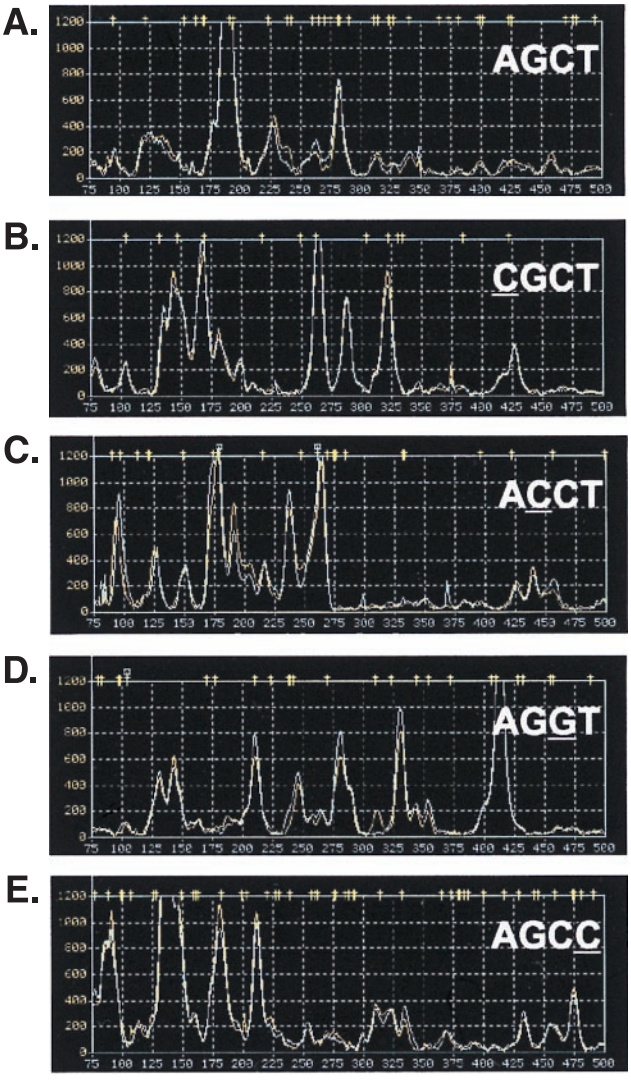

Fig. 3. TOGA reproducibility and primer specificity. Product pool profiles generated by AGCT and single nucleotide variants at each of the four locations within $N_1N_2N_3N_4$; the varied nucleotide is underlined. The 1,200-unit amplitude scale has been selected for the *y* axis. Each trace shows the overlay of the product profiles from two independent libraries generated from the same RNA sample prepared from serum-depleted human osteosarcoma cells.

and a protein synthesis inhibitor, anisomycin, to cause reentry into mitosis and superinduction of the mRNAs transcribed from immediate early genes (12). The RNA samples were divided, and replicate libraries and sets of TOGA products were prepared on different days by using different thermocyclers of the robotic processing station. We observed 20–40 products per lane with the 256 primers. No products were detected if the $5PRIMER_3N_1N_2N_3N_4$ was omitted from the second PCR step (not shown). Only selected data will be discussed here; complete analysis of the data set and the details of the growth conditions of the cell line will be reported separately.

When the pairs of noncontemporaneous replicate traces were superimposed, they were nearly indistinguishable (Fig. 3), demonstrating the reproducibility of the procedures. For each different primer, the array of products appeared to be distinct. To determine the degree of specificity in the series of 5PRIMER-dependent reactions, we examined the products generated with $5PRIMER_3N_1N_2N_3N_4$s that differed at single positions. Single-nucleotide substitutions at either the N_1, N_2, N_3, or N_4 position (compare Fig. 3 *A* with *B*, *C*, *D*, and *E*, respectively) resulted in apparently independent product arrays, with very few comigrating products as would be expected merely by chance, suggesting that high fidelity primer binding in the two PCR steps was efficiently parsing the cDNA subsets.

Using TOGA. The major utility of TOGA is for comparing mRNA expression profiles for two or more tissue samples. We compared the effect of the serum starvation/replenishment experiment on the panels of products generated by TOGA. The majority of products from the pair of samples comigrated and were of comparable amplitudes. Fewer than 10% of the products had amplitudes that differed by a factor of 2 or more, and these were distributed approximately equally between species induced and species repressed by serum replenishment.

Many products, some of which were differentially represented in the two panels, appeared to migrate in positions coincident with DST hash marks extracted from GenBank and, thus, had candidate identities: for example, the fragments indicated by the guidelines in Figs. 2 and 4*A* fell under human DST icons corresponding to TINUR, an NGFI-B/nur77-related orphan receptor, and NF-κB, a transcription factor, both known immediate early genes. To test these candidate identities, oligonucleotides (GATCGAATCCGGGGCATCCTGGATTTAGAA and GATCGAATCCGGCCCGCCTGAATCATTCTC) were synthesized corresponding to the $5PRIMER_3N_1N_2N_3N_4$ for each sub-subpool extended 3′ with an additional 14 nt from the sequences adjacent to the $N_1N_2N_3N_4$ nucleotides in the GenBank sequences for human TINUR and NF-κB, respectively. These were paired with the fluorescent 3PRIMER in PCRs by using the matching TOGA N_1 cDNA pools as substrates. For each of these two examples, a single product was generated that comigrated with the candidate TOGA products and exhibited a differential distribution similar to that of the corresponding TOGA product (Fig. 4*B* shows the NF-κB result). This simple extended primer experiment confirmed the identities of the two candidates. To obtain independent confirmation of the TINUR and NF-κB identities, we excised each product from preparative separation gels run in parallel with the TOGA analytical gel, eluted the candidate cDNA fragments, amplified them by PCR with 3PRIMER and its appropriate $5PRIMER_3N_1N_2N_3N_4$, and cloned the products in the TA cloning vector (Invitrogen). The nucleotide sequences of the inserts were determined by using primers that annealed to flanking vector sequences and were found to be identical to those of the DSTs from GenBank. Thus, TOGA sorted these known mRNAs properly according to the sequences adjacent to their most 3′ *Msp*I sites.

The expression of NF-κB in the serum starvation/replenishment paradigm was tested independently by performing Northern blot analyses on the same RNAs used to prepare the original TOGA samples. Differential RNA expression values comparable to those suggested by the TOGA data were observed (Fig. 4*C* and Table 1). The concentrations of this mRNA in the presence and absence of serum were measured at 6.5 and 0.7 parts in 100,000, indicating that such concentrations are easily within the range of detection by the TOGA process.

Many TOGA products, including some that were differentially represented in the two experimental samples, did not fall under DST hash marks, suggesting that these corresponded to RNAs not in GenBank: an example, OS51, is shown in Fig. 4*E*. We determined the nucleotide sequence of the excised OS51 product (accession no. AF178682) and found it to be novel in BLAST searches of GenBank and dbEST databases. Extended primer analysis based on the newly determined sequence generated a band that comigrated with and exhibited a differential distribution similar to that of the corresponding TOGA product (Fig. 4*F*). Northern blot analysis using the isolated fragment as probe (Fig. 4*G* and Table 1) demonstrated that the novel mRNA was

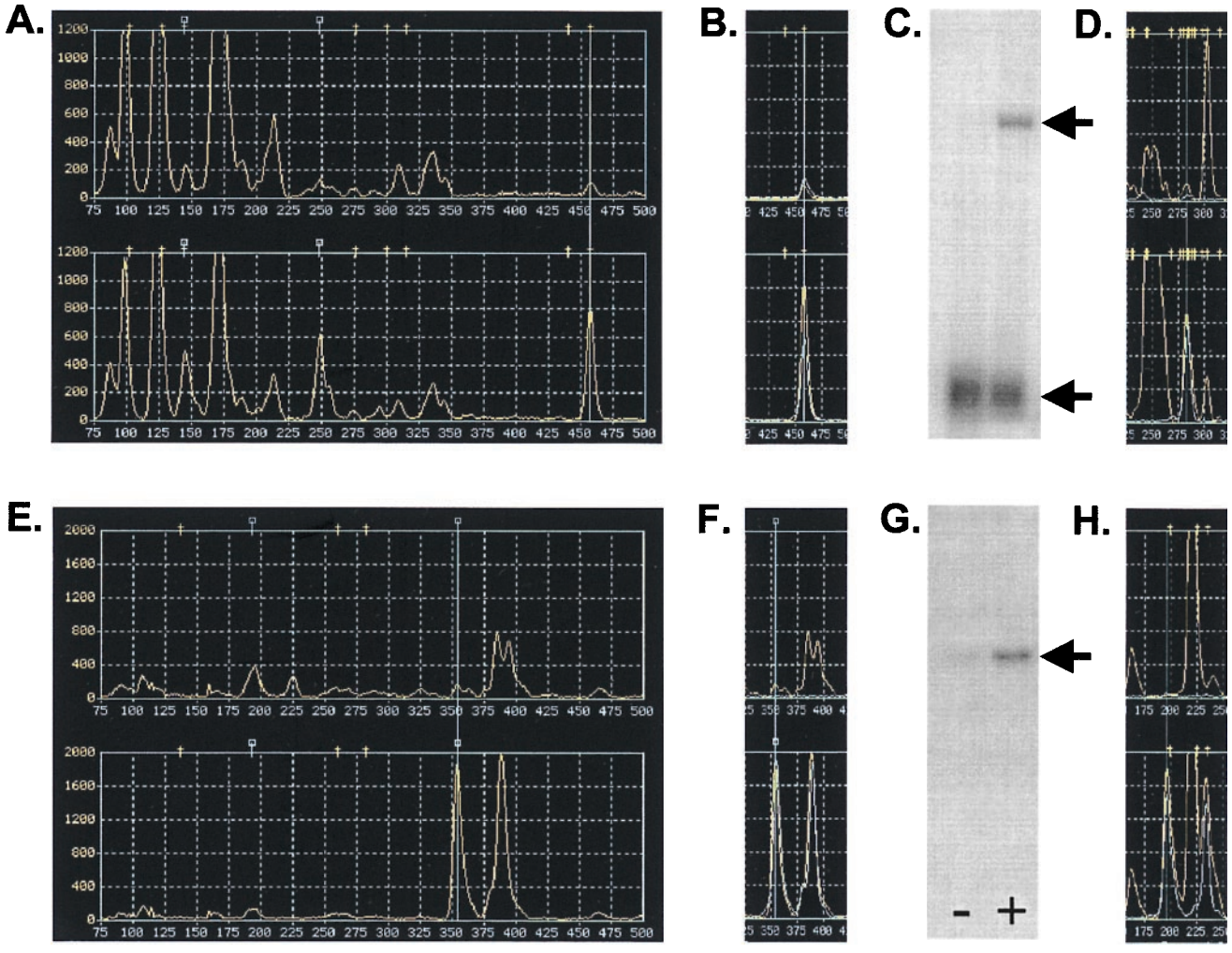

Fig. 4. TOGA candidate verification, fragment capture, and validation. Pool profiles generated from serum-depleted (*Upper*) and serum-replenished (*Lower*) cells by $N_1N_2N_3N_4$ primers CCCG (*A*) and ATCG (*E*) in the *Msp*I vector; amplitude scales are 1,200 and 2,000, respectively. In *A*, *B*, *D–F*, and *H*, the guideline is placed over a peak of greater amplitude in the trace from the serum-replenished sample. In *A*, these peaks comigrate with hash marks corresponding to the virtual DST for NF-κB; in *E*, there was no comigrating DST (square hash marks represent nonpublic DSTs). In *B* and *F*, the region of the traces containing the highlighted peaks is shown overlaid with the products generated by the 14-nt extended primers specific for NF-κB and OS51, respectively, synthesized based on the sequences in the corresponding RSFs (accession nos. M58603 and AF178682). The OS51 extension primer was GATCGAATCCGATCGCTCCTGGTTCCTCGGA. Northern blots for NF-κB (*C*) and OS51 (*G*) were prepared from RNA samples from osteosarcoma cells depleted of (−) or replenished with (+) serum. The lower band in *C* represents the mRNA for ribosomal protein S20, which was used as a normalization standard. Arrows indicate the relevant blot bands. In *D* and *H*, the relevant local regions of the corresponding *Sau*3AI vector panels for each RNA is shown. In *E*, *F*, and *H*, note the alternative poly(A) variant that migrates as 40 nt longer than the OS51 DST.

differentially represented in the experimental RNA samples in the manner suggested by the TOGA measurements and had mRNA concentrations of 0.17 and 3.74 parts per 100,000 before and after serum treatment, respectively. Thus, TOGA detects and properly sorts both known and previously undescribed mRNAs of abundances at least as low as 1 part in 100,000 and gives reliable measurements of their relative concentrations. Some mRNA species might be amplified disproportionately by the TOGA steps, but such putative effects should be uniform across related samples such that the intensity of a band in one sample can be compared with its intensity in others. Interestingly, the extended primer experiment additionally revealed a product 40 nt longer than the original OS51 product that was coordinately regulated with OS51. Sequence analysis demonstrated this to be an alternative poly(A) site selection variant of the OS51 mRNA.

The TOGA program described above would not encounter the products of every gene once and only once. Some mRNAs use alternative polyadenylation sites or contain internal strings of A residues, and these would give rise to more than one fragment with the panel of 256 primers; hence, these genes would be scored more than once. Assuming random nucleotide distribution in the 3′ ends of mRNAs, approximately 25% will not contain a target sequence in their 3′ 500 nt for digestion by *Msp*I; hence, their PCR-amplified fragments will be too large to resolve discretely on sequencing gels and would go unsurveyed if only a single restriction enzyme were used to generate 3′ end fragments. There are other mRNAs that will not be scored: those whose TOGA products comigrate with other fragments. For displays with 30 products, mostly of low amplitude, resolved over 500 nt, approximately 15% would migrate more closely than 3 nt to another fragment, making their discrimination as distinct products problematical. Cumulatively, these considerations suggest that approximately 60% of the mRNAs expressed at a concentration of greater than 1 part in 100,000 would be scored in one pass of the method. To approach actual closure, we investigated the use of additional restriction endonucleases.

Table 1. Measurements of product abundance

	*Msp*I			*Sau*3AI			Northern		
DST	−S	+S	+/−	−S	+S	+/−	−S	+S	+/−
NF-κB	111	764	6.9	211	1,330	6.3	0.70	6.5	9.3
OS51	159	1,880	11.8	35	1,820	52	0.17	3.74	22.0

The peak amplitudes of the two DSTs under study from the serum-depleted (−S) and serum-replenished (+S) samples were determined from the *Msp*I and *Sau*3AI browser panels, and their ratios (+/−) were calculated. The Northern blots contained, in addition to the 2-μg poly(A)-enriched RNA samples from the serum-depleted or serum-replenished cells, a dilution series of cRNAs produced from the isolated, cloned DSTs. The intensities of the hybridization signals were compared to determine the mass ratio, expressed as parts per 100,000.

We repeated the serum starvation/replenishment study by using *Sau*3AI rather than *Msp*I to generate 3′ cDNA fragments and cloned these between the *Bam*HI and *Not*I sites in pDGT4, a pBC SK$^+$-derived vector engineered to contain a site for *Bam*HI in its polylinker. We repeated the TOGA process by using primers with sequence differences to accommodate the enzyme substitution and obtained a comparable panel of TOGA results. Using their nucleotide sequences, we predicted *Sau*3AI DSTs for the NF-κB and OS51 mRNAs and examined the appropriate addresses for these DSTs. In each case (Fig. 4 *D* and *H*), we observed a fragment pair whose amplitude ratio (Table 1) was similar to that for its corresponding *Msp*I DST ratio. The alternative poly(A) variant of OS51 also was observed. We excised each of the fragments and found its nucleotide sequence to match that of the predicted DST. This *Sau*3AI iteration of TOGA allows an independent sample of approximately 60% of the mRNAs to be detected. Iterations with four different restriction endonucleases would be required to score 98% of the mRNAs at least once.

Discussion

The TOGA methodology sorts mRNAs on the basis of the identity of an 8-nt sequence (for *Msp*I, $CCGGN_1N_2N_3N_4$) and the distance of that sequence from the poly(A) tail. Because amplification depends on the nucleotide sequence of an mRNA and not its prevalence in a given tissue, the method, in principal, allows for the accounting of nearly all mRNAs present at concentrations above its detection threshold (closure). The automation of TOGA, made possible by the simplicity of its concept, enables a program of systematic and rapid accumulation of data about expression of mRNAs. The browser, also made possible by the simplicity of the principle, allows instantaneous assessments as to the potential identities or novelty of differentially expressed mRNAs detected by the process.

We thank Jeanette Quan, Wendy Erickson-Hirons, Melissa Almazan, Sharon Simons, William Strauss, Serena Nguyen, Patria Danielson, and Marie Callahan for excellent technical assistance and Dennis Grace for work on fluorescence normalization parameters. This work was supported in part by National Institutes of Health Grants NS33396 and GM32355 and Digital Gene Technologies.

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270,** 467–470.
2. Lockhart, D. J., Dong H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., *et al.* (1996) *Nat. Biotechnol.* **14,** 1675–1680.
3. Liang, P. & Pardee, A. B. (1992) *Science* **257,** 967–971.
4. Welsh, J., Chada, K., Dalal, S. S., Cheng, R., Ralph D. & McClelland, M. (1992) *Nucleic Acids Res.* **20,** 4965–4970.
5. Prashar, Y. & Weissman, S. M. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 659–663.
6. Shimkets, R. A., Lowe, D. G., Tai, J. T.-N., Sehl, P., Jin, H., Yang, R., Predki, P. F., Rothberg, B. E. G., Murtha, M. T., Roth, M. E., *et al.* (1999) *Nat. Biotechnol.* **17,** 798–803.
7. Gubler, U. & Hoffman, B. (1983) *Gene* **25,** 263–269.
8. Schibler, K., Tosi, M., Pittet, A. C., Fabiani, L. & Wellauer, P. K. (1980) *J. Mol. Biol.* **142,** 93–116.
9. Harris, F. J. (1978) *Proc. Inst. Electric Electronics Eng.* **66,** 51–83.
10. Erlander, M. G., Dopazo, A., Foye, P. E. & Sutcliffe, J. G. (1994) in *Identification of Transcribed Sequences,* eds. Hochgeschwender, U. & Gardiner, K. (Plenum, New York), pp. 261–271.
11. Sutcliffe, J. G. & Erlander, M. G. (1995) U.S. Patent 5,459,037; divisional 5,807,680.
12. Cochran, B. H., Reffel, A. C. & Stiles, C. D. (1983) *Cell* **33,** 939–947.