Method For Indexing And Determining The Relative Concentration Of Expressed Messenger RNAs

(MBHB No. 98,429-A)

BACKGROUND OF THE INVENTION

This invention is directed to methods for simultaneous identification of differentially expressed mRNAs, as well as measurements of their relative concentrations.

A complete characterization of the protein molecules that make up an organism would be useful, e.g. for the improved design of drugs, the selection of optimal treatment of individual patients, and for the development of more compatible biomaterials. Such a characterization of expressed proteins would include their identification, sequence determination, demonstration of their anatomical sites of expression, elucidation of their biochemical activities, and understanding of how these activities determine organismic physiology. For medical applications, the description should also include information about how the concentration of each protein changes in response to pharmaceutical or toxic agents.

Let us consider the scope of the problem: How many genes are there? The issue of how many genes are expressed in a mammal is still unsettled after at least two decades of study. There are few direct studies that address patterns of gene expression in different tissues. Mutational load studies (J.O. Bishop, "The Gene Numbers Game," Cell 2:81-86 (1974); T. Ohta & M. Kimura, "Functional Organization of Genetic Material as a Product of Molecular Evolution," Nature 223:118-119 (1971)) have suggested that there are between $3x10^4$ and 10^5 essential genes.

Before cDNA cloning techniques, information on gene expression came from RNA complexity studies: analog measurements (measurements in bulk) based on observations of mixed populations of RNA molecules with different specificities in abundances. To an unexpected extent, early analog complexity studies were distorted by hidden complications of the fact that the molecules in each tissue that make up

most of its mRNA mass comprise only a small fraction of its total complexity. Later, cDNA cloning allowed digital measurements (i.e., sequence-specific measurements on individual species) to be made; hence, more recent concepts about mRNA expression are based upon actual observations of individual RNA species.

Brain, liver, and kidney are the mammalian tissues that have been most extensively studied by analog RNA complexity measurements. The lowest estimates of complexity are those of Hastie and Bishop (N.D. Hastie & J. B. Bishop, "The Expression of Three Abundance Classes of Messenger RNA in Mouse Tissues," Cell 9:761-774 (1976)), who suggested that $26x10^6$ nucleotides of the $3x10^9$ base pair rodent genome were expressed in brain, $23x10^6$ in liver, and $22x10^6$ in kidney, with nearly complete overlap in RNA sets. This indicates a very minimal number of tissue-specific mRNAs. However, experience has shown that these values must clearly be underestimates, because many mRNA molecules, which were probably of abundances below the detection limits of this early study, have been shown to be expressed in brain but detectable in neither liver nor kidney. Many other researchers (J.A. Bantle & W.E. Hahn, "Complexity and Characterization of Polyadenylated RNA in the Mouse Brain," Cell 8:139-150 (1976); D.M. Chikaraishi, "Complexity of Cytoplasmic Polyadenylated and Non-Adenylated Rat Brain Ribonucleic Acids," Biochemistry 18:3249-3256 (1979)) have measured analog complexities of between $100-200x10^6$ nucleotides in brain, and 2-to-3-fold lower estimates in liver and kidney. Of the brain mRNAs, 50-65% are detected in neither liver nor kidney. These values have been supported by digital cloning studies (R.J. Milner & J.G. Sutcliffe, "Gene Expression in Rat Brain," Nucl. Acids Res. 11:5497-5520 (1983)).

Analog measurements on bulk mRNA suggested that the average mRNA length was between 1400-1900 nucleotides. In a systematic digital analysis of brain mRNA length using 200 randomly selected brain cDNAs to measure RNA size by northern blotting (Milner & Sutcliffe, supra), it was found that, when the mRNA size data were weighted for RNA prevalence, the average length was 1790 nucleotides, the same as that determined by analog measurements. However, the mRNAs that made up most of the brain mRNA complexity had an average length of 5000 nucleotides. Not only were the rarer brain RNAs longer, but they tended to be brain specific, while

the more prevalent brain mRNAs were more ubiquitously expressed and were much shorter on average.

These concepts about mRNA lengths have been corroborated more recently from the length of brain mRNA whose sequences have been determined (J.G. Sutcliffe, "mRNA in the Mammalian Central Nervous System," Annu. Rev. Neurosci. 11:157-198 (1988)). Thus, the 1-2x10^8 nucleotide complexity and 5000-nucleotide average mRNA length calculates to an estimated 30,000 mRNAs expressed in the brain, of which about 2/3 are not detected in liver or kidney. Brain apparently accounts for a considerable portion of the tissue-specific genes of mammals. Most brain mRNAs are expressed at low concentration. There are no total-mammal mRNA complexity measurements, nor is it yet known whether 5000 nucleotides is a good mRNA-length estimate for non-neural tissues. A reasonable estimate of total gene number might be between 50,000 and 100,000.

What is most needed to advance by a chemical understanding of physiological function is a menu of protein sequences encoded by the genome plus the cell types in which each is expressed. At present, protein sequences can be reliably deduced only from cDNAs, not from genes, because of the presence of intervening sequences (introns) in the genomic sequences. Even the complete nucleotide sequence of a mammalian genome will not substitute for characterization of its expressed sequences. Therefore, a systematic strategy for collecting transcribed sequences and demonstrating their sites of expression is needed. Such a strategy would be of particular use in determining sequences expressed differentially within the brain. It is necessarily an eventual goal of such a study to achieve closure; that is, to identify all mRNAs. Closure can be difficult to obtain due to the differing prevalence of various mRNAs and the large number of distinct mRNAs expressed by many distinct tissues. The effort to obtain it allows one to obtain a progressively more reliable description of the dimensions of gene space.

Studies carried out in the laboratory of Craig Venter (M.D. Adams et al., "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project," Science 252:1651-1656 (1991); M.D. Adams et al., "Sequence Identification

of 2,375 Human Brain Genes," Nature 355:632-634 (1992)) have resulted in the isolation of randomly chosen cDNA clones of human brain mRNAs, the determination of short single-pass sequences of their 3'-ends, about 300 base pairs, and a compilation of some 2500 of these as a database of "expressed sequence tags." This database, while useful, fails to provide any knowledge of differential expression. It is therefore important to be able to recognize genes based on their overall pattern of expression within regions of brain and other tissues and in response to various paradigms, such as various physiological or pathological states or the effects of drug treatment, rather than simply their expression in a single tissue.

Other work has focused on the use of the polymerase chain reaction (PCR) to establish a database. Williams et al. (J.G.K. Williams et al., "DNA Polymorphisms Amplified by Arbitrary Primers Are Useful as Genetic Markers," Nucl. Acids Res. 18:6531-6535 (1990)) and Welsh & McClelland (J. Welsh & McClelland, "Genomic Fingerprinting Using Arbitrarily Primed PCR and a Matrix of Pairwise Combinations of Primers," Nucl. Acids Res. 18:7213-7218 (1990) showed that single 10-mer primers of arbitrarily chosen sequences, i.e., any 10-mer primer off the shelf, when used for PCR with complex DNA templates such as human, plant, yeast, or bacterial genomic DNA, gave rise to an array of PCR products. The priming events were demonstrated to involve incomplete complementarity between the primer and the template DNA. Presumably, partially mismatched primer-binding sites are randomly distributed through the genome. Occasionally, two of these sites in opposing orientation were located closely enough together to give rise to a PCR product band. There were on average 8-10 products, which varied in size from about 0.4 to about 4 kb and had different mobilities for each primer. The array of PCR products exhibited differences among individuals of the same species. These authors proposed that the single arbitrary primers could be used to produce restriction fragment length polymorphism (RFLP)-like information for genetic studies. Others have applied this technology (S.R. Woodward et al., "Random Sequence Oligonucleotide Primers Detect Polymorphic DNA Products Which Segregate in Inbred Strains of Mice," Mamm. Genome 3:73-78 (1992); J.H. Nadeau et al., "Multilocus Markers for Mouse Genome Analysis: PCR Amplification Based on Single Primers of Arbitrary Nucleotide Sequence," Mamm. Genome 3:55-64 (1992)).

Two groups (J. Welsh et al., "Arbitrarily Primed PCR Fingerprinting of RNA," Nucl. Acids Res. 20:4965-4970 (1992); P. Liang & A.B. Pardee, "Differential Display of Eukaryotic Messenger RNA by Means of the Polymerase Chain Reaction," Science 257:967-971 (1992)) adapted the method to compare mRNA populations. In the study of Liang and Pardee, this method, called mRNA differential display, was used to compare the population of mRNAs expressed by two related cell types, normal and tumorigenic mouse A31 cells. For each experiment, they used one arbitrary 10-mer as the 5′-primer and an oligonucleotide complementary to a subset of poly A tails as a 3′ anchor primer, performing PCR amplification in the presence of ^{35}S-dNTPs on cDNAs prepared from the two cell types. The products were resolved on sequencing gels and 50-100 bands ranging from 100-500 nucleotides were observed. The bands presumably resulted from amplification of cDNAs corresponding to the 3′-ends of mRNAs that contain the complement of the 3′ anchor primer and a partially mismatched 5′ primer site, as had been observed on genomic DNA templates. For each primer pair, the pattern of bands amplified from the two cDNAs was similar, with the intensities of about 80% of the bands being indistinguishable. Some of the bands were more intense in one or the other of the PCR samples; a few were detected in only one of the two samples.

Further studies (P. Liang et al., "Distribution and Cloning of Eukaryotic mRNAs by Means of Differential Display: Refinements and Optimization," Nucl. Acids Res. 21:3269-3275 (1993)) have demonstrated that the procedure works with low concentrations of input RNA (although it is not quantitative for rarer species), and the specificity resides primarily in the last nucleotide of the 3′ anchor primer. At least a third of identified differentially detected PCR products correspond to differentially expressed RNAs, with a false positive rate of at least 25%.

If all of the 50,000 to 100,000 mRNAs of the mammal were accessible to this arbitrary-primer PCR approach, then about 80-95 5′ arbitrary primers and 12 3′ anchor primers would be required in about 1000 PCR panels and gels to give a likelihood, calculated by the Poisson distribution, that about two-thirds of these mRNAs would

be identified.

It is unlikely that all mRNAs are amenable to detection by this method for the following reasons. For an mRNA to surface in such a survey, it must be prevalent enough to produce a signal on the autoradiograph and contain a sequence in its 3' terminus 500 nucleotides capable of serving as a site for mismatched primer binding and priming. The more prevalent an individual mRNA species, the more likely it would be to generate a product. Thus, prevalent species may give bands with many different arbitrary primers. Because this latter property would contain an unpredictable element of chance based on selection of the arbitrary primers, it would be difficult to approach closure by the arbitrary primer method. Also, for the information to be portable from one laboratory to another and reliable, the mismatched priming must be highly reproducible under different laboratory conditions using different PCR machines, with the resulting slight variation in reaction conditions. As the basis for mismatched priming is poorly understood, this is a drawback of building a database from data obtained by the Liang & Pardee differential display method.

U.S. Patents Numbers 5,459,037 ('037) and 5,807,680 ('680) describe an improved method of differential display of mRNA species that reduces the uncertain aspect of 5'-end generation and allows data to be absolutely reproducible in different settings. The method does not depend on potentially irreproducible mismatched priming, reduces the number of PCR panels and gels required for a complete survey, and allows double-strand sequence data to be rapidly accumulated. Furthermore, the improved method also reduces the number of concurrent signals obtained from the same species of mRNA. The '037 and '680 patents are hereby incorporated by reference as part of this disclosure.

There remains a need for further improvements of the method disclosed in the '037 patent. For example, the specificity of the method could be improved by decreasing mispriming during the synthesis of complimentary DNA molecules and during PCR reactions. Furthermore, the technique could be further refined so that it is more reproducible, more sensitive and easier to use. Preferably the technique would

provide the ability to use sequences obtained to form databases, and to scan nucleotide data bases such as GenBank to recognize sequence identities and similarities using computer programs such as BLASTN and BLASTX.

SUMMARY

We have developed an improved method for the simultaneous sequence-specific identification of mRNAs in a mRNA population. The improved method sorts mRNAs on the basis of an identity or address determined by 1) a partial nucleotide sequence of length $a + b$, where a is the length in bases of the restriction endonuclease recognition site and b is the number of parsing bases, where $6 \geq b \geq 3$, and 2) the distance of that partial sequence from the poly(A) tail. Typically the identity or address is determined by a partial sequence that includes a four base recognition site for a restriction endonuclease and four parsing bases. In one preferred embodiment, the recognition site for a restriction endonuclease is MspI, and the partial sequence is C-C-G-G-N_1-N_2-N_3-N_4. Because it is dependent upon the nucleotide sequence of an mRNA and not its prevalence in a given tissue, the method can account for all mRNAs present at concentrations above its detection threshold. In contrast to differential display and RAP-PCR methodologies, there is no uncertain aspect to the generation of 5' ends.

According to one preferred embodiment of the method of the present invention (Figure 1), the cDNA libraries produced from each of the mRNA samples contain copies of the extreme 3' ends, from the most distal site for MspI 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 exactly defined 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, an important fundamental concept of the approach. Messenger RNAs lacking MspI-recognition sequences are not represented, but these are relatively rare. These mRNAs are captured by applying the method using a different restriction endonuclease that recognizes a different four base recognition

sequence.

Another aspect of such embodiments of the present invention is the use of sequences adjacent to the 3' restriction endonuclease site, in one preferred embodiment, a MspI site, to sort the cDNAs in at least two successive PCR steps. The first PCR step utilizes a primer that anneals with sequences derived from the vector, e.g., pBC SK$^+$, but extends across the CGG of the non-regenerated MspI site to include the first adjacent nucleotide (N_1) of the insert. This step segregates the starting population of mRNAs into 4 subpools. In a second PCR step, each of the 4 subpools produced by the first PCR step is further segregated by division into 64 for a total of 256 subsubpools by using more insert-invasive primers ($N_1N_2N_3N_4$). A fluorescent label is incorporated into the products for their detection by laser-induced fluorescence by using fluorescent labeled 3'PCR primers in the final PCR step.

In a preferred embodiment, a separation technique such as electrophoresis is used to resolve the labeled molecules of the PCR product into distinct bands of measurable intensities and corresponding to measurable lengths. Suitable separation techniques include gel electrophoresis, capillary electrophoresis, HPLC, MALDI mass spectroscopy and other suitable separations techniques known in the art that are capable of single base resolution over the range of 50 – 500 bases are encompassed by the present invention.

In one preferred embodiment, each final PCR reaction product is thus assigned an identity or address based upon an 8-nucleotide sequence including the four base restriction endonuclease site plus four parsing bases (e.g., C-C-G-G-N_1-N_2-N_3-N_4) and the distance of that sequence from the junction between the end of the message and the first A of the polyA tail at the 3' end of the mRNA. When the nucleotide sequence of a PCR product fragment, either experimentally determined or determined from a database sequence, is known, the fragment is referred to as a digital sequence tag (DST): that is, a 3'-end EST (expressed sequence tag) derived by the method of the present invention. The intensity of the separated band of labeled PCR product fragments, detected using an appropriate method, preferably laser-induced fluorescence (but radioactive or magnetic labeling and detection may be used) is

quantified and stored for each PCR product fragment in a database with the address assigned for that PCR product fragment. The intensity of the separated band of labeled PCR product fragments is proportional to the starting amount of mRNA corresponding to that PCR product fragment.

In general, the method of the present invention comprises:

(a) preparing a double-stranded cDNA population from an mRNA population using a mixture of anchor primers, each anchor primer having a 5' terminus and a 3' terminus and including: (i) a tract of from 7 to 40 T residues; (ii) a site for cleavage by a first restriction endonuclease that recognizes more than six bases, the site for cleavage being located towards the 5'-terminus relative to the tract of T residues; (iii) a first stuffer segment of from 4 to 40 nucleotides, the first stuffer segment being located towards the 5'-terminus relative to the site for cleavage by the first restriction endonuclease; (iv) a second stuffer segment interposed between the site for cleavage by a first restriction endonuclease that recognizes more than six bases and the tract of T residues, and (v) phasing residues located at the 3' terminus of each of the anchor primers selected from the group consisting of -V, -V-N, and -V-N-N, wherein V is a deoxyribonucleotide selected from the group consisting of A, C, and G; and N is a deoxyribonucleotide selected from the group consisting of A, C, G, and T, the mixture including anchor primers containing all possibilities for V and N;

(b) cleaving the double-stranded cDNA population with the first restriction endonuclease and a second restriction endonuclease, the second restriction endonuclease recognizing a four-nucleotide sequence, to form a population of double-stranded cDNA molecules having first and second termini, respectively;

(c) inserting each double-stranded cDNA molecule from step (b) into a vector in an orientation that is antisense with respect to a bacteriophage-specific promoter within the vector to form a population of constructs containing the inserted cDNA molecules, thereby defining 5' and 3' flanking vector sequences adjacent to the 5' terminus of the sense strand of the inserted cDNA and the 3' terminus of the sense strand respectively, and said constructs having a 3' flanking vector sequence at least 15 nucleotides in length between said first restriction endonuclease site and a site

defining transcription initiation in said promoter;

(d) Transforming a host cell with the vector into which the cleaved cDNA has been inserted to produce vectors containing cloned inserts;

(e) generating linearized fragments containing the inserted cDNA molecules by digestion of the constructs produced in step (c) with at least one restriction endonuclease that does not recognize sequences in either the inserted cDNA molecules or in the bacteriophage-specific promoter, but does recognize sequences in the vector, such that the resulting linearized fragments have a 5' flanking vector sequence of at least 15 nucleotides into the vector 5' to the double-stranded cDNA molecule's second terminus;

(f) generating a cRNA preparation of antisense cRNA transcripts by incubating the linearized fragments with a bacteriophage-specific RNA polymerase capable of initiating transcription from the bacteriophage-specific promoter;

(g) generating first-strand cDNA by transcribing the cRNA using a reverse transcriptase and a 5' RT primer being 15 to 30 nucleotides in length and comprising a nucleotide sequence that is complementary to the 5' flanking vector sequence;

(h) generating a first set of PCR products by dividing the first-strand cDNA into a first series of subpools and using the first-strand cDNA as templates for a first polymerase chain reaction with a first 3' PCR-primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a first 5' PCR-primer defined as having a 3'-terminus consisting of $-N_1$, wherein "N" is one of the four deoxyribonucleotides A, C, G, or T, the first 5' PCR-primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the first 5' PCR-primer's complementarity extending into one nucleotide of the insert-specific nucleotides of the cRNA, wherein a different one of the first 5' PCR primers is used in each of four different subpools;

(i) generating a second set of PCR products by further dividing the first set of PCR products in each of the first series of subpools into a second series of subpools and using the first set of PCR products as templates for a second polymerase chain reaction with a second 3' PCR primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a second 5' PCR primer defined as having a 3'-terminus consisting of $-N_1-N_x$, wherein N_1 is identical to the N_1 used in the first polymerase chain reaction for that subpool, "N" is as in step (h), and "x" is an integer from 1 to 5, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the primer's complementarity extending across into the insert-specific nucleotides of the cRNA in a number of nucleotides equal to "x" +1, wherein a different one of the second 5' PCR primers is used in different subpools of the second series of subpools and wherein there are 4^X subpools in the second series of subpools for each of the subpools in the first set of subpools;

(j) resolving the second set of PCR products to generate a display of sequence-specific products representing the 3'-ends of mRNAs present in the mRNA population.

In one preferred embodiment, a biotin moiety is conjugated to the anchor primers, preferably to the 5' terminus of the anchor primers. In such an embodiment, the first restricted cDNA is separated from the remainder of the cDNA in step (b) by contacting the first restricted cDNA with a streptavidin-coated substrate. Suitable streptavidin-coated substrates include microtitre plates, PCR tubes, polystyrene beads, paramagnetic polymer beads and paramagnetic porous glass particles. A preferred streptavidin-coated substrate is a suspension of paramagnetic polymer beads (Dynal, Inc., Lake Success, NY).

In one embodiment, the 3 nucleotides at the 3' end of the first 5' PCR primer are joined by phosophodiesterase-resistant linkages, preferably phosphorothioate linkages. In a further embodiment, the 3 nucleotides at the 3' end of the second 5' PCR primer are joined by phosophodiesterase-resistant linkages, preferably

phosphorothioate linkages. Preferably, the 3 nucleotides at the 3' end of both the first and second 5' PCR primers are joined by phosphorothioate linkages.

Typically, one of the primers for the second PCR reaction is conjugated to a fluorescent label. A suitable fluorescent label is selected from the group consisting of spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 6-carboxylic acid, 3',6'-dihydroxy-6-carboxyfluorescein (6-FAM, ABI); spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 5-carboxylic acid, 3',6'-dihydroxy-5-carboxyfluorescein (5-FAM, Molecular Probes); spiro(isobenzofuran-1(3H), 9'-(9H)-xanthen)-3-one, 3',6'-dihydroxy-fluorescein (FAM, Molecular Probes); 9-(2,5-dicarboxyphenyl)-3,6- bis(dimethylamino)-xanthylium (6-carboxytetramethylrhodamine (6-TAMRA), Molecular Probes); 3,6-diamino-9-(2-carboxyphenyl)-xanthylium (Rhodamine Green™, Molecular Probes); spiro[isobenzofuran-1(3H), 9'-xanthene]-6-carboxylic acid,5'-dichloro-3',6'-dihydroxy-2',7'-dimethoxy-3-oxo-(JOE, Molecular Probes); 1H,5H,11H,15H-xantheno[2,3,4-ij:5,6,7-i'j']diquinolizin- 8-ium, -(2,4-disulfophenyl)-2,3,6,7,12,13,16,17-octahydro-, inner salt (Texas Red, Molecular Probes); 6-((4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-s-indacene-3-propionyl) amino) hexanoic acid (BODIPY FL-X, Molecular Probes); 6-((4,4-difluoro-1,3-dimethyl-5-(4-methoxyphenyl)-4-bora-3a,4a-diaza-s-indacene-3-propionyl)amino)hexanoic acid (BODIPY TMR-X, Molecular Probes); 6-(((4-(4,4-difluoro-5-(2-thienyl)-4-bora-3a,4a-diaza-s-indacene-3-yl) phenoxy)acetyl) amino)-hexanoic acid (BODIPY TR-X, Molecular Probes); 4,4-difluoro-4-bora-3a,4a-diaza-s-indacene-3-pentanoic acid (BODIPY FL-C_5, Molecular Probes); 4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-s-indacene-3-propanoic acid (BODIPY FL, Molecular Probes); 4,4-difluoro-5-phenyl-4-bora-3a,4a-diaza-s-indacene-3-propionic acid (BODIPY 581/591, Molecular Probes); 4,4-difluoro-5-(4-phenyl-1,3-butadienyl)-4-bora-3a,4a-diaza-s-indacene-3-

propionic acid (BODIPY 564/570, Molecular Probes);
4,4-difluoro-5-styryl-4-bora-3a,4a-diaza-s-indacene-3-propionic acid;
6-(((4,4-difluoro-5-(2-thienyl)-4-bora-3a,4a-diaza-s-indacene-3-yl)styryloxy)acetyl) aminohexanoic acid (BODIPY 630/650, Molecular Probes);
6-(((4,4-difluoro-5-(2-pyrrolyl)-4-bora-3a,4a-diaza-s-indacene-3-yl) styryloxy)acetyl) aminohexanoic acid (BODIPY 650/665, Molecular Probes); and
9-(2,4(or 2,5)-dicarboxyphenyl)-3,6- bis(dimethylamino)- xanthylium, inner salt (TAMRA, Molecular Probes). Other suitable fluorescent labels, including 4, 7, 2', 4', 5', 7' hexachloro 6-carboxyfluorescein ("HEX," ABI), "NED" (ABI) and 4, 7, 2', 7' tetrachloro 6-carboxyfluorescein ("TET," ABI) are known in the art.

Typically, the phasing residues in step (a) have a 3' terminus of -V-N-N. In other embodiments, the phasing residues in step (a) have a 3' terminus of –V or –V-N.

In a preferred embodiment, the "x" in step (i) is 3. Preferably, the phasing residues in step (a) are -V-N-N and the "x" in step (i) is 3.

Typically, the anchor primers each have from 8 to 18 T residues in the tract of T residues. In one preferred embodiment, the anchor primers each have 18 T residues in the tract of T residues. In other embodiments, the anchor primers each have from 8 to 18 T residues, preferably from 8 to 16 T residues, more preferably from 8 to 14 T residues, most preferably from 8 to 12 T residues, in the tract of T residues. In another preferred embodiment, the anchor primers each have 12 T residues in the tract of T residues.

Typically, the first stuffer segment of the anchor primers is 14 residues in length. In one embodiment, the first stuffer segment has the nucleotide sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C (SEQ ID NO: 1). In a preferred embodiment, the first stuffer segment has the nucleotide sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A (SEQ ID NO: 2).

Typically, the bacteriophage-specific promoter is selected from the group consisting of T3 promoter, T7 promoter and SP6 promoter. Preferably, the bacteriophage-specific promoter is T3 promoter.

In one embodiment, the primer for priming of transcription of cDNA from cRNA has the sequence A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G (SEQ ID NO: 14). In another embodiment, the primer for priming of transcription of cDNA from cRNA has the sequence A-G-C-T-C-T-G-T-G-G-T-G-A-G-G-A-T-C (SEQ ID NO: 28). In further embodiment, the primer for priming of transcription of cDNA from cRNA has the sequence T-C-G-A-C-T-G-T-G-G-T-G-A-G-C-A-T-G (SEQ ID NO: 35).

In one embodiment, the vector is the plasmid pBC SK+ cleaved with ClaI and NotI and the 3' PCR primer in steps (h) and (i) is G-A-G-C-T-C-C-A-C-C-G-C-G-T (SEQ ID NO: 47). In another embodiment, the vector is the plasmid pBC SK+ cleaved with ClaI and NotI and the 3' PCR primer in steps (h) and (i) is G-A-G-C-T-C-G-T-T-T-T-C-C-C-C-A-G (SEQ ID NO: 48).

Typically, the first restriction endonuclease that recognizes more than six bases is selected from the group consisting of AscI, BaeI, FseI, NotI, PacI, PmeI PpuMI, RsrII, SapI, SexAI, SfiI, SgfI, SgrAI, SrfI, Sse8387I and SwaI. A preferred first restriction endonuclease that recognizes more than six bases is NotI.

Typically, the second restriction endonuclease recognizing a four-nucleotide sequence is selected from the group consisting of MboI, DpnII, Sau3AI, Tsp509I, HpaII, BfaI, Csp6I, MseI, HhaI, NlaIII, TaqI, MspI, MaeII and HinPII. Preferred second restriction endonucleases recognizing a four-nucleotide sequence are MspI, Sau3AI and NlaIII.

Typically, the restriction endonuclease used in step (e) has a nucleotide sequence recognition that includes the four-nucleotide sequence of the second restriction endonuclease used in step (b). In one embodiment, the second restriction endonuclease is MspI and the restriction endonuclease used in step (e) is Sma I. In

another embodiment, the second restriction endonuclease is TaqI and the restriction endonuclease used in step (e) is XhoI. In an alternative embodiment, the second restriction endonuclease is HinPII and the restriction endonuclease used in step (e) is NarI. In yet another embodiment, the second restriction endonuclease is MaeII and the restriction endonuclease used in step (e) is AatII.

Typically, the vector of step (c) is in the form of a circular DNA molecule having first and second vector restriction endonuclease sites flanking a vector stuffer sequence, and further comprising the step of digesting the vector with restriction endonucleases that cleave the vector at the first and second vector restriction endonuclease sites. Preferably, the vector stuffer sequence includes an internal vector stuffer restriction endonuclease site between the first and second vector restriction endonuclease sites.

One suitable host cell is Escherichia coli.

Typically, step (e) includes digestion of the vector with a restriction endonuclease which cleaves the vector at the internal vector stuffer restriction endonuclease site.

Typically, the restriction endonuclease used in step (e) also cleaves the vector at the internal vector stuffer restriction endonuclease site.

For other restriction endonucleases, a general scheme for linearizing a pSK vector without a suitable restriction endonuclease having a six base recognition site containing an internal four base recognition site comprises: (i) dividing the plasmid containing the insert into two fractions, a first fraction cleaved with the restriction endonuclease XhoI and a second fraction cleaved with the restriction endonuclease SalI; (ii) recombining the first and second fractions after cleavage; (iii) dividing the recombined fractions into thirds and cleaving the first third with the restriction endonuclease HindIII, the second third with the restriction endonuclease BamHI, and the third with the restriction endonuclease EcoRI; and (iv) recombining the thirds after digestion in order to produce a population of linearized fragments of which about one-

sixth of the population corresponds to the product of cleavage by each of the possible combinations of enzymes.

Typically, the mRNA population has been enriched for polyadenylated mRNA species.

Typically, the resolving of the amplified fragments in step (j) is conducted by electrophoresis to display the products. Preferably, the intensity of products displayed after electrophoresis is about proportional to the abundances of the mRNAs corresponding to the products in the original mixture. In a preferred embodiment, the method further comprises a step of determining the relative abundance of each mRNA in the original mixture from the intensity of the product corresponding to that mRNA after electrophoresis.

Typically, the step of resolving the polymerase chain reaction amplified fragments by electrophoresis comprises electrophoresis of the fragments on multiple gels.

In one embodiment of the invention, the method further comprises the steps of:

(k) eluting at least one cDNA corresponding to a mRNA from an electropherogram in which bands representing the 3'-ends of mRNAs present in the sample are displayed;

(l) amplifying the isolated PCR product in a polymerase chain reaction;

(m) cloning the amplified isolated PCR product into a plasmid;

(n) producing DNA corresponding to the cloned isolated PCR product from the plasmid; and

(o) sequencing the cloned isolated PCR product.

Another embodiment of the present invention comprises the steps of:

(a) isolating an mRNA population;

(b) preparing a double-stranded cDNA population from the mRNA population using a mixture of anchor primers, each anchor primer having a 5' terminus and a 3'

terminus and including: (i) a tract of from 7 to 40 T residues; (ii) a site for cleavage by a first restriction endonuclease that recognizes more than six bases, the site for cleavage being located towards the 5'-terminus relative to the tract of T residues; (iii) a first stuffer segment of from 4 to 40 nucleotides, the first stuffer segment being located towards the 5'-terminus relative to the site for cleavage by the first restriction endonuclease; (iv) a second stuffer segment interposed between the site for cleavage by a first restriction endonuclease that recognizes more than six bases and the tract of T residues, and (v) phasing residues –V-N-N located at the 3' terminus of each of the anchor primers, wherein V is a deoxyribonucleotide selected from the group consisting of A, C, and G; and N is a deoxyribonucleotide selected from the group consisting of A, C, G, and T, the mixture including anchor primers containing all possibilities for V and N;

(c) cleaving the double-stranded cDNA population with the first restriction endonuclease and a second restriction endonuclease, the second restriction endonuclease recognizing a four-nucleotide sequence, to form a population of double-stranded cDNA molecules having first and second termini, respectively;

(d) inserting each double-stranded cDNA molecule from step (b) into a vector in an orientation that is sense with respect to a T3 promoter within the vector to form a population of constructs containing the inserted cDNA molecules, thereby defining 5' and 3' flanking vector sequences adjacent to the 5' terminus of the sense strand of the inserted cDNA and the 3' terminus of the sense strand respectively, and said constructs having a 5' flanking vector sequence at least 15 nucleotides in length between said second restriction endonuclease site and a site defining transcription initiation in said promoter:

(e) transforming Escherichia coli with the vector into which the cleaved cDNA has been inserted to produce vectors containing cloned inserts;

(f) generating linearized fragments containing the inserted cDNA molecules by digestion of the constructs produced in step (c) with at least one restriction endonuclease that does not recognize sequences in either the inserted cDNA molecules or in the T3 promoter;

(g) generating a cRNA preparation of sense cRNA transcripts by incubating the linearized fragments with a T3 RNA polymerase capable of initiating transcription from the T3 promoter;

(h) generating first-strand cDNA by transcribing the cRNA using a reverse transcriptase and a 3' RT primer being 15 to 30 nucleotides in length and comprising a nucleotide sequence that is complementary to the 3' flanking vector sequence;

(i) generating a first set of PCR products by dividing the first-strand cDNA into a first series of subpools and using the first-strand cDNA as templates for a first polymerase chain reaction with a first 3' PCR-primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences 3' to the first restriction endonuclease site and a first 5' PCR-primer defined as having a 3'-terminus consisting of $-N_1$, wherein "N" is one of the four deoxyribonucleotides A, C, G, or T, the first 5' PCR-primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the first 5' PCR-primer's complementarity extending into one nucleotide of the insert-specific nucleotides of the cRNA, wherein a different one of the first 5' PCR primers is used in each of four different subpools;

(j) generating a second set of PCR products by further dividing the first set of PCR products in each of the first series of subpools into a second series of subpools and using the first set of PCR products as templates for a second polymerase chain reaction with a second 3' PCR primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector 3' to the first restriction endonuclease site and a second 5' PCR primer defined as having a 3'-terminus consisting of $-N_1-N_x$, wherein N_1 is identical to the N_1 used in the first polymerase chain reaction for that subpool, "N" is as is step (i), and "x" is an integer selected from the group consisting of 3 and 4, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the primer's complementarity extending across into the insert-specific nucleotides of the cRNA in a number of nucleotides equal to "x" =1, wherein a different one of the second 5' PCR primers is used in different subpools of the second series of subpools and wherein there are 4^X subpools in the second series of subpools

(k) resolving the second set of PCR products to generate a display of sequence-specific products representing the 3'-ends of mRNAs present in the mRNA population.

Typically, the mixture of 48 anchor primers has the sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-

T-T-T-T-T-V-N-N (SEQ ID NO: 5). In a preferred embodiment, the mixture of 48 anchor primers has the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 8).

Typically, the mixture of 12 anchor primers has the sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 4). In a preferred embodiment, the mixture of 12 anchor primers has the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 7).

Typically, the mixture of 3 anchor primers has the sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 3). In a preferred embodiment, the mixture of 3 anchor primers has the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 6).

In a preferred embodiment, the first restriction endonuclease is MspI and the second restriction endonuclease is NotI.

Typically, the first 5' PCR-primer is G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N (SEQ ID NO: 22).

In a preferred embodiment, the 3'PCR primer in the second polymerase chain reaction is the nucleotide of SEQ ID NO: 47 conjugated to a fluorescent label, more preferably, the nucleotide of SEQ ID NO: 47 conjugated to 6-FAM.

Suitable values of "x" in step (i) are integers from 1 to 5. Preferably, the "x" in step (i) is 3.

Typically, a method for detecting a change in the pattern of mRNA expression in a tissue associated with a physiological or pathological change comprising the steps of:

(a) obtaining a first sample of normal or neoplastic tissue that is not subject to

This Sheet Has Been Replaced



the physiological or pathological change;

(b) isolating an mRNA population from the first sample;

(c) determining the pattern of mRNA expression in the first sample of the tissue by performing steps (a)-(j) of the general method to generate a first display of sequence-specific products representing the 3'-ends of mRNAs present in the first sample;

(d) obtaining a second sample of the tissue that has been subject to the physiological or pathological change;

(e) isolating an mRNA population from the second sample;

(f) determining the pattern of mRNA expression in the second sample of the tissue by performing steps (a)-(j) of the general method to generate a second display of sequence-specific products representing the 3'-ends of mRNAs present in the second sample; and

(g) comparing multiple displays to determine the effect of the physiological or pathological change on the pattern of mRNA expression in the tissue.

Typically more than two samples are compared. In preferred embodiments 3, more preferably at least 4, samples are taken at multiple times and compared.

Typically, the physiological or pathological change is selected from the group consisting of Alzheimer's disease, parkinsonism, ischemia, alcohol addiction, drug addiction, schizophrenia, amyotrophic lateral sclerosis, multiple sclerosis, depression, and bipolar manic-depressive disorder.

Typically, the physiological or pathological change is associated with learning or memory, emotion, glutamate neurotoxicity, feeding behavior, olfaction, vision, movement disorders, viral infection, electroshock therapy, the administration of a drug or the toxic side effects of drugs.

Typically, the physiological or pathological change is selected from the group consisting of circadian variation, aging, and long term potentiation. In general, the physiological or pathological change is selected from processes mediated by transcription factors, intracellular second messengers, hormones, neurotransmitters,

growth factors and neuromodulators. Alternatively, the physiological or pathological change is selected from processes mediated by cell-cell contact, cell-substrate contact, cell-extracellular matrix contact and contact between cell membranes and cytoskeleton.

Preferably, the normal or neoplastic tissue comprises cells taken or derived from an organ or organ system selected from the group consisting of the cardiovascular system, the lymphatic system, the respiratory system, the digestive system, the peripheral nervous system, the central nervous system, the enteric nervous system, the endocrine system, the integument (including skin, hair and nails), the skeletal system (including bone and muscle), the urinary system and the reproductive system.

In preferred embodiments, the normal or neoplastic tissue comprises cells taken or derived from the group consisting of epithelia, endothelia, mucosa, glands, blood, lymph, connective tissue, cartilage, bone, smooth muscle, skeletal muscle, cardiac muscle, neurons, glial cells, spleen, thymus, pituitary, thyroid, parathyroid, adrenal cortex, adrenal medulla, adrenal cortex, pineal, skin, hair, nails, teeth, liver, pancreas, lung, kidney, bladder, ureter, breast, ovary, uterus, vagina, testes, prostate, penis, eye and ear.

Typically, the normal or neoplastic tissue is derived from a structure within the central nervous system selected from the group consisting of retina, cerebral cortex, olfactory bulb, thalamus, hypothalamus, anterior pituitary, posterior pituitary, hippocampus, nucleus accumbens, amygdala, striatum, cerebellum, brain stem, suprachiasmatic nucleus, and spinal cord.

Typically, a method of detecting a difference in action of a drug to be screened and a known compound comprising the steps of-

(a) obtaining a first sample of tissue from an organism treated with a compound of known physiological function;

(b) isolating an mRNA population from the first sample;

(c) determining the pattern of mRNA expression in the first sample of the

tissue by performing steps (a)-(j) of the general method to generate a first display of sequence-specific products representing the 3'-ends of mRNAs present in the first sample;

(d) obtaining a second sample of tissue from an organism treated with a drug to be screened for a difference in action of the drug and the known compound;

(e) isolating an mRNA population from the first sample;

(f) determining the pattern of mRNA expression in the second sample of the tissue by performing steps (a)-(j) of the general method to generate a second display of sequence-specific products representing the 3'-ends of mRNAs present in the second sample; and

(g) comparing the first and second displays in order to detect the presence of mRNA species whose expression is not affected by the known compound but is affected by the drug to be screened, thereby indicating a difference in action of the drug to be screened and the known compound.

Typically, the drug to be screened is selected from the group consisting of antidepressants, neuroleptics, tranquilizers, anticonvulsants, monoamine oxidase inhibitors, stimulants, anti-parkinsonism agents, skeletal muscle relaxants, analgesics, local anesthetics, cholinergics, antiviral agents, antispasmodics, steroids, and non-steroidal anti-inflammatory drugs.

More generally, the terms "drug to be screened" and "drug to be tested" are used herein to refer to a broad class of useful chemical and therapeutic agents including physiologically active steroids, antibiotics, antifungal agents, antibacterial agents, antineoplastic agents, analgesics and analgesic combinations, anorexics, anthelmintics, antiarthritics, antiasthia agents, anticonvulsants, antidepressants, antidiabetic agents, antidiarrheals, antihistamines, anti-inflammatory agents, antimigraine preparations, antimotion sickness preparations, antinauseants, antiparkinsonism drugs, antipruritics, antipsychotics, antipyretics, antispasmodics, including gastrointestinal and urinary; anticholinergics, sympathomimetics, xanthine derivatives, cardiovascular preparations including calcium channel blockers, betablockers, antiarrhythmics, antihypertensives diuretics, vasodilators including general, coronary, peripheral and cerebral; central nervous system stimulants, cough and cold preparations, decongestants, hormones, hypnotics, immunosuppressives, muscle relaxants, parasympatholytics,

parasympathomimetics, psychostimulants, sedatives, tranquilizers, allergens, antihistaminic agents, anti-inflammatory agents, physiologically active peptides and proteins, ultraviolet screening agents, perfumes, insect repellents, hair dyes, and the like. The term "physiologically active" in describing the agents contemplated herein is used in a broad sense to comprehend not only agents having a direct pharmacological effect on the host but also those having an indirect or observable effect which is useful in the medical arts, e.g., the coloring or opacifying of tissue for diagnostic purposes, the screening of ultraviolet radiation from the tissues and the like.

For instance, typical fungistatic and fungicidal agents include thiabendazole, chloroxine, amphotericin, candicidin, fungimycin, nystatin, chlordantoin, clotrimazole, ethonam nitrate, miconazole nitrate, pyrrolnitrin, salicylic acid, fezatione, ticlatone, tolnaftate, triacetin, zinc, pyrithione and sodium pyrithione.

Steroids include cortisone, cortodoxone, fluoracetonide, fludrocortisone, difluorsone diacetate, flurandrenolone acetonide, medrysone, amcinafel, amcinafide, betamethasone and its esters, chloroprednisone, clorcortelone, descinolone, desonide, dexamethasone, dichlorisone, difluprednate, flucloronide, flumethasone, flunisolide, fluocinonide, flucortolone, fluoromethalone, fluperolone, fluprednisolone, meprednisone, methylmeprednisone, paramethasone, prednisolone and predisone.

Antibacterial agents include sulfonamides, penicillins, cephalosporins, penicillinase, erythromycins, linomycins, vancomycins, tetracyclines, chloramphenicols, streptomycins, and the like. Specific examples of antibacterials include erythromycin, erythromycin ethyl carbonate, erythromycin estolate, erythromycin glucepate, erythromycin ethylsuccinate, erythromycin lactobionate, lincomycin, clindamycin, tetracycline, chlortetracycline, demeclocycline, doxycycline, methacycline, oxytetracycline, minocycline, and the like.

Peptides and proteins include, in particular, small to medium-sized peptides, e.g., insulin, vasopressin, oxytocin, growth factors, cytokines as well as larger proteins such as human growth hormone.

Other agents encompass a variety of therapeutic agents such as the xanthines, triamterene and theophylline, the antitumor agents, 5-fluorouridinedeoxyriboside, 6-mercaptopurinedeoxyriboside, vidarabine, the narcotic analgesics, hydromorphone, cyclazine, pentazocine, bupomorphine, the compounds containing organic anions, heparin, prostaglandins and prostaglandin-like compounds, cromolyn sodium, carbenoxolone, the polyhydroxylic compounds, dopamine, dobutamine, l-dopa, a-methyldopa, angiotensin antagonists, polypeptides such as bradykinin, insulin, adrenocorticotrophic hormone (ACTH), enkephalins, endorphins, somatostatin, secretin and miscellaneous compounds such as tetracyclines, bromocriptine, lidocaine, cimetidine or any related compounds.

Other agents include iododeoxyuridine, podophyllin, theophylline, isoproterenol, triamcinolone acetonide, hydrocortisone, indomethacin, phenylbutazone paraaminobenzoic acid, aminopropionitrile and penicillamine.

The foregoing list is by no means intended to be exhaustive, and any physiologically active agent may be tested by the method of the present invention.

Typically, a database is constructed comprising the data produced by the quantitation of the display of sequence-specific PCR products. Typically, the database further comprises data concerning sequence relationships, gene mapping and cellular distributions.

In one embodiment, the invention provides a method for recognizing sequence identities and similarities between the sequence of 3'-ends of mRNA molecules present in a sample and a database of sequences, comprising the steps of:

(a) preparing a double-stranded cDNA population from an mRNA population using a mixture of anchor primers, each anchor primer having a 5' terminus and a 3' terminus and including: (i) a tract of from 7 to 40 T residues; (ii) a site for cleavage by a first restriction endonuclease that recognizes more than six bases, the site for cleavage being located towards the 5'-terminus relative to the tract of T residues; (iii) a first stuffer segment of from 4 to 40 nucleotides, the first stuffer segment being located towards the 5'-terminus relative to the site for cleavage by the

first restriction endonuclease; (iv) a second stuffer segment interposed between the site for cleavage by a first restriction endonuclease that recognizes more than six bases and the tract of T residues, and (v) phasing residues located at the 3' terminus of each of the anchor primers selected from the group consisting of -V, -V-N, and -V-N-N, wherein V is a deoxyribonucleotide selected from the group consisting of A, C, and G; and N is a deoxyribonucleotide selected from the group consisting of A, C, G, and T, the mixture including anchor primers containing all possibilities for V and N;

(b) cleaving the double-stranded cDNA population with the first restriction endonuclease and a second restriction endonuclease, the second restriction endonuclease recognizing a four-nucleotide sequence, to form a population of double-stranded cDNA molecules having first and second termini, respectively;

(c) inserting each double-stranded cDNA molecule from step (b) into a vector in an orientation that is antisense with respect to a bacteriophage-specific promoter within the vector to form a population of constructs containing the inserted cDNA molecules, thereby defining 5' and 3' flanking vector sequences adjacent to the 5' terminus of the sense strand of the inserted cDNA and the 3' terminus of the sense strand respectively, and said constructs having a 3' flanking vector sequence at least 15 nucleotides in length between said first restriction endonuclease site and a site defining transcription initiation in said promoter;

(d) transforming a host cell with the vector into which the cleaved cDNA has been inserted to produce vectors containing cloned inserts;

(e) generating linearized fragments containing the inserted cDNA molecules by digestion of the constructs produced in step (c) with at least one restriction endonuclease that does not recognize sequences in either the inserted cDNA molecules or in the bacteriophage-specific promoter, but does recognize sequences in the vector, such that the resulting linearized fragments have a 5' flanking vector sequence of at least 15 nucleotides into the vector 5' to the double-stranded cDNA molecule's second terminus;

(f) generating a cRNA preparation of antisense cRNA transcripts by incubating the linearized fragments with a bacteriophage-specific RNA polymerase capable of initiating transcription from the bacteriophage-specific promoter;

(g) generating first-strand cDNA by transcribing the cRNA using a reverse transcriptase and a 5' RT primer being 15 to 30 nucleotides in length and comprising

a nucleotide sequence that is complementary to the 5' flanking vector sequence;

(h) generating a first set of PCR products by dividing the first-strand cDNA into a first series of subpools and using the first-strand cDNA as templates for a first polymerase chain reaction with a first 3' PCR-primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a first 5' PCR-primer defined as having a 3'-terminus consisting of $-N_1$, wherein "N" is one of the four deoxyribonucleotides A, C, G, or T, the first 5' PCR-primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the first 5' PCR-primer's complementarity extending into one nucleotide of the insert-specific nucleotides of the cRNA, wherein a different one of the first 5' PCR primers is used in each of four different subpools;

(i) generating a second set of PCR products by further dividing the first set of PCR products in each of the first series of subpools into a second series of subpools and using the first set of PCR products as templates for a second polymerase chain reaction with a second 3' PCR primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a second 5' PCR primer defined as having a 3'-terminus consisting of $-N_1-N_x$, wherein N_1 is identical to the N_1 used in the first polymerase chain reaction for that subpool, "N" is as is step h, and "x" is an integer from 1 to 5, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the primer's complementarity extending across into the insert-specific nucleotides of the cRNA in a number of nucleotides equal to "x" + 1, wherein a different one of the second 5' PCR primers is used in different subpools of the second series of subpools and wherein there are 4^x subpools in the second series of subpools for each of the subpools in the first set of subpools;

(j) resolving the second set of PCR products to generate a display of sequence-specific products representing the 3'-ends of mRNAs present in the mRNA population.

(k) eluting at least one cDNA corresponding to a mRNA from an electropherogram in which bands representing the 3'-ends of mRNAs present in the

sample are displayed;

(l) amplifying the eluted cDNA in a polymerase chain reaction;

(m) cloning the amplified cDNA into a plasmid;

(n) producing DNA corresponding to the cloned DNA from the plasmid;

(o) determining the sequence of the cloned cDNA;

(p) determining corresponding nucleotide sequences from a database of nucleotide sequences, said corresponding nucleotide sequences being delimited by the most distal recognition site for the second endonuclease and the beginning of the poly(A) tail; and

(q) comparing the sequence of the cloned cDNA to the corresponding nucleotide sequences, thereby recognizing sequence identities and similarities between the sequence of 3'-ends of mRNA molecules present in a sample and a database of sequences.

Typically, the method further comprises the step of

(r) comparing the length and amount of the PCR products in a two dimensional graphical display.

In general, the method also comprises the steps of

(s) determining the expected length of the corresponding nucleotide sequence, which is equal to the sum of the lengths of the corresponding nucleotide sequence determined from the database, the length of the 5'PCR sequence hybridizable to vector sequence, the length of the remaining anchor primer sequence, an intervening segment of vector sequence and the length of the 3'PCR sequence hybridizable to vector sequence; and

(t) comparing the length of the PCR product to the determined expected length of the corresponding nucleotide sequence, wherein the expected length of corresponding nucleotide sequence is indicated in the two dimensional graphical display by the use of a graphical symbol or text character. Suitable graphical symbols include vertical and horizontal lines, intersecting short line segments such as crosses and "x", and geometric figures, including circles and polygons.

In another embodiment, the invention provides a method for recognizing sequence identities and similarities between the sequence of a cDNA fragment

corresponding to a mRNA molecule present in a sample and a database of sequences, comprising the steps of:

eluting a cDNA fragment corresponding to a mRNA molecule present in a sample;

amplifying the eluted cDNA fragment in a polymerase chain reaction to produce an amplified cDNA fragment;

cloning the amplified cDNA fragment into a plasmid;

producing a DNA molecule corresponding to the cloned cDNA fragment;

sequencing the produced DNA molecule, thereby determining the sequence of the eluted cDNA fragment; and

comparing the sequence of the eluted cDNA fragment to the sequences in a database thereby recognizing sequence identities and similarities.

Typically, the step of comparing the sequence of the eluted cDNA fragment to the sequences in a database is performed using a computer. Typically, the method also comprises the additional step of displaying the results of the comparison graphically.

In general, sequence identities and similarities between the sequence of a cDNA fragment corresponding to a mRNA molecule present in a sample and a database of sequences are recognized by a method comprising the steps of :

eluting a cDNA fragment corresponding to a mRNA molecule present in a sample, where the cDNA fragment has a length determined by the position of a restriction endonuclease recognition site and a poly(A) tail of the mRNA molecule;

determining a partial sequence of the cDNA fragment by performing a polymerase chain reaction with a 5' PCR primer corresponding to the sequence of the restriction endonuclease recognition site and

comparing the determined partial sequence of the eluted cDNA fragment and the length of the cDNA fragment to the sequences in a database thereby recognizing sequence identities and similarities.

In another embodiment, the present invention provides a method of producing a transformed polynucleotide sequence database entry, comprising the steps of:

choosing a source sequence from a polynucleotide sequence database entry;

locating a poly(A) tail sequence within the source sequence;

locating an endonuclease recognition site sequence within the source sequence that is closest to the first recognition site;

determining an index sequence consisting of about two to about six nucleotides adjacent to the endonuclease recognition site;

determining a correlate sequence within the source sequence, said correlate sequence including the sequence bounded by the poly(A) tail and the endonuclease recognition site and including at least part of the endonuclease recognition site;

determining the length of the correlate sequence; and

storing information concerning the location and sequence of the poly(A) tail, the location and sequence of the endonuclease recognition site, and the length of the correlate sequence in relation to the source sequence, thereby producing a transformed database entry. Typically the method includes the step of displaying graphically the length of the correlate sequence in relation to the index sequence. Preferably the restriction endonuclease is chosen from the group consisting of MspI, TaqI and HinP1I.

The invention also provides a method of improving the resolution of the length and amount of PCR products by diminishing background that is due to amplification of untargeted cDNAs comprising the steps of:

selecting a sample of a cRNA population, wherein each cRNA molecule comprises insert sequence and vector-derived sequence;

performing reverse transcription using a reverse transcription primer that hybridizes to the vector-derived sequence and that extends about five nucleotides to about six nucleotides into the insert sequence to produce a cDNA reverse transcription product;

subdividing the cDNA reverse transcription product;

performing at least one polymerase chain reaction using the subdivided cDNA reverse transcription product, a 3'PCR primer and a 5' PCR primer that hybridizes to the vector-derived sequence and extends about seven nucleotides to about nine nucleotides into the insert sequence to produce a PCR product, thereby diminishing background that is due to amplification of untargeted cDNAs.

Typically there are sixteen pools of reverse transcription reactions and there are sixteen different reverse transcription primers. In general, there are 4^x subpools of

polymerase chain reactions, where X is the difference between the number of nucleotides that the 5' PCR primer extends into the insert sequence and the number of nucleotides that the reverse transcription primer extends into the insert sequence.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other features, aspects, and advantages of the present invention will become better understood with reference to the following description, appended claims, and accompanying drawings where:

Figure 1 is a diagrammatic depiction of the improved method of the present invention showing the various stages of priming, cleavage, cloning, antisense RNA transcription and amplification showing the sequences of anchor and other primers schematically – see text for complete sequences;

Figure 2 is a diagrammatic depiction of an embodiment of the improved method using biotinylated anchor primers with streptavidin coated substrate and showing the various stages of priming, cleavage, cloning, antisense RNA transcription and amplification showing the sequences of anchor and other primers schematically – see text for complete sequences;

Figure 3 is a plot of relative abundance of labeled PCR products versus product length in base pairs using a fluorescent detection system, showing analysis of PCR products obtained using a 5' PCR primer C-G-A-C-G-G-T-A-T-C-G-G-G-G-T-G (SEQ ID NO: 42), starting from mRNA samples from serum-starved (A) and serum-added (B) human MG63 cells, data from (A) and (B) were overlaid in the bottom panel (C) using software for comparison of relative expression levels between samples;

Figure 4 is a plot comparing the relative abundance of labeled PCR products versus product length in base pairs using a fluorescent detection system for the method employing two PCR steps versus the method employing only one PCR step, showing the results obtained from analysis of mRNA extracted from serum-starved (A and C) and serum-added (B and D) MG63 osteosarcoma cells using either one PCR step (A-D) or two PCR steps (E-H), presenting data from 5' PCR primers 109T (C-G-A-C-G-G-T-A-T-C-G-G-T-G-C-A, SEQ ID NO: 43) and 45A (C-G-A-C-G-G-T-A-T-C-G-G-A-G-C-A, SEQ ID NO: 44), which differ only at the N1 position (in bold), for serum starved (os-) and serum added (os+) samples, showing that the PCR products generated with 109T and 45A appear to be nearly identical from templates produced by the one PCR step method (A-D), whereas the products detected following PCR from templates produced using the two PCR step method are overall quite distinct (E-H);

Figure 5 is a plot comparing the relative abundance of labeled PCR products versus product length in base pairs using a fluorescent detection system for the comparing results obtained using the standard method depicted in Figure 1 and the magnetic bead embodiment of the method depicted in Figure 2, showing that data from the magnetic bead embodiment display a marked increase in reproducibility across samples (similarity of fragments generated and consistency of intensity values) compared to data derived from the standard embodiment of the method;

Figure 6 is graph showing a linear relationship between cRNA concentration and the peak amplitude of the resulting PCR product for several different tissues;

Figure 7 shows the nucleotide sequences and restriction maps of the multiple cloning sites of plasmids pBC SK$^+$ /DGT1, pBS SK$^+$ /DGT2, pBS SK$^+$ /DGT3, pBC SK$^+$ /DGT4 and pBS SK$^+$ /DGT5; and

Figure 8 is a diagrammatic depiction of an embodiment of the improved method using biotinylated anchor primers with streptavidin coated substrate and showing the various stages of priming, cleavage, cloning, sense RNA transcription and amplification showing the sequences of anchor and other primers schematically –

see text for complete sequences.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

We have developed an improved method for simultaneous sequence-specific identification and display of mRNAs in a mRNA population which has a number of applications in determination of the mechanisms of drug action, drug screening, the study of physiological and pathological conditions, and genomic mapping. The improved method and its applications are discussed below.

I. SIMULTANEOUS SEQUENCE-SPECIFIC IDENTIFICATION OF mRNAs

A method according to the present invention, based on the polymerase chain reaction (PCR) technique, provides means for visualization of nearly every mRNA expressed by normal or neoplastic eukaryotic cells or tissue as a distinct band on a gel whose intensity corresponds roughly to the concentration of the mRNA. The method is based on the observation that virtually all mRNAs conclude with a 3'-poly (A) tail but does not rely on the specificity of primer binding to the tail.

The improved method is schematically illustrated in three embodiments in Figures 1, 2, and 8. In general, the improved method comprises:

(a) preparing a double-stranded cDNA population from an mRNA population using a mixture of anchor primers, each anchor primer having a 5' terminus and a 3' terminus and including: (i) a tract of from 7 to 40 T residues; (ii) a site for cleavage by a first restriction endonuclease that recognizes more than six bases, the site for cleavage being located towards the 5'-terminus relative to the tract of T residues; (iii) a first stuffer segment of from 4 to 40 nucleotides, the first stuffer segment being located towards the 5'-terminus relative to the site for cleavage by the first restriction endonuclease; (iv) a second stuffer segment interposed between the site for cleavage by a first restriction endonuclease that recognizes more than six bases and the tract of T residues, and (v) phasing residues located at the 3' terminus of each of the anchor primers selected from the group consisting of -V, -V-N, and -V-N-N, preferably –V-N-N, wherein V is a deoxyribonucleotide selected from the group consisting of A, C,

and G; and N is a deoxyribonucleotide selected from the group consisting of A, C, G, and T, the mixture including anchor primers containing all possibilities for V and N;

(b) cleaving the double-stranded cDNA population with the first restriction endonuclease and a second restriction endonuclease, the second restriction endonuclease recognizing a four-nucleotide sequence, to form a population of double-stranded cDNA molecules having first and second termini, respectively;

(c) inserting each double-stranded cDNA molecule from step (b) into a vector in an orientation that is antisense with respect to a bacteriophage-specific promoter within the vector to form a population of constructs containing the inserted cDNA molecules, thereby defining 5' and 3' flanking vector sequences adjacent to the 5' terminus of the sense strand of the inserted cDNA and the 3' terminus of the sense strand respectively, and said constructs having a 3' flanking vector sequence at least 15 nucleotides in length between said first restriction endonuclease site and a site defining transcription initiation in said promoter;

(d) Transforming a host cell with the vector into which the cleaved cDNA has been inserted to produce vectors containing cloned inserts;

(e) generating linearized fragments containing the inserted cDNA molecules by digestion of the constructs produced in step (c) with at least one restriction endonuclease that does not recognize sequences in either the inserted cDNA molecules or in the bacteriophage-specific promoter, but does recognize sequences in the vector, such that the resulting linearized fragments have a 5' flanking vector sequence of at least 15 nucleotides into the vector 5' to the double-stranded cDNA molecule's second terminus;

(f) generating a cRNA preparation of antisense cRNA transcripts by incubating the linearized fragments with a bacteriophage-specific RNA polymerase capable of initiating transcription from the bacteriophage-specific promoter;

(g) generating first-strand cDNA by transcribing the cRNA using a reverse transcriptase and a 5' RT primer being 15 to 30 nucleotides in length and comprising a nucleotide sequence that is complementary to the 5' flanking vector sequence;

(h) generating a first set of PCR products by dividing the first-strand cDNA into a first series of subpools and using the first-strand cDNA as templates for a first polymerase chain reaction with a first 3' PCR-primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction

endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a first 5' PCR-primer defined as having a 3'-terminus consisting of $-N_1$, wherein "N" is one of the four deoxyribonucleotides A, C, G, or T, the first 5' PCR-primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the first 5' PCR-primer's complementarity extending into one nucleotide of the insert-specific nucleotides of the cRNA, wherein a different one of the first 5' PCR primers is used in each of four different subpools;

(i) generating a second set of PCR products by further dividing the first set of PCR products in each of the first series of subpools into a second series of subpools and using the first set of PCR products as templates for a second polymerase chain reaction with a second 3' PCR primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a second 5' PCR primer defined as having a 3'-terminus consisting of $-N_1-N_x$, wherein N_1 is identical to the N_1 used in the first polymerase chain reaction for that subpool, "N" is as is step h, and "x" is an integer from 1 to 5, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the primer's complementarity extending across into the insert-specific nucleotides of the cRNA in a number of nucleotides equal to "x" + 1, wherein a different one of the second 5' PCR primers is used in different subpools of the second series of subpools and wherein there are 4^x subpools in the second series of subpools for each of the subpools in the first set of subpools; and

(j) resolving the second set of PCR products to generate a display of sequence-specific products representing the 3'-ends of mRNAs present in the mRNA population.

In an alternative embodiment, step (c) above comprises inserting each double-stranded cDNA molecule from step (b) into a vector in an orientation that is sense with respect to a bacteriophage-specific promoter within the vector to form a population of constructs containing the inserted cDNA molecules (Figure 8).

A. Preparation of Double-Stranded cDNA

The first step in the method requires an mRNA population. Methods of

extraction of RNA are well-known in the art and are described, for example, in J. Sambrook et al., "Molecular Cloning: A Laboratory Manual" (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1989), vol. 1, ch. 7, "Extraction, Purification, and Analysis of Messenger RNA from Eukaryotic Cells," incorporated herein by this reference. Other isolation and extraction methods are also well-known. Typically, isolation is performed in the presence of chaotropic agents such as guanidinium chloride or guanidinium thiocyanate, although other detergents and extraction agents can alternatively be used.

Typically, the mRNA is isolated from the total extracted RNA by chromatography over oligo(dT)-cellulose or other chromatographic media that have the capacity to bind the polyadenylated 3'-portion of mRNA molecules. Alternatively, but less preferably, total RNA can be used. However, it is generally preferred to isolate poly(A)$^+$ RNA.

Double-stranded cDNAs are then prepared from the mRNA population using a mixture of anchor primers to initiate reverse transcription. Each anchor primer has a 5' terminus and a 3' terminus and including: (i) a tract of from 7 to 40 T residues; (ii) a site for cleavage by a first restriction endonuclease that recognizes more than six bases, the site for cleavage being located towards the 5'-terminus relative to the tract of T residues; (iii) a first stuffer segment of from 4 to 40 nucleotides, the first stuffer segment being located towards the 5'-terminus relative to the site for cleavage by the first restriction endonuclease; (iv) a second stuffer segment interposed between the site for cleavage by a first restriction endonuclease that recognizes more than six bases and the tract of T residues, and (v) phasing residues located at the 3' terminus of each of the anchor primers selected from the group consisting of -V, -V-N, and -V-N-N, wherein V is a deoxyribonucleotide selected from the group consisting of A, C, and G; and N is a deoxyribonucleotide selected from the group consisting of A, C, G, and T, the mixture including anchor primers containing all possibilities for V and N where the phasing residues in the mixture are defined by one of -V, -V-N, or -V-N-N. Where the anchor primers have phasing residues of –V, the mixture comprises a mixture of three anchor primers. Where the anchor primers have phasing residues of –V-N, the mixture comprises a mixture of twelve anchor primers. Where the anchor

primers have phasing residues of –V-N-N, the mixture comprises a mixture of 48 anchor primers.

Typically, the anchor primers each have 18 T residues in the tract of T residues, end in -V-N-N, and have a first stuffer segment of 14 residues in length. Preferred sequences of the first stuffer segment are selected from the group consisting of A-A-C-T-G-G-A-A-G-A-A-T-T-C (SEQ ID NO: 1) and G-A-A-T-T-C-A-A-C-T-G-G-A-A (SEQ ID NO: 2). Typically, the site for cleavage by a restriction endonuclease that recognizes more than six bases is the NotI cleavage site.

One preferred set of three anchor primers has the sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 3). Another preferred set of twelve anchor primers has the sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 4). A further preferred set of 48 anchor primers has the sequences A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 5).

In a preferred embodiment, the set of 3 anchor primers has the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 6). In another preferred embodiment, the set of 12 anchor primers has the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 7). In an especially preferred embodiment, the set of 48 anchor primers has the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 8).

One member of this mixture of anchor 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′-end point for each species.

This reaction is carried out under conditions for the preparation of double-

This Sheet Has Been Replaced



stranded cDNA from mRNA that are well-known in the art. Such techniques are described, for example, in Volume 2 of J. Sambrook et al., "Molecular Cloning: A Laboratory Manual", entitled "Construction and Analysis of cDNA Libraries." Suitable reverse transcriptases include those from avian myeloblastosis virus (AMV) and Moloney murine leukemia virus (MMLV). A preferred reverse transcriptase is the MMLV reverse transcriptase.

In preferred embodiments of the invention magnetic beads are used to improve the preparation of the cDNA population (Figures 2 and 8). Typically, the biotin moiety is conjugated to the 5' terminus of the anchor primer and the first restricted cDNA is separated from the remainder of the cDNA by contacting the first restricted cDNA with a streptavidin-coated substrate, such as number of streptavidin coated magnetic beads.

B. Cleavage of the cDNA Sample With Restriction Endonucleases

The cDNA sample is cleaved with two restriction endonucleases. The first restriction endonuclease recognizes a site having more than six bases and cleaves at a single site within each member of the mixture of anchor primers. The second restriction endonuclease is an endonuclease that recognizes a 4-nucleotide sequence. Such endonucleases typically cleave at multiple sites in most cDNAs. Typically, the first restriction endonuclease is NotI and the second restriction endonuclease is MspI. The enzyme NotI does not cleave within most cDNAs. This is desirable to minimize the loss of cloned inserts that would result from cleavage of the cDNAs at locations other than in the anchor site.

Alternatively, the second restriction endonuclease can be TaqI, MaeII or HinP1I. The use of the above three restriction endonucleases can detect rare mRNAs that are not cleaved by MspI. The second restriction endonuclease generates a 5'-overhang compatible for cloning into the desired vector, as discussed below. This cloning, for the vector chosen from the group consisting of pBC SK$^+$, pBS SK$^+$, pBC SK$^+$/DGT1, pBS SK$^+$/DGT2 and pBS SK$^+$/DGT3, is into the ClaI site, as discussed below.

Alternatively, the second restriction endonuclease can be Sau3AI. The use of this restriction endonuclease can also detect rare mRNAs that are not cleaved by MspI. The second restriction endonuclease generates a 5'-overhang compatible for cloning into the desired vector, as discussed below. This cloning for the vector pBC SK$^+$/DGT4 is into the BamHI site, as discussed below.

Alternatively, the second restriction endonuclease can be NlaIII. The use of this restriction endonuclease can also detect rare mRNAs that are not cleaved by MspI. The second restriction endonuclease generates a 5'-overhang compatible for cloning into the desired vector, as discussed below. This cloning for the vector pBS SK$^+$/DGT5, is into the SphI site, as discussed below.

Alternatively, other suitable restriction endonucleases can be used to detect cDNAs not cleaved by the above restriction endonucleases. Suitable second restriction endonucleases recognizing a four-nucleotide sequence are MboI, DpnII, Sau3AI, Tsp509I, HpaII, BfaI, Csp6I, MseI, HhaI, NlaIII, TaqI, MspI, MaeII and HinPII.

Suitable first restriction endonucleases that recognize more than six bases are AscI, BaeI, FseI, NotI, PacI, PmeI PpuMI, RsrII, SapI, SexAI, SfiI, SgfI, SgrAI, SrfI, Sse8387I and SwaI.

Conditions for digestion of the cDNA are well-known in the art and are described, for example, in J. Sambrook et al., "Molecular Cloning: A Laboratory Manual," Vol. 1, Ch. 5, "Enzymes Used in Molecular Cloning."

C. Insertion of Cleaved cDNA into a Vector

The cDNA sample cleaved with the first and second restriction endonucleases is then inserted into a vector. In general, a suitable vector includes a multiple cloning site having a NotI restriction endonuclease site. A suitable vector is the plasmid pBC SK$^+$ that has been cleaved with the restriction endonucleases ClaI and NotI. The

vector contains a bacteriophage-specific promoter. Typically, the promoter is a T3 promoter, a SP6 promoter, or a T7 promoter. A preferred promoter is a bacteriophage T3 promoter. The cleaved cDNA is inserted into the promoter in an orientation that is antisense with respect to the bacteriophage-specific promoter (Figures 1 and 2). In another preferred embodiment, the cleaved cDNA is inserted into the promoter in an orientation that is sense with respect to the bacteriophase-specific promoter (Figure 8). In a preferred embodiment, the vector includes a multiple cloning site having a nucleotide sequence chosen from the group consisting of SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 11, SEQ ID NO: 12 and SEQ ID NO: 13.

Preferred vectors are based on the plasmid vector pBluescript (pBS or pBC) SK+ (Stratagene) in which a portion of the nucleotide sequence from positions 656 to 764 was removed and replaced with a sequence of at least 110 nucleotides including a NotI restriction endonuclease site. This region, designated the multiple cloning site (MCS), spans the portion of the nucleotide sequence from the SacI site to the KpnI site.

A suitable plasmid vector, such as pBC SK^+ or pBS SK^+ (Stratagene), was digested with suitable restriction endonuclease to remove at least 100 nucleotides of the multiple cloning site. In the case of pBS SK^+, suitable restriction endonucleases for removing the multiple cloning site are SacI and KpnI. A cDNA portion comprising a new multiple cloning site, having ends that are compatible with NotI and ClaI after digestion with first and second restriction endonucleases was cloned into the vector to form a suitable plasmid vector. Preferred cDNA portions comprising new multiple cloning sites include those having the nucleotide sequences described in SEQ ID NO: 9, SEQ ID NO: 10 and SEQ ID NO: 11. cDNA clones are linearized by digestion with a single restriction endonuclease that recognizes a sequence having more than six bases that includes the four nucleotide sequence of the second restriction endonuclease site.

A preferred plasmid vector, referred to herein as pBC SK^+/DGT1, comprises the MCS of SEQ ID NO:9. The pairs for second restriction endonuclease and

linearization restriction endonuclease (of step E, below) are, respectively: MspI and SmaI; HinP1I and NarI; TaqI and XhoI; MaeII and AatII.

Another preferred plasmid vector, referred to herein as pBS SK+/DGT2, comprises the MCS of SEQ ID NO:10, and was prepared as described above for pBC SK+/DGT1. The multiple cloning site does not accept cDNA inserts produced using MaeII. Thus, for pBS SK+/DGT2, the pairs for second restriction endonuclease and linearization restriction endonuclease (of step E, below) are, respectively: MspI and SmaI; HinP1I and NarI; and TaqI and XhoI.

Another preferred plasmid vector, referred to herein as pBS SK+/DGT3, comprises the MCS of SEQ ID NO:11. The pairs for second restriction endonuclease and linearization restriction endonuclease (of step E, below) are, respectively: MspI and SmaI; HinP1I and NarI; TaqI and XhoI; MaeII and AatII.

Another preferred plasmid vector, referred to herein as pBC SK+/DGT4, comprises the MCS of SEQ ID NO:12. The pair of second restriction endonuclease and linearization restriction endonuclease (of step E, below) enzymes suitable for use with this vector are, respectively, Sau3AI and BglII.

Another preferred plasmid vector, referred to herein as pBS SK+/DGT5, comprises the MCS of SEQ ID NO:13. The pair of second restriction endonuclease and linearization restriction endonuclease (of step E, below) enzymes suitable for use with this vector are, respectively, NlaIII and NcoI.

In a preferred embodiment, the vector includes a vector stuffer sequence that comprises an internal vector stuffer restriction endonuclease site between the first and second vector restriction endonuclease sites. In one such an embodiment, the linearization step includes digestion of the vector with a restriction endonuclease which cleaves the vector at the internal vector stuffer restriction endonuclease site. In another such embodiment, the restriction endonuclease used in the linearization step also cleaves the vector at the internal vector stuffer restriction endonuclease site.

D. Transformation of a Suitable Host Cell

The vector into which the cleaved DNA has been inserted is then used to transform a suitable host cell that can be efficiently transformed or transfected by the vector containing the insert. Suitable host cells for cloning are described, for example, in Sambrook et al., "Molecular Cloning: A Laboratory Manual," supra. Typically, the host cell is prokaryotic. A particularly suitable host cell is a strain of E. coli. A suitable E. coli strain is MC1061. Preferably, a small aliquot is also used to transform E. coli strain XL1-Blue so that the percentage of clones with inserts is determined from the relative percentages of blue and white colonies on X-gal plates. Only libraries with in excess of $5x10^5$ recombinants are typically acceptable.

E. Generation of Linearized Fragments

Plasmid preparations are then made from each of the cDNA libraries. Linearized fragments are then generated by digestion with at least one restriction endonuclease.

In one embodiment, vector is the plasmid pBC SK^+ and MspI is used both as the second restriction endonuclease and as the linearization restriction endonuclease.

In another embodiment, vector is the plasmid pBC SK^+, the second restriction endonuclease is chosen from the group consisting of MspI, MaeII, TaqI and HinP1I and the linearization is accomplished by a first digestion with SmaI followed by a second digestion with a mixture of KpnI and ApaI.

In other embodiments the vector is chosen from the group consisting of pBC

SK+ /DGT1, pBS SK+ /DGT2, pBS SK+ /DGT3, pBC SK+ /DGT4 and pBS SK+ /DGT5. In such embodiments, one suitable enzyme combination is provided where the second restriction endonuclease is MspI and the restriction endonuclease used in the linearization step is Sma I. Another suitable combination is provided where the second restriction endonuclease is TaqI and the restriction endonuclease used in the linearization step is XhoI. A further suitable combination is provided where the second restriction endonuclease is HinP1I and the restriction endonuclease used in the linearization step is NarI. Yet another suitable combination is provided where the second restriction endonuclease is MaeII and the restriction endonuclease used in the linearization step is AatII. If the vector is pBC SK+ /DGT4, another suitable combination is provided by Sau3AI as the second restriction endonuclease and BglII as the restriction endonuclease used in the linearization step. If the vector is pBS SK+ /DGT5, another suitable combination is provided by NlaIII as the second restriction endonuclease and NcoI as the restriction endonuclease used in the linearization step.

In general, in the linearization step, described in detail in Section F, below, any plasmid vector lacking a cDNA insert was cleaved at the 6-nucleotide recognition site (underlined in Figure 7A) for SmaI, NarI, XhoI, or AatII found between the NotI site and the ClaI site and the recognition site having more than six bases for SmaI, NarI, XhoI or AatII sites found 3' to the ClaI site. In contrast, plasmid vectors containing inserts would be cleaved at the 6-nucleotide recognition site for SmaI, NarI, XhoI or AatII sites found 3' to the ClaI site.

F. Generation of cRNA

The next step is a generation of a cRNA preparation of antisense cRNA transcripts. This is performed by incubation of the linearized fragments with an RNA polymerase capable of initiating transcription from the bacteriophage-specific promoter. Typically, as discussed above, the promoter is a T3 promoter, and the polymerase is therefore T3 RNA polymerase. The polymerase is incubated with the linearized fragments and the four ribonucleoside triphosphates under conditions suitable for synthesis (Ambion, Austin, TX).

G. Transcription of First-Strand cDNA

First-strand cDNA is transcribed using Moloney murine leukemia virus

(MMLV) reverse transcriptase (Life Technologies, Gaithersburg, MD). With this reverse transcriptase annealing is performed at 42°C, and the transcription reaction at 42°C. The reaction uses a primer which is 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence.

In another embodiment, the cRNA is transcribed using a thermostable reverse transcriptase and a primer as described below. A preferred transcriptase is the avian recombinant reverse transcriptase, known as ThermoScript RT, available from Life Technologies (Gaithersburg, MD).

This promotes high fidelity complementarity between the primer and the cRNA. The primer used is at least 15 nucleotides in length, corresponding in sequence to the 3'-end of the bacteriophage-specific promoter.

Another suitable transcriptase is the recombinant reverse transcriptase from Thermus thermophilus, known as rTth, available from Perkin-Elmer (Norwalk, CT).

Where the bacteriophage-specific promoter is the T3 promoter, the primers typically have the sequence A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G (SEQ ID NO: 14) or G-A-G-C-T-C-C-A-C-C-G-C-G-G-T (SEQ ID NO: 47).

H. Generating First PCR Product

The next step is the use of the product of transcription as a template for a polymerase chain reaction with a first set of primers as described below to produce polymerase chain reaction amplified fragments.

In general, the product of first-strand cDNA transcription is used as a template for a polymerase chain reaction with a first 3' PCR primer and a first 5' PCR primer to produce polymerase chain reaction amplified fragments. The first 3' PCR primer typically is 15 to 30 nucleotides in length, and is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter. The first 5'-PCR primers have a 3' terminus consisting of $-N_1$ where "N_1" is one of the four

deoxyribonucleotides A, C, G, or T, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the primer's complementarity extending into one nucleotide of the insert-specific nucleotides of the cRNA, wherein a different one of the first 5' PCR primers is used in each of four different subpools.

When the vector is the plasmid pBC SK$^+$ cleaved with ClaI and NotI, a suitable 3'-PCR primer is selected from the group consisting of G-A-G-C-T-C-C-A-C-C-G-C-G-G-T (SEQ ID NO: 47) and G-A-G-C-T-C-G-T-T-T-T-C-C-C-A-G (SEQ ID NO: 48). Where the bacteriophage-specific promoter is the T3 promoter, a suitable 5'-PCR primer can have the sequence G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N (SEQ ID NO: 22) where in a given reaction N is either A, G, C, or T.

Typically, PCR is performed using a PCR program of 15 seconds at 94°C for denaturation, 15 seconds at 50°C - 65°C for annealing, and 30 seconds at 72°C for synthesis on a suitable thermocycler such as the PTC-200 (MJ Research) or the Perkin-Elmer 9600 (Perkin-Elmer Cetus, Norwalk, CT). The annealing temperature is optimized for the specific nucleotide sequence of the primer, using principles well known in the art. The high temperature annealing step minimizes artifactual mispriming by the first 5'-PCR primer at its 3'-end and promotes high fidelity copying.

I. Generating Second PCR Product

The next step is the use of the products of the first PCR reaction as templates for a second polymerase chain reaction with a second set of primers as described below to produce a second set of polymerase chain reaction amplified fragments.

In general, the product of first PCR reaction is used as a template for a polymerase chain reaction with a second 3' PCR primer and a second 5'-PCR primer to produce polymerase chain reaction amplified fragments. The second 3' PCR primer typically is 15 to 30 nucleotides in length, and is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining

transcription initiation by the bacteriophage-specific promoter. The second 5' PCR primer is defined as having a 3'-terminus consisting of $-N_1-N_x$, wherein N_1 is identical to the N_1 used in the first polymerase chain reaction for that subpool, "N" is as is step (H), and "x" is an integer from 1 to 5, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the primer's complementarity extending across into the insert-specific nucleotides of the cRNA in a number of nucleotides equal to "x" + 1, wherein a different one of the second 5' PCR primers is used in different subpools of the second series of subpools and wherein there are 4^x subpools in the second series of subpools for each of the subpools in the first set of subpools.

In another embodiment, the primers used are: (a) a second 3' PCR primer that corresponds in sequence to a sequence in the vector adjoining the site of insertion of the cDNA sample in the vector; and (b) a 5'-PCR primer selected from the group consisting of: (i) the first 5' PCR primer which was used in the first PCR reaction for that subpool; (ii) the first 5' PCR primer from which the first-strand cDNA was made for that subpool extended at its3'-terminus by an additional residue –N; (iii) the first 5' PCR primer used for that subpool extended at its 3' terminus by two additional residues -N-N, (iv) the first 5' PCR primer used for that subpool extended at its 3' terminus by three additional residues –N-N-N; and (v) the first 5' PCR primer used for that subpool extended at its 3' terminus by four additional residues -N-N-N-N, wherein N can be any of A, C, G, or T.

Suitable 3' PCR primers are selected from the group consisting of G-A-G-C-T-C-C-A-C-C-G-C-G-G-T (SEQ ID NO: 47) and G-A-G-C-T-C-G-T-T-T-T-C-C-C-A-G (SEQ ID NO: 48).

Where the bacteriophage-specific promoter is the T3 promoter, a suitable 5'-PCR primer is chosen from the group consisting of the sequences:

A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N (SEQ ID NO: 16);

A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N (SEQ ID NO: 17);

A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N (SEQ ID NO: 18);

G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N (SEQ ID NO: 22);

G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N (SEQ ID NO: 23);
T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N (SEQ ID NO: 24);
C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N (SEQ ID NO: 25);
G-A-C-G-G-T-A-T-C-G-G-N-N-N-N-N (SEQ ID NO: 26);
A-C-G-G-T-A-T-C-G-G-N-N-N-N-N-N (SEQ ID NO: 16);
A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N-N (SEQ ID NO: 19); and
A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N-N-N (SEQ ID NO: 20).

Typically, PCR is performed using a PCR program of 15 seconds at 94°C for denaturation, 15 seconds at 50°C - 65°C for annealing, and 30 seconds at 72°C for synthesis on a suitable thermocycler such as the PTC-200 (MJ Research) or the Perkin-Elmer 9600 (Perkin-Elmer Cetus, Norwalk, CT). The annealing temperature is optimized for the specific nucleotide sequence of the primer, using principles well known in the art. The high temperature annealing step minimizes artifactual mispriming by the 5′-primer at its 3′-end and promotes high fidelity copying.

In preferred embodiments detection methods utilizing non-radioactive labels can also be used. For non-radioactive detection methods, one of the primers for the second PCR reaction is preferably conjugated to a fluorescent label. A suitable fluorescent label is selected from the group consisting of
spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 6-carboxylic acid, 3',6'-dihydroxy-6-carboxyfluorescein (6-FAM, ABI);
spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 5-carboxylic acid, 3',6'-dihydroxy-5-carboxyfluorescein (5-FAM, Molecular Probes);
spiro(isobenzofuran-1(3H), 9'-(9H)-xanthen)-3-one, 3',6'-dihydroxy-fluorescein (FAM, Molecular Probes);
9-(2,5-dicarboxyphenyl)-3,6- bis(dimethylamino)-xanthylium (6-carboxytetramethylrhodamine (6-TAMRA), Molecular Probes);
3,6-diamino-9-(2-carboxyphenyl)-xanthylium (Rhodamine Green™, Molecular Probes);
spiro[isobenzofuran-1(3H), 9'-xanthene]-6-carboxylic acid,5'-dichloro-3',6'-dihydroxy-2',7'-dimethoxy-3-oxo-(JOE, Molecular Probes);
1H,5H,11H,15H-xantheno[2,3,4-ij:5,6,7-i'j']diquinolizin- 8-ium, -(2,4-

disulfophenyl)-2,3,6,7,12,13,16,17-octahydro-, inner salt (Texas Red, Molecular Probes);
6-((4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-s-indacene-3-propionyl) amino) hexanoic acid (BODIPY FL-X, Molecular Probes);
6-((4,4-difluoro-1,3-dimethyl-5-(4-methoxyphenyl)-4-bora-3a,4a-diaza-s-indacene-3-propionyl)amino)hexanoic acid (BODIPY TMR-X, Molecular Probes); 6-(((4-(4,4-difluoro-5-(2-thienyl)-4-bora-3a,4a-diaza-s-indacene-3-yl) phenoxy)acetyl) amino)-hexanoic acid (BODIPY TR-X, Molecular Probes);
4,4-difluoro-4-bora-3a,4a-diaza-s-indacene-3-pentanoic acid (BODIPY FL-C_5, Molecular Probes);
4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-s-indacene-3-propanoic acid (BODIPY FL, Molecular Probes);
4,4-difluoro-5-phenyl-4-bora-3a,4a-diaza-s-indacene-3-propionic acid (BODIPY 581/591, Molecular Probes);
4,4-difluoro-5-(4-phenyl-1,3-butadienyl)-4-bora-3a,4a-diaza-s-indacene-3-propionic acid (BODIPY 564/570, Molecular Probes);
4,4-difluoro-5-styryl-4-bora-3a,4a-diaza-s-indacene-3-propionic acid;
6-(((4,4-difluoro-5-(2-thienyl)-4-bora-3a,4a-diaza-s-indacene-3-yl)styryloxy)acetyl) aminohexanoic acid (BODIPY 630/650, Molecular Probes);
6-(((4,4-difluoro-5-(2-pyrrolyl)-4-bora-3a,4a-diaza-s-indacene-3-yl) styryloxy)acetyl) aminohexanoic acid (BODIPY 650/665, Molecular Probes); and
9-(2,4(or 2,5)-dicarboxyphenyl)-3,6- bis(dimethylamino)- xanthylium, inner salt (TAMRA, Molecular Probes). Other suitable fluorescent labels, including 4, 7, 2', 4', 5', 7' hexachloro 6-carboxyfluorescein ("HEX," ABI), 4, 7, 2', 7' tetrachloro 6-carboxyfluorescein ("TET," ABI) and "NED" (ABI) are known in the art.

A preferred fluorescent label is spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 6-carboxylic acid, 3',6'-dihydroxy-6-carboxyfluorescein (6-FAM).

In alternative embodiments, autoradiographic detection methods can be used. In one embodiment, the PCR is performed in the presence of ^{35}S-dATP Alternatively,

the PCR amplification can be carried out in the presence of a radionuclide labeled deoxyribonucleoside triphosphate, such as [^{32}P]dCTP or [^{33}P]dCTP. However, for autoradiographic detection it is generally preferred to use a ^{35}S-labeled deoxyribonucleoside triphosphate for maximum resolution.

In an alternative embodiment, the detection method employs oligonucleotides that are labeled with magnetic particles that are used and detected as described in U.S. Patent No. 5,656,429, the teachings of which are incorporated by reference.

In one preferred embodiment, the 3 nucleotides at the 3' end of the first or second 5' PCR primer are joined by phosphorothioate linkages. See, Mullins, J. I., de Noronha, C. M. Amplimers with 3'-terminal phosphorothioate linkages resist degradation by vent polymerase and reduce Taq polymerase mispriming. PCR Methods Appl 1992 2(2):131-136; Ott, J. and Eckstein, F. Protection of oligonucleotide primers against degradation by DNA polymerase I. Biochemistry 1987 26(25):8237-8241; Uhlmann, E., Ryte, A., and Peyman, A. Studies on the mechanism of stabilization of partially phosphorothioated oligonucleotides against nucleolytic degradation. Antisense Nucleic Acid Drug Dev. 1997 7(4):345-350; Schreiber, G., Koch, E. M., and Neubert, W. J. Selective protection of in vitro synthesized cDNA against nucleases by incorporation of phosphorothioate-analogues. Nucleic Acids Res. 1985 13(21):7663-7672.

J. Electrophoresis

The polymerase chain reaction amplified fragments are then resolved by a separation method such as electrophoresis to display bands representing the 3'-ends of mRNAs present in the sample.

Electrophoretic techniques for resolving PCR amplified fragments are well-understood in the art and need not be further recited here in detail. The corresponding PCR products are resolved in denaturing DNA sequencing gels and visualized by laser induced fluorescence. Alternatively, the corresponding PCR products are resolved using capillary electrophoresis and visualized by laser induced fluorescence.

In one preferred embodiment, one of the primers for the second PCR reaction is conjugated to a fluorescent label. A suitable fluorescent label is selected from the group consisting of

spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 6-carboxylic acid, 3',6'-dihydroxy-6-carboxyfluorescein (6-FAM, ABI);

spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 5-carboxylic acid, 3',6'-dihydroxy-5-carboxyfluorescein (5-FAM, Molecular Probes);

spiro(isobenzofuran-1(3H), 9'-(9H)-xanthen)-3-one, 3',6'-dihydroxy-fluorescein (FAM, Molecular Probes);

9-(2,5-dicarboxyphenyl)-3,6- bis(dimethylamino)-xanthylium (6-carboxytetramethylrhodamine (6-TAMRA), Molecular Probes);

3,6-diamino-9-(2-carboxyphenyl)-xanthylium (Rhodamine Green™, Molecular Probes);

spiro[isobenzofuran-1(3H), 9'-xanthene]-6-carboxylic acid,5'-dichloro-3',6'-dihydroxy-2',7'-dimethoxy-3-oxo-(JOE, Molecular Probes);

1H,5H,11H,15H-xantheno[2,3,4-ij:5,6,7-i'j']diquinolizin- 8-ium, -(2,4-disulfophenyl)-2,3,6,7,12,13,16,17-octahydro-, inner salt (Texas Red, Molecular Probes);

6-((4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-s-indacene-3-propionyl) amino) hexanoic acid (BODIPY FL-X, Molecular Probes);

6-((4,4-difluoro-1,3-dimethyl-5-(4-methoxyphenyl)-4-bora-3a,4a-diaza-s-indacene-3-propionyl)amino)hexanoic acid (BODIPY TMR-X, Molecular Probes);

6-(((4-(4,4-difluoro-5-(2-thienyl)-4-bora-3a,4a-diaza-s-indacene-3-yl) phenoxy)acetyl) amino)-hexanoic acid (BODIPY TR-X, Molecular Probes);

4,4-difluoro-4-bora-3a,4a-diaza-s-indacene-3-pentanoic acid (BODIPY FL-C_5, Molecular Probes);

4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-s-indacene-3-propanoic acid (BODIPY FL, Molecular Probes);

4,4-difluoro-5-phenyl-4-bora-3a,4a-diaza-s-indacene-3-propionic acid (BODIPY 581/591, Molecular Probes);

4,4-difluoro-5-(4-phenyl-1,3-butadienyl)-4-bora-3a,4a-diaza-s-indacene-3-propionic acid (BODIPY 564/570, Molecular Probes);

4,4-difluoro-5-styryl-4-bora-3a,4a-diaza-s-indacene-3-propionic acid;
6-(((4,4-difluoro-5-(2-thienyl)-4-bora-3a,4a-diaza-s-indacene-3-yl)styryloxy)acetyl) aminohexanoic acid (BODIPY 630/650, Molecular Probes);
6-(((4,4-difluoro-5-(2-pyrrolyl)-4-bora-3a,4a-diaza-s-indacene-3-yl) styryloxy)acetyl) aminohexanoic acid (BODIPY 650/665, Molecular Probes); and
9-(2,4(or 2,5)-dicarboxyphenyl)-3,6- bis(dimethylamino)- xanthylium, inner salt (TAMRA, Molecular Probes). Other suitable fluorescent labels, including 4, 7, 2', 4', 5', 7' hexachloro 6-carboxyfluorescein ("HEX," ABI), NED (ABI) and 4, 7, 2', 7' tetrachloro 6-carboxyfluorescein ("TET," ABI) are known in the art.

Typically, fluorescence is used to detect the resolved cDNA species. However, other detection methods, such as phosphorimaging or autoradiography, or magnetic detection, can also be used.

According to the scheme, the cDNA libraries produced from each of the mRNA samples contain copies of the extreme 3'-ends from the most distal site for MspI to the beginning of the poly(A) tail of 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 exactly defined by sequence, their lengths are uniform for each species allowing their later visualization as discrete bands on a gel, regardless of the tissue source of the mRNA.

Typically, the intensity of products displayed after electrophoresis is about proportional to the abundances of the mRNAs corresponding to the products in the original mixture.

Typically, the method further comprises a step of determining the relative abundance of each mRNA in the original mixture from the intensity of the product corresponding to that mRNA after electrophoresis.

II. APPLICATIONS OF THE METHOD FOR DISPLAY OF mRNA

PATTERNS

The method described above for the detection of patterns of mRNA expression in a tissue and the resolving of these patterns by gel electrophoresis has a number of applications. One of these applications is its use for the detection of a change in the pattern of mRNA expression in a tissue associated with a physiological or pathological change. In general, this method comprises:

(1) obtaining a first sample of a tissue that is not subject to the physiological or pathological change;

(2) determining the pattern of mRNA expression in the first sample of the tissue by performing the method of simultaneous sequence-specific identification of mRNAs corresponding to members of an antisense cRNA pool representing the 3′-ends of a population of mRNAs as described above to generate a first display of bands representing the 3′-ends of mRNAs present in the first sample;

(3) obtaining a second sample of the tissue that has been subject to the physiological or pathological change;

(4) determining the pattern of mRNA expression in the second sample of the tissue by performing the method of simultaneous sequence-specific identification of mRNAs corresponding to members of an antisense cRNA pool representing the 3′-ends of a population of mRNAs as described above to generate a second display of bands representing the 3′-ends of mRNAs present in the second sample; and

(5) comparing the first and second displays to determine the effect of the physiological or pathological change on the pattern of mRNA expression in the tissue.

Typically, the comparison is made in adjacent lanes of a single gel.

Typically, a database comprising the data produced by the quantitation of the display of sequence-specific products is constructed and maintained using suitable computer hardware and computer software. Preferably, such a database further comprises data concerning sequence relationships, gene mapping and cellular distributions. In preferred embodiments the length and at least part of the nucleotide sequence of the PCR products are compared to expected values determined from a database of nucleotide sequences.

The tissue can be derived from the central nervous system. In particular, it can be derived from a structure within the central nervous system that is the retina, cerebral cortex, olfactory bulb, thalamus, hypothalamus, anterior pituitary, posterior pituitary, hippocampus, nucleus accumbens, amygdala, striatum, cerebellum, brain stem, suprachiasmatic nucleus, or spinal cord. When the tissue is derived from the central nervous system, the physiological or pathological change can be any of Alzheimer's disease, parkinsonism, ischemia, alcohol addiction, drug addiction, schizophrenia, amyotrophic lateral sclerosis, multiple sclerosis, depression, and bipolar manic-depressive disorder. Alternatively, the method of the present invention can be used to study circadian variation, aging, or long-term potentiation, the latter affecting the hippocampus. Additionally, particularly with reference to mRNA species occurring in particular structures within the central nervous system, the method can be used to study brain regions that are known to be involved in complex behaviors, such as learning and memory, emotion, drug addiction, glutamate neurotoxicity, feeding behavior, olfaction, viral infection, vision, and movement disorders.

This method can also be used to study the results of the administration of drugs and/or toxins to an individual by comparing the mRNA pattern of a tissue before and after the administration of the drug or toxin. Results of electroshock therapy can also be studied.

Alternatively, the tissue can be from an organ or organ system that includes the cardiovascular system, the pulmonary system, the digestive system, the peripheral nervous system, the liver, the kidney, skeletal muscle, and the reproductive system, or from any other organ or organ system of the body. For example, mRNA patterns can be studied from liver, heart, kidney, or skeletal muscle. Additionally, for any tissue, samples can be taken at various times so as to discover a circadian effect of mRNA expression. Thus, this method can ascribe particular mRNA species to involvement in particular patterns of function or malfunction.

Preferably, the normal or neoplastic tissue comprises cells taken or derived

from an organ or organ system selected from the group consisting of the cardiovascular system, the lymphatic system, the respiratory system, the digestive system, the peripheral nervous system, the central nervous system, the enteric nervous system, the endocrine system, the integument (including skin, hair and nails), the skeletal system (including bone and muscle), the urinary system and the reproductive system.

In preferred embodiments, the normal or neoplastic tissue comprises cells taken or derived from the group consisting of epithelia, endothelia, mucosa, glands, blood, lymph, connective tissue, cartilage, bone, smooth muscle, skeletal muscle, cardiac muscle, neurons, glial cells, spleen, thymus, pituitary, thyroid, parathyroid, adrenal cortex, adrenal medulla, adrenal cortex, pineal, skin, hair, nails, teeth, liver, pancreas, lung, kidney, bladder, ureter, breast, ovary, uterus, vagina, testes, prostate, penis, eye and ear.

Similarly, the mRNA resolution method of the present invention can be used as part of a method of screening for a side effect of a drug. In general, such a method comprises:

(1) obtaining a first sample of tissue from an organism treated with a compound of known physiological function;

(2) determining the pattern of mRNA expression in the first sample of the tissue by performing the method of simultaneous sequence-specific identification of mRNAs corresponding to members of an antisense cRNA pool representing the 3'-ends of a population of mRNAs, as described above, to generate a first display of bands representing the 3'-ends of mRNAs present in the first sample;

(3) obtaining a second sample of tissue from an organism treated with a drug to be screened for a side effect;

(4) determining the pattern of mRNA expression in the second sample of the tissue by performing the method of simultaneous sequence-specific identification of mRNAs corresponding to members of an antisense cRNA pool representing the 3'-ends of a population of mRNAs, as described above, to generate a second display of bands representing the 3'-ends of mRNAs present in the second sample; and

(5) comparing the first and second displays in order to detect the presence of

mRNA species whose expression is not affected by the known compound but is affected by the drug to be screened, thereby indicating a difference in action of the drug to be screened and the known compound and thus a side effect.

In particular, this method can be used for drugs affecting the central nervous system, such as antidepressants, neuroleptics, tranquilizers, anticonvulsants, monoamine oxidase inhibitors, and stimulants. However, this method can in fact be used for any drug that may affect mRNA expression in a particular tissue. For example, the effect on mRNA expression of anti-parkinsonism agents, skeletal muscle relaxants, analgesics, local anesthetics, cholinergics, antispasmodics, steroids, non-steroidal anti-inflammatory drugs, antiviral agents, or any other drug capable of affecting mRNA expression can be studied, and the effect determined in a particular tissue or structure.

A further application of the method of the present invention is in obtaining the sequence of the 3'-ends of mRNA species that are displayed. In general, a method of obtaining the sequence comprises:

(1) eluting at least one cDNA corresponding to a mRNA from an electropherogram in which bands representing the 3'-ends of mRNAs present in the sample are displayed;

(2) amplifying the eluted cDNA in a polymerase chain reaction;

(3) cloning the amplified cDNA into a plasmid;

(4) producing DNA corresponding to the cloned DNA from the plasmid; and

(5) sequencing the cloned cDNA.

The cDNA that has been excised can be amplified with the primers previously used in the second PCR step. The cDNA can then be cloned into pCR II (Invitrogen, San Diego, CA) by TA cloning and ligation into the vector. Minipreps of the DNA can then be produced by standard techniques from subclones and a portion denatured and split into two aliquots for automated sequencing by the dideoxy chain termination method of Sanger. A commercially available sequencer can be used, such as a ABI sequencer, for automated sequencing. This will allow the determination of complementary sequences for most cDNAs studied, in the length range of 50-500 bp,

These partial sequences can then be used to scan nucleotide data bases such as GenBank using suitable computer equipment to recognize sequence identities and similarities using comparison and analysis programs such as BLASTN and BLASTX. Because this method generates sequences from only the 3'-ends of mRNAs it is expected that open reading frames (ORFs) would be encountered only occasionally. For example, the 3'-untranslated regions of brain mRNAs are on average longer than 1300 nucleotides (J.G. Sutcliffe, 1988, supra). Potential ORFs can be examined for signature protein motifs.

The cDNA sequences obtained can then be used to design primer pairs for semiquantitative PCR to confirm tissue expression patterns. Selected products can also be used to isolate full-length cDNA clones for further analysis. Primer pairs can be used for SSCP-PCR (single strand conformation polymorphism-PCR) amplification of genomic DNA. For example, such amplification can be carried out from a panel of interspecific backcross mice to determine linkage of each PCR product to markers already linked. This can result in the mapping of new genes and can serve as a resource for identifying candidates for mapped mouse mutant loci and homologous human disease genes. SSCP-PCR uses synthetic oligonucleotide primers that amplify, via PCR, a small (100-200 bp) segment. (M. Orita et al., "Detection of Polymorphisms of Human DNA by Gel Electrophoresis as Single-Strand Conformation Polymorphisms," Proc. Natl. Acad. Sci. USA 86: 2766-2770 (1989); M. Orita et al., "Rapid and Sensitive Detection of Point Mutations in DNA Polymorphisms Using the Polymerase Chain Reaction," Genomics 5: 874-879 (1989)).

The excised fragments of cDNA can be radiolabeled by techniques well-known in the art for use in probing a northern blot or for in situ hybridization to verify mRNA distribution and to learn the size and prevalence of the corresponding full-length mRNA. The probe can also be used to screen a cDNA library to isolate clones for more reliable and complete sequence determination. The labeled probes can also be used for any other purpose, such as studying in vitro expression.

III. PANELS AND DEGENERATE MIXTURES OF PRIMERS

Another aspect of the present invention is panels of primers and degenerate mixtures of primers suitable for the practice of the present invention. These include:

(1) a panel of primers comprising 16 primers of the sequence A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N (SEQ ID NO: 16), wherein N is one of the four deoxyribonucleotides A, C, G, or T;

(2) a panel of primers comprising 64 primers of the sequences A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N (SEQ ID NO: 17),

(3) a panel of primers comprising 256 primers of the sequences A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N (SEQ ID NO: 18);

(4) a panel of primers comprising 1024 primers of the sequences A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N-N (SEQ ID NO: 19);

(5) a panel of primers comprising 4096 primers of the sequences A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N-N-N (SEQ ID NO: 20);

(6) a panel of primers comprising 3 primers of the sequences A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 3);

(7) a panel of primers comprising 12 primers of the sequences A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 4), wherein V is a deoxyribonucleotide selected from the group consisting of A, C, and G;

(8) a panel of primers comprising 48 primers of the sequences A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 5);

(9) a panel of primers comprising 3 primers of the sequences G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 6);

(10) a panel of primers comprising 12 primers of the sequences G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 7);

This Sheet Has Been Replaced



(11) a panel of primers comprising 48 primers of the sequences G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 8);

(12) a panel of primers comprising 4 different oligonucleotides each having the sequence G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N (SEQ ID NO: 22);

(13) a panel of primers comprising 16 different oligonucleotides each having the sequence G-T-C-G-A-C-G-G-T-A-T-C-G-G-N-N (SEQ ID NO: 23);

(14) a panel of primers comprising 64 different oligonucleotides each having the sequence T-C-G-A-C-G-G-T-A-T-C-G-G-N-N-N (SEQ ID NO: 24);

(15) a panel of primers comprising 256 different oligonucleotides each having the sequence C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N (SEQ ID NO: 25);

(16) a panel of primers comprising 1024 different oligonucleotides each having the sequence G-A-C-G-G-T-A-T-C-G-G-N-N-N-N-N (SEQ ID NO: 26);

(17) a panel of primers comprising 4096 different oligonucleotides each having the sequence A-C-G-G-T-A-T-C-G-G-N-N-N-N-N-N (SEQ ID NO: 27);

(18) a degenerate mixture of primers comprising a mixture of 3 primers of the sequences A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 2), each of the 3 primers being present in about an equimolar quantity;

(19) a degenerate mixture of primers comprising a mixture of 12 primers of the sequences A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 4), each of the 12 primers being present in about an equimolar quantity;

(20) a degenerate mixture of primers comprising a mixture of 48 primers of the sequences A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 5), each of the 48 primers being present in about an equimolar quantity;

(21) a degenerate mixture of primers comprising a mixture of 3 primers of the sequences G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 6), each of the 3 primers being present in about an equimolar quantity;

(22) a degenerate mixture of primers comprising a mixture of 12 primers of the sequences G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-

This Sheet Has Been Replaced



T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 7), each of the 12 primers being present in about an equimolar quantity; and

(23) a degenerate mixture of primers comprising a mixture of 48 primers of the sequences G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 8), each of the 48 primers being present in about an equimolar quantity.

IV. SPECIFIC EXAMPLES OF PREFERED EMBODIMENTS

Example 1:

Application of the Improved Method.

The improved method of the present invention is based upon the observation that virtually all eukaryotic mRNAs conclude with a poly(A) tail, but, unlike differential display (Liang, P. and A.B. Pardee (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967-971), the method of the present invention uses the specificity of primer binding to the tail only to fix a site on each mRNA, not to subdivide mRNAs into pools. The improved method is illustrated in three embodiments in Figures 1, 2 and 8.

In general, double-stranded cDNA is generated from poly(A)-enriched cytoplasmic RNA extracted from the tissue samples of interest using an equimolar mixture of all 48 5'-biotinylated anchor primers of a set to initiate reverse transcription (Figures 2 and 8) (Gubler, U. and B. Hoffman (1983) A simple and very efficient method for generating cDNA libraries. Gene 25:263-269) (Schibler, K., M. Tosi, A.C. Pittet, L. Fabiani and P.K. Wellauer (1980) Tissue-specific expression of mouse amylase genes. J. Mol. Biol. 142:93-116). One such suitable set is A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 5), where V is A, C or G and N is A, C, G or T. One member of this mixture of 48 anchor 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, resulting in biotinylated double stranded cDNA.

This Sheet Has Been Replaced

Each biotinylated double stranded cDNA sample was cleaved with the restriction endonuclease MspI, which recognizes the sequence CCGG. The 3' fragments of cDNA were then isolated by capture of the biotinylated cDNA fragments on a streptavidin-coated substrate. Suitable streptavidin-coated substrates include microtitre plates, PCR tubes, polystyrene beads, paramagnetic polymer beads and paramagnetic porous glass particles. A preferred streptavidin-coated substrate is a suspension of paramagnetic polymer beads (Dynal, Inc., Lake Success, NY).

After washing the streptavidin-coated substrate and captured biotinylated cDNA fragments, the cDNA fragment product was released by digestion with NotI, which cleaves at an 8-nucleotide sequence within the anchor primers but rarely within the mRNA-derived portion of the cDNAs. The 3' MspI-NotI fragments, which are of uniform length for each mRNA species, were directionally ligated into ClaI-, NotI-cleaved plasmid pBC SK$^+$ (Stratagene, La Jolla, CA) in an antisense orientation with respect to the vector's T3 promoter, and the product used to transform Escherichia coli SURE cells (Stratagene). The ligation regenerates the NotI site, but not the MspI site. Each library contained in excess of 5 x 10^5 recombinants to ensure a high likelihood that the 3' ends of all mRNAs with concentrations of 0.001% or greater were multiply represented. Plasmid preps (Qiagen) were made from the cDNA library of each sample under study.

An aliquot of each library was digested with MspI, which effects linearization by cleavage at several sites within the parent vector while leaving the 3' cDNA inserts and their flanking sequences, including the T3 promoter, intact. The product was incubated with T3 RNA polymerase (MEGAscript kit, Ambion) to generate antisense cRNA transcripts of the cloned inserts containing known vector sequences abutting the MspI and NotI sites from the original cDNAs.

This step avoids contamination of each cRNA sample 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 MspI between

its ClaI and NotI sites and, therefore, the MspI digestion step eliminated the 5' tag from cRNAs transcribed from insertless plasmids, rendering them inert in the product amplification steps described below. Plasmid DNA was removed from the mixture of antisense cRNA transcripts by incubation with RNase-free DNase.

At this stage, each of the cRNA preparations was processed in a three-step fashion. In step one, 250ng of cRNA was converted to first-strand cDNA using the 5' RT primer (5PRIMER in Figures 1 and 2 and 8) A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G, (SEQ ID NO: 14). In step two, 400 pg of cDNA product was used as PCR template in four separate reactions with each of the four 5' PCR primers of the form G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N (SEQ ID NO: 22), each paired with an "universal" 3' PCR primer G-A-G-C-T-C-C-A-C-C-G-C-G-G-T (SEQ ID NO: 47), using the program

94 degrees Celsius, 15 seconds;
65 degrees Celsius, 15 seconds;
72 degrees Celsius, 60 seconds;
20 cycles.

In step three, the product of each subpool was further divided into 64 subsubpools (2ng in 20µl) for the second PCR reaction, with 100 ng each of the fluoresceinated "universal" 3' PCR primer, the oligonucleotide G-A-G-C-T-C-C-A-C-C-G-C-G-G-T (SEQ ID NO: 47) conjugated to 6-FAM and the appropriate 5' PCR primer of the form C-G-A-C-G-G-T-A-T-C-G-G-N-N-N-N (SEQ ID NO:25), using the program

94 degrees Celsius, 15 seconds;
X degrees Celsius, 15 seconds;
72 degrees Celsius, 30 seconds,
30 cycles,

that included an annealing step at a temperature X slightly above the T_m of each 5' PCR primer to minimize artifactual mispriming and promote high fidelity copying. Each polymerase chain reaction step was performed in the presence of TaqStart antibody (Clonetech).

The products from the final polymerase chain reaction step for each of the tissue samples were resolved on a series of denaturing DNA sequencing gels using the automated ABI Prizm 377 sequencer. Data were collected using the GeneScan software package (ABI) and normalized for amplitude and migration. Complete execution of this series of reactions generated 64 product subpools for each of the four pools established by the N_1 5' PCR primers, for a total of 256 product subpools for the entire N_4 5' PCR primer set.

To summarize, in this embodiment of the improved method (Figure 2), reverse transcriptase was used to generate a cDNA pool from cRNA with a non-parsing primer 5' RT primer of the form 5PRIMER (SEQ ID NO:14), Taq DNA polymerase was employed in PCR (20 cycles) to generate double stranded cDNA subpools with the 5' PCR primer 5PRIMERN1 (SEQ ID NO:11) as 5'-PCR primer and 3' PCR primer (SEQ ID NO:47). The final PCR was carried out for 30 cycles using 2ng of DNA template and 100ng of each $5PRIMER_3N_1N_2N_3N_4$ primer (SEQ ID NO: 25) and 3' PCR primer (SEQ ID NO:47) conjugated to 6-FAM.

Two mRNA samples from serum-starved (Figure 3, panel A) and serum-added (Figure 3, panel B) human MG63 osteosarcoma cells were analyzed. The data shown were generated with a 5'-PCR primer (C-G-A-C-G-G-T-A-T-C-G-G-G-G-T-G, SEQ ID NO: 42) paired with the "universal" 3' primer (SEQ ID NO:47) labeled with 6-carboxyfluorescein (6FAM, ABI) at the 5' terminus. PCR reaction products were resolved by gel electrophoresis on 4.5% acrylamide gels and fluorescence data acquired on ABI377 automated sequencers. Data were analyzed using Genescan software (Perkin-Elmer). In the three panels shown above, relative abundance of labeled PCR products is plotted (Y-axis = relative fluorescence units) versus product length in base pairs. The high reproducibility of the method is shown in the bottom panel, which shows data from panels (A) and (B) overlaid using Genescan software for comparison of relative expression levels between samples.

The major application of the present invention 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. The

majority of products from the pair of samples co-migrated and were of comparable amplitudes. Fewer than 10% of the products had amplitudes that differed by a factor of two or more, and these were approximately equally distributed between species induced by serum replenishment and those repressed by replenishment.

Many products, some of which were differentially represented in the two panels, appeared to migrate in positions coincident with predicted DSTs based on data extracted from GenBank, thus had candidate identities. To test these candidate identities, oligonucleotides were synthesized corresponding to the 5PRIMER$_3$N$_1$N$_2$N$_3$N$_4$ (SEQ ID NO: 25) for each candidate extended at the 3' end with an additional 14 nucleotides from the sequences adjacent to the terminal MspI sites in the GenBank sequences. These were paired with the fluorescent 3PRIMER (SEQ ID NO: 47) in PCRs using the N_1 cDNA as substrate.

Example 2:

Parsing Specificity In Embodiments of the Method Using One PCR Step and Two PCR Steps: Analysis Of PCR Products.

The advantages of the embodiments of basic method that include two PCR steps were demonstrated using serum-starved and serum-added MG63 cells. For the two PCR step variant of the basic method (Figure 1 and Figure 2), reverse transcriptase was used to generate a cDNA pool from cRNA with a non-parsing primer (N0) of the form 5PRIMER (SEQ ID NO: 14); Taq DNA polymerase was then employed in PCR (20 cycles) to generate double stranded cDNA subpools with 5PRIMERN1 (SEQ ID NO: 22) as 5'-PCR primer and 3' PCR primer (SEQ ID NO: 47). In the one PCR step modification, reverse transcriptase was used to generate 4 cDNA subpools from cRNA by initiating transcription with one of the four N1 primers of the form 5PRIMERN1 (SEQ ID NO: 22). In both methods, the final PCR was carried out for 30 cycles using 2ng of DNA template and 100ng of each 5' PCR primer (SEQ ID NO: 25) and 6-FAM labeled 3' PCR primer (SEQ ID NO:47).

Labeled PCR fragments were resolved by electrophoresis on automated DNA sequencers (ABI377) and analyzed by Genescan software. The results are presented

in Figure 4. Data from primer 109T (C-G-A-C-G-G-T-A-T-C-G-G-T-G-C-A, SEQ ID NO: 43) and 45A (C-G-A-C-G-G-T-A-T-C-G-G-A-G-C-A, SEQ ID NO: 44), which differ only at the N1 position (in bold), are shown for both serum starved (Figs. 4A, 4C, 4E and 4G) and serum added (Figs. 4B, 4D, 4F and 4H) samples.

The PCR products generated with 109T and 45A appear to be nearly identical from templates produced by the one PCR step variant (compare Fig. 4A to Fig. 4C, and Fig. 4B to Fig. 4D). In contrast, the products detected following PCR from templates produced using the two PCR step method are overall quite distinct (compare Fig. 4E to Fig. 4G, and Fig. 4F to Fig. 4H). The two PCR step embodiment of the method thus provides a substantial improvement over the closest previously available method.

Example 3:

Parsing Specificity In Embodiments of the Method Using One PCR Step and Two PCR Steps: Cloning And Sequence Data.

The method of the present invention was performed on serum-starved and serum-treated MG63 cells using either the one PCR step (Table I) or two PCR step (Table II) embodiments. In the experiment shown in Table I, reverse transcriptase was used to generate four cDNA subpools from cRNA by initiating transcription with one of the set of four N1 5' PCR primers (SEQ ID NO: 22). For Table II, reverse transcriptase was used to generate a cDNA pool from cRNA with a non-parsing 5' RT primer (SEQ ID NO: 14). Taq DNA polymerase was used in PCR (20 cycles) to generate double stranded cDNA subpools with 5' PCR primer (SEQ ID NO: 22) and as 3' PCR primer (SEQ ID NO: 47). The final PCR in both Table I and Table II was performed identically with the complete series of 256 5'-PCR primers paired (SEQ ID NO: 25) with 6FAM-labeled 3' PCR primer (SEQ ID NO: 47) using 2ng input cDNA template. From the PCR reaction displays, differentially regulated molecules were identified and isolated for cloning and sequencing purposes.

DNA sequence data was obtained for individual clones and gene identification determined following database searches using the BLAST algorithm. In the tables, clones found to be exact matches to known human genes are listed by gene name and

GenBank locus ID. The fidelity of the parsing step using 5PRIMERN1 (SEQ ID NO: 22) in either reverse transcription (Table I) or PCR reactions (Table II) was assessed by tabulating the sequence match of the clone at the N1 position to the GenBank sequence. In the two-step method, 5/22 clones matched correctly at the N1 position (essentially at random), whereas with the three-step procedure, all clones were found to match correctly with the corresponding GenBank sequence data.

Table I:

PARSING SPECIFICITY WITH ONE PCR STEP

GENE NAME	GenBank LOCUS ID	N1 POSITION MATCH
Nma	HSU23070	YES
CDE1 binding protein	HSCDEIBPA	YES
Laminin receptor homolog	S35960	YES
U1 snRNP-specific C protein	HSU1RNPC	YES
Ubiquitin	HSUBA52P	YES
MAD-3	HUMMAD3A	NO
α-tubulin	HSTUBB2	NO
Id1	HSID1	NO
NNMT	HSNNMT2	NO
BFGF	HUMGFB	NO
SC35	HUMSC35A	NO
Ribosomal protein S14	HUMRPS14	NO
Ribosomal protein L30	HUMRPL30A	NO
Na/K ATPase B3	HSU51478	NO
Ribosomal protein L37A	HSRPL37A	NO
IRF-2	HSIRF2	NO
SRp20	HUMSRP20	NO
Glyoxalase II	HSHAGH1	NO
pim-1 oncogene	HUMPIM1	NO
Endothelin-1	HUMEDN1B	NO
Metallothionein II	HUMMETIIPS	NO
CRP3 homolog	S63168	NO

Table II:

PARSING SPECIFICITY WITH TWO PCR STEPS

GENE NAME	GenBank LOCUS ID	N1 POSITION MATCH
MAD-3	HUMMAD3A	YES
Id1	HSID1	YES
Na/K ATPase B3	HSU51478	YES
pim-1 oncogene	HUMPIM1	YES
endothelin-1	HUMEDN1B	YES
ribosomal protein S20	HUMRPS20	YES
ribosomal protein S10	HUMRPS10	YES
GADD45	HUMGADD45	YES
AP-2	HSAP2	YES
beta-2 microglobulin	HUMB2MO2	YES
RDC-1	HSU67784	YES
56K autoantigen	HUM56KAUTO	YES
NFKB1	HSNFX24	YES
Lon protease-like protein	HSLONP	YES
nucleotide binding protein	U01833	YES
insulinoma gene	HUMIDB	YES
histone 2A.2	HUMH2A2A	YES

Note that five gene products highlighted in bold, **MAD-3, Id1, Na/K ATPase B3, pim-1 oncogene and endothelin-1** were isolated in both experiments, and in every case the two PCR step method produced a match at the N_1 position, while the one PCR step method did not. The two PCR step method thus provides a substantial improvement over the closest previously available method.

Example 4:

Improved Resolution Obtained Using Biotinylated Anchor Primers

As noted above, in one preferred embodiment, anchor primers are biotinylated at their 5' end (compare Figures 1 and 2). Biotinylated cDNA fragments can be captured using a streptavidin-coated substrate, preferably streptavidin-coated paramagnetic beads (Dynal). Figure 5 compares the results from the standard basic method to those obtained using anchor primers labeled with magnetic beads.

cDNA libraries were constructed using the standard technique (as outlined in Figure 1) and the magnetic bead alternative embodiment (see Figure 2) from 2 μg mRNA aliquots from five separate samples of striatum from haloperidol treated mice taken in a time series (0, 0.75, 7 hours, 10 and 14 days). The results are shown in Figure 5A-E (standard) and 5F-J (magnetic bead). The results from 5'PCR primer 170G (C-G-A-C-G-G-T-A-T-C-G-G-G-G-T, SEQ ID NO: 45) and the 6-FAM labeled 3'PCR primer (SEQ ID NO: 47) are shown in both cases for comparison. Relative abundance of labeled PCR products is plotted (Y-axis arbitrary fluorescence units) versus PCR product length (base pairs). The data from the magnetic bead libraries (Figure 5F-J) show greater reproducibility across samples in the time series (both in similarity of fragments and consistency of intensity values) and fewer apparently spurious short (100-125 bp) fragments compared to the data from the standard library technique (Figure 5A-E).

Example 5:

Demonstration of linearity in the three-step method:

Relationship of PCR product peak height to input cRNA concentration.

To determine the linear amplification range, a single cRNA species was spiked into 4 independent cRNA pools and processed by the method of the present invention as shown in Figure 2. The results are shown in Figure 6. The peak height (in relative fluorescence units) corresponding to the synthetic RNA was measured and plotted versus input concentration for the 4 samples. Data shown are averages from triplicate determinations; the error bars indicate the range of ± one standard error of the mean.

A SalI-NotI cDNA fragment (SEQ ID NO: 51) was cloned into the library

vector pBCSK+, linerarized and cRNA produced by transcription from the T3 promoter synthetic cRNA was constructed to give rise to a peak of known size (492bp) in PCR. Varying amounts of cRNA (0, 25, 100, or 250pg) were introduced into a 250ng pool of cRNA prior to reverse transcription with the N_0 primer (SEQ ID NO:14). 400pg of cDNA was used as template for PCR reactions with 5' PCR primer (SEQ ID NO: 22) and 3' PCR primers (SEQ ID NO:47), respectively. A 2ng aliquot of cDNA was used in a final PCR with 5' PCR primer 221C (C-G-A-C-G-G-T-A-T-C-G-G-C-T-C-A, SEQ ID NO: 46) and 3' PCR primer (SEQ ID NO:47). The results depicted in Figure 6 demonstrate that for a given tissue type, the peak height of the PCR product is proportional to the input RNA concentration.

The foregoing is intended to be illustrative of the present invention, but not limiting. Numerous variations and modifications of the present invention may be effected without departing from the true spirit and scope of the invention.

We claim:

1. An improved method for simultaneous sequence-specific identification of mRNAs in an mRNA population comprising the steps of:

(a) preparing a double-stranded cDNA population from an mRNA population using a mixture of anchor primers, each anchor primer having a 5' terminus and a 3' terminus and including: (i) a tract of from 7 to 40 T residues; (ii) a site for cleavage by a first restriction endonuclease that recognizes more than six bases, the site for cleavage being located towards the 5'-terminus relative to the tract of T residues; (iii) a first stuffer segment of from 4 to 40 nucleotides, the first stuffer segment being located towards the 5'-terminus relative to the site for cleavage by the first restriction endonuclease; (iv) a second stuffer segment interposed between the site for cleavage by a first restriction endonuclease that recognizes more than six bases and the tract of T residues, and (v) phasing residues located at the 3' terminus of each of the anchor primers selected from the group consisting of -V, -V-N, and -V-N-N, wherein V is a deoxyribonucleotide selected from the group consisting of A, C, and G; and N is a deoxyribonucleotide selected from the group consisting of A, C, G, and T, the mixture including anchor primers containing all possibilities for V and N;

(b) cleaving the double-stranded cDNA population with the first restriction endonuclease and a second restriction endonuclease, the second restriction endonuclease recognizing a four-nucleotide sequence, to form a population of double-stranded cDNA molecules having first and second termini, respectively;

(c) inserting each double-stranded cDNA molecule from step (b) into a vector in an orientation that is antisense with respect to a bacteriophage-specific promoter within the vector to form a population of constructs containing the inserted cDNA molecules, thereby defining 5' and 3' flanking vector sequences adjacent to the 5' terminus of the sense strand of the inserted cDNA and the 3' terminus of the sense strand respectively, and said constructs having a 3' flanking vector sequence at least 15 nucleotides in length between said first restriction endonuclease site and a site defining transcription initiation in said promoter;

(d) transforming a host cell with the vector into which the cleaved cDNA has been inserted to produce vectors containing cloned inserts;

(e) generating linearized fragments containing the inserted cDNA molecules by digestion of the constructs produced in step (c) with at least one

This Sheet Has Been Replaced

restriction endonuclease that does not recognize sequences in either the inserted cDNA molecules or in the bacteriophage-specific promoter, but does recognize sequences in the vector, such that the resulting linearized fragments have a 5' flanking vector sequence of at least 15 nucleotides into the vector 5' to the double-stranded cDNA molecule's second terminus;

(f) generating a cRNA preparation of antisense cRNA transcripts by incubating the linearized fragments with a bacteriophage-specific RNA polymerase capable of initiating transcription from the bacteriophage-specific promoter;

(g) generating first-strand cDNA by transcribing the cRNA using a reverse transcriptase and a 5' RT primer being 15 to 30 nucleotides in length and comprising a nucleotide sequence that is complementary to the 5' flanking vector sequence;

(h) generating a first set of PCR products by dividing the first-strand cDNA into a first series of subpools and using the first-strand cDNA as templates for a first polymerase chain reaction with a first 3' PCR-primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a first 5' PCR-primer defined as having a 3'-terminus consisting of $-N_1$, wherein "N" is one of the four deoxyribonucleotides A, C, G, or T, the first 5' PCR-primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the first 5' PCR-primer's complementarity extending into one nucleotide of the insert-specific nucleotides of the cRNA, wherein a different one of the first 5' PCR primers is used in each of four different subpools;

(i) generating a second set of PCR products by further dividing the first set of PCR products in each of the first series of subpools into a second series of subpools and using the first set of PCR products as templates for a second polymerase chain reaction with a second 3' PCR primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a second 5' PCR primer defined as having a 3'-terminus consisting of $-N_1-N_x$, wherein N_1 is identical to the N_1 used in the first polymerase chain reaction for that subpool, "N" is as is step h, and "x" is an integer from 1 to 5, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking

vector sequence with the primer's complementarity extending across into the insert-specific nucleotides of the cRNA in a number of nucleotides equal to "x" + 1, wherein a different one of the second 5' PCR primers is used in different subpools of the second series of subpools and wherein there are 4^x subpools in the second series of subpools for each of the subpools in the first set of subpools; and

(j) resolving the second set of PCR products to generate a display of sequence-specific products representing the 3'-ends of mRNAs present in the mRNA population.

2. The method of claim 1 wherein a biotin moiety is conjugated to the anchor primers.

3. The method of claim 2 wherein the biotin moiety is conjugated to the 5' terminus of the anchor primer.

4. The method of claim 2 wherein the first restricted cDNA is separated from the remainder of the cDNA in step b of claim 1 by contacting the first restricted cDNA with a streptavidin-coated substrate.

5. The method of the claim 1 wherein the 3 nucleotides at the 3' end of the first 5' PCR primer are joined by phosphorothioate linkages.

6. The method of the claim 1 wherein the 3 nucleotides at the 3' end of the second 5' PCR primer are joined by phosphorothioate linkages.

7. The method of the claim 1 wherein the 3 nucleotides at the 3' end of the first and second 5' PCR primers are joined by phosphorothioate linkages.

8. The method of claim 1 wherein one of the primers for the second PCR reaction is conjugated to a fluorescent label.

9. The method of claim 8 wherein the fluorescent label is selected from the group consisting of

spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 6-carboxylic acid, 3',6'-dihydroxy-6-carboxyfluorescein;

spiro(isobenzofuran-1(3H),9'-(9H)-xanthen)-3-one, 5-carboxylic acid, 3',6'-dihydroxy-5-carboxyfluorescein;

spiro(isobenzofuran-1(3H), 9'-(9H)-xanthen)-3-one, 3',6'-dihydroxy-fluorescein;

9-(2,5-dicarboxyphenyl)-3,6- bis(dimethylamino)-xanthylium

6-carboxytetramethylrhodamine;

3,6-diamino-9-(2-carboxyphenyl)-xanthylium;
spiro[isobenzofuran-1(3H), 9'-xanthene]-6-carboxylic acid,5'-dichloro-3',6'-dihydroxy-2',7'-dimethoxy-3-oxo-;
1H,5H,11H,15H-xantheno[2,3,4-ij:5,6,7-i'j']diquinolizin- 8-ium, -(2,4-disulfophenyl)-2,3,6,7,12,13,16,17-octahydro-, inner salt;
6-((4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-s-indacene-3-propionyl) amino) hexanoic acid;
6-((4,4-difluoro-1,3-dimethyl-5-(4-methoxyphenyl)-4-bora-3a,4a-diaza-s-indacene-3-propionyl)amino)hexanoic acid;
6-(((4-(4,4-difluoro-5-(2-thienyl)-4-bora-3a,4a-diaza-s-indacene-3-yl) phenoxy)acetyl) amino)-hexanoic acid;
4,4-difluoro-4-bora-3a,4a-diaza-s-indacene-3-pentanoic acid;
4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-s-indacene-3-propanoic acid;
4,4-difluoro-5-phenyl-4-bora-3a,4a-diaza-s-indacene-3-propionic acid;
4,4-difluoro-5-(4-phenyl-1,3-butadienyl)-4-bora-3a,4a-diaza-s-indacene-3-propionic acid;
4,4-difluoro-5-styryl-4-bora-3a,4a-diaza-s-indacene-3-propionic acid;
6-(((4,4-difluoro-5-(2-thienyl)-4-bora-3a,4a-diaza-s-indacene-3-yl)styryloxy)acetyl) aminohexanoic acid;
6-(((4,4-difluoro-5-(2-pyrrolyl)-4-bora-3a,4a-diaza-s-indacene-3-yl) styryloxy)acetyl) aminohexanoic acid;
9-(2,4(or 2,5)-dicarboxyphenyl)-3,6- bis(dimethylamino)- xanthylium, inner salt; and
4, 7, 2', 4', 5', 7' hexachloro 6-carboxyfluorescein and 4, 7, 2', 7' tetrachloro 6-carboxyfluorescein.

10. The method of claim 1 wherein the host cell is an Escherichia coli cell.

11. The method of claim 1 wherein the phasing residues in step (a) are -V-N-N.

12. The method of claim 1 wherein the phasing residues in step (a) are -V-N.

13. The method of claim 1 wherein the phasing residues in step (a) are -V.

14. The method of claim 1 wherein the "x" in step (i) is 3.

15. The method of claim 1 wherein the "x" in step (i) is 1.

16. The method of claim 1 wherein the phasing residues in step (a) are -V-N-N and the "x" in step (i) is 3.

17. The method of claim 1 wherein the phasing residues in step (a) are –V and the "x" in step (i) is 2.

18. The method of claim 1 wherein the anchor primers each have 18 T residues in the tract of T residues.

19. The method of claim 1 wherein the first stuffer segment of the anchor primers is 14 residues in length.

20. The method of claim 1 wherein the sequence of the first stuffer segment is
G-A-A-T-T-C-A-A-C-T-G-G-A-A (SEQ ID NO: 2).

21. The method of claim 1 wherein the bacteriophage-specific promoter is selected from the group consisting of T3 promoter, T7 promoter and SP6 promoter.

22. The method of claim 1 wherein the bacteriophage-specific promoter is T3 promoter.

23. The method of claim 1 wherein the primer for priming of transcription of cDNA from cRNA has the sequence A-G-G-T-C-G-A-C-G-G-T-A-T-C-G-G (SEQ ID NO: 14).

24. The method of claim 1 wherein the vector is the plasmid pBC SK+ cleaved with ClaI and NotI and the 3' PCR primer in steps (h) and (i) is G-A-G-C-T-C-C-A-C-C-G-C-G-G-T (SEQ ID NO: 47).

25. The method of claim 1 wherein the vector is the plasmid pBC SK+ cleaved with ClaI and NotI and the 3' PCR primer in steps (h) and (i) is G-A-G-C-T-C-G-T-T-T-T-C-C-C-A-G (SEQ ID NO: 48).

26. The method of claim 1 wherein the second restriction endonuclease recognizing a four-nucleotide sequence is MspI.

27. The method of claim 1 wherein the second restriction endonuclease recognizing a four-nucleotide sequence is selected from the group consisting of MboI, DpnII, Sau3AI, Tsp509I, HpaII, BfaI, Csp6I, MseI, HhaI, NlaIII, TaqI, MspI, MaeII Sau3AI, BglII and HinPII.

28. The method of claim 1 wherein the first restriction endonuclease that recognizes more than six bases is selected from the group consisting of AscI, BaeI,

FseI, NotI, PacI, PmeI PpuMI, RsrII, SapI, SexAI, SfiI, SgfI, SgrAI, SrfI, Sse8387I and SwaI.

29. The method of claim 1 wherein the first restriction endonuclease that recognizes more than six bases is NotI.

30. The method of claim 1 wherein the restriction endonuclease used in step (e) has a nucleotide sequence recognition that includes the four-nucleotide sequence of the second restriction endonuclease used in step (b).

31. The method of claim 30 wherein the second restriction endonuclease is MspI and the restriction endonuclease used in step (e) is Sma I.

32. The method of claim 30 wherein the second restriction endonuclease is TaqI and the restriction endonuclease used in step (e) is XhoI.

33. The method of claim 30 wherein the second restriction endonuclease is HinPII and the restriction endonuclease used in step (e) is NarI.

34. The method of claim 30 wherein the second restriction endonuclease is MaeII and the restriction endonuclease used in step (e) is AatII.

35. The method of claim 30 wherein the second restriction endonuclease is Sau3AI and the restriction endonuclease used in step (e) is BglII.

36. The method of claim 30 wherein the second restriction endonuclease is NlaIII and the restriction endonuclease used in step (e) is NcoI.

37. A vector suitable for the practice of the method of claim 1 wherein the vector of step (c) is in the form of a circular DNA molecule having first and second vector restriction endonuclease sites flanking a vector stuffer sequence, and further comprising the step of digesting the vector with restriction endonucleases that cleave the vector at the first and second vector restriction endonuclease sites.

38. The vector of claim 37 wherein the vector stuffer sequence includes an internal vector stuffer restriction endonuclease site between the first and second vector restriction endonuclease sites.

39. The vector of claim 38 wherein the step (e) includes digestion of the vector with a restriction endonuclease which cleaves the vector at the internal vector stuffer restriction endonuclease site.

40. A vector chosen from the group consisting of plasmids pBC SK^+ /DGT1, pBS SK^+ /DGT2, pBS SK^+ /DGT3, pBC SK^+ /DGT4 and pBS SK^+ /DGT5.

41. A vector comprising a mutiple cloning site chosen from the group

consisting of SEQ ID NO: 9, SEQ ID NO: 10, SEQ ID NO: 11, SEQ ID NO: 12 and SEQ ID NO: 13.

42. The method of claim 1 wherein the mRNA population has been enriched for polyadenylated mRNA species.

43. The method of claim 1 wherein the resolving in step (j) of the amplified fragments is conducted by electrophoresis to display the products.

44. The method of claim 43 wherein the intensity of products displayed after electrophoresis is about proportional to the abundances of the mRNAs corresponding to the products in the original mixture.

45. The method of claim 43 further comprising a step of determining the relative abundance of each mRNA in the original mixture from the intensity of the product corresponding to that mRNA after electrophoresis.

46. The method of claim 43 wherein the step of resolving the polymerase chain reaction amplified fragments by electrophoresis comprises electrophoresis of the fragments on at least two gels.

47. The method of claim 40 further comprising the steps of:

(k) eluting at least one cDNA corresponding to a mRNA from an electropherogram in which bands representing the 3'-ends of mRNAs present in the sample are displayed;

(l) amplifying the eluted cDNA in a polymerase chain reaction;

(m) cloning the amplified cDNA into a plasmid;

(n) producing DNA corresponding to the cloned DNA from the plasmid; and

(o) sequencing the cloned cDNA.

48. An improved method for simultaneous sequence-specific identification of mRNAs in a mRNA population comprising the steps of:

(a) isolating an mRNA population;

(b) preparing a double-stranded cDNA population from the mRNA population using a mixture of anchor primers, each anchor primer having a 5' terminus and a 3' terminus and including: (i) a tract of from 7 to 40 T residues; (ii) a site for cleavage by a first restriction endonuclease that recognizes eight bases, the site for cleavage being located towards the 5'-terminus relative to the tract of T residues; (iii) a first stuffer segment of from 4 to 40 nucleotides, the first stuffer segment being

This Sheet Has Been Replaced

located towards the 5'-terminus relative to the site for cleavage by the first restriction endonuclease; (iv) a second stuffer segment interposed between the site for cleavage by a first restriction endonuclease that recognizes more than six bases and the tract of T residues, and (v) phasing residues defined by one of -V, -V-N, or -V-N-N located at the 3' terminus of each of the anchor primers, wherein V is a deoxyribonucleotide selected from the group consisting of A, C, and G; and N is a deoxyribonucleotide selected from the group consisting of A, C, G, and T, the mixture including anchor primers containing all possibilities for V and N;

(c) cleaving the double-stranded cDNA population with the first restriction endonuclease and a second restriction endonuclease, the second restriction endonuclease recognizing a four-nucleotide sequence, to form a population of double-stranded cDNA molecules having first and second termini, respectively;

(d) inserting each double-stranded cDNA molecule from step (b) into a vector in an orientation that is antisense with respect to a T3 promoter within the vector to form a population of constructs containing the inserted cDNA molecules, thereby defining 5' and 3' flanking vector sequences adjacent to the 5' terminus of the sense strand of the inserted cDNA and the 3' terminus of the sense strand respectively, and said constructs having a 3' flanking vector sequence at least 15 nucleotides in length between said first restriction endonuclease site and a site defining transcription initiation in said promoter;

(e) transforming Escherichia coli with the vector into which the cleaved cDNA has been inserted to produce vectors containing cloned inserts;

(f) generating linearized fragments containing the inserted cDNA molecules by digestion of the constructs produced in step (c) with at least one restriction endonuclease that does not recognize sequences in either the inserted cDNA molecules or in the T3 promoter;

(g) generating a cRNA preparation of antisense cRNA transcripts by incubating the linearized fragments with a T3 RNA polymerase capable of initiating transcription from the T3 promoter;

(h) generating first-strand cDNA by transcribing the cRNA using a reverse transcriptase and a 5' RT primer being 15 to 30 nucleotides in length and comprising a nucleotide sequence that is complementary to the 5' flanking vector sequence;

This Sheet Has Been Replaced

(i) generating a first set of PCR products by dividing the first-strand cDNA into a first series of subpools and using the first-strand cDNA as templates for a first polymerase chain reaction with a first 3' PCR-primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the T3-specific promoter and a first 5' PCR-primer defined as having a 3'-terminus consisting of $-N_1$, wherein "N" is one of the four deoxyribonucleotides A, C, G, or T, the first 5' PCR-primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the first 5' PCR-primer's complementarity extending into one nucleotide of the insert-specific nucleotides of the cRNA, wherein a different one of the first 5' PCR primers is used in each of four different subpools;

(j) generating a second set of PCR products by further dividing the first set of PCR products in each of the first series of subpools into a second series of subpools and using the first set of PCR products as templates for a second polymerase chain reaction with a second 3' PCR primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the T3-specific promoter and a second 5' PCR primer defined as having a 3'-terminus consisting of $-N_1-N_x$, wherein N_1 is identical to the N_1 used in the first polymerase chain reaction for that subpool, "N" is as is step i, and "x" is an integer selected from the group consisting of 3 and 4, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the primer's complementarity extending across into the insert-specific nucleotides of the cRNA in a number of nucleotides equal to "x" + 1, wherein a different one of the second 5' PCR primers is used in different subpools of the second series of subpools and wherein there are 4^X subpools in the second series of subpools for each of the subpools.

(k) resolving the second set of PCR products to generate a display of sequence-specific products representing the 3'-ends of mRNAs present in the mRNA population.

49. The method of claim 48 wherein the mixture of 48 anchor primers have the sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 5).

50. The method of claim 48 wherein the mixture of 48 anchor primers have the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N-N (SEQ ID NO: 8).

51. The method of claim 48 wherein the mixture of 12 anchor primers have the sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 4).

52. The method of claim 48 wherein the mixture of 12 anchor primers have the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V-N (SEQ ID NO: 7).

53. The method of claim 48 wherein the mixture of 3 anchor primers have the sequence A-A-C-T-G-G-A-A-G-A-A-T-T-C-G-C-G-G-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 3).

54. The method of claim 48 wherein the mixture of 3 anchor primers have the sequence G-A-A-T-T-C-A-A-C-T-G-G-A-A-G-C-G-G-C-C-C-G-C-A-G-G-A-A-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-T-V (SEQ ID NO: 6).

55. The method of claim 48 wherein the first restriction endonuclease is MspI and the second restriction endonuclease is NotI.

56. The method of claim 48 wherein the first 5' PCR-primer is G-G-T-C-G-A-C-G-G-T-A-T-C-G-G-N (SEQ ID NO: 22).

57. The method of claim 48 wherein the first 3' PCR primer and the second 3' PCR-primer are G-A-G-C-T-C-C-A-C-C-G-C-G-G-T (SEQ ID NO: 47).

58. The method of claim 48 wherein the "x" in step (j) is 3.

59. The method of claim 48 wherein the "x" in step (j) is 4.

60. A method for detecting a change in the pattern of mRNA expression in a tissue associated with a physiological or pathological change comprising the steps of.

(a) obtaining a first sample of a tissue that is not subject to the physiological or pathological change;

(b) isolating an mRNA population from the first sample;

(c) determining the pattern of mRNA expression in the first sample of the tissue by performing steps (a)-(j) of claim 1 to generate a first display of sequence-specific products representing the 3'-ends of mRNAs present in the first sample;

This Sheet Has Been Replaced

(d) obtaining a second sample of a tissue that has been subject to the physiological or pathological change;

(e) isolating an mRNA population from the second sample;

(f) determining the pattern of mRNA expression in the second sample of the tissue by performing steps (a)-(j) of claim 1 to generate a second display of sequence-specific products representing the 3'-ends of mRNAs present in the second sample; and

(g) comparing the first and second displays to determine the effect of the physiological or pathological change on the pattern of mRNA expression in the tissue.

61. The method of claim 60 wherein the physiological or pathological change is selected from the processes mediated by transcription factors, intracellular second messengers, hormones, neurotransmitters, growth factors, neuromodulators, cell-cell contact, cell-substrate contact, cell-extracellular matrix contact and contact between cell membranes and cytoskeleton.

62. The method of claim 60 wherein the tissue is derived from the central nervous system.

63. The method of claim 62 wherein the physiological or pathological change is selected from the group consisting of Alzheimer's disease, parkinsonism, ischemia, alcohol addiction, drug addiction, schizophrenia, amyotrophic lateral sclerosis, multiple sclerosis, depression, and bipolar manic-depressive disorder.

64. The method of claim 62 wherein the physiological or pathological change is associated with learning or memory, emotion, glutamate neurotoxicity, feeding behavior, olfaction, vision, movement disorders, viral infection, electroshock therapy, or the administration of a drug or toxin.

65. The method of claim 60 wherein the physiological or pathological change is selected from the group consisting of circadian variation, aging, and long term potentiation.

66. The method of claim 60 wherein the tissue is derived from a structure within the central nervous system selected from the group consisting of retina, cerebral cortex, olfactory bulb, thalamus, hypothalamus, anterior pituitary, posterior pituitary, hippocampus, nucleus accumbens, amygdala, striatum, cerebellum, brain stem, suprachiasmatic nucleus, and spinal cord.

67. The method of claim 60 wherein the tissue is normal or neoplastic tissue from an organ or organ system selected from the group consisting of the cardiovascular system, the pulmonary system, the digestive system, the peripheral nervous system, the liver, the kidney, skeletal muscle, and the reproductive system.

68. The method of claim 60 wherein the tissue is normal or neoplastic tissue that comprises cells taken or derived from an organ or organ system selected from the group consisting of the cardiovascular system, the lymphatic system, the respiratory system, the digestive system, the peripheral nervous system, the central nervous system, the enteric nervous system, the endocrine system, the integument (including skin, hair and nails), the skeletal system (including bone and muscle), the urinary system and the reproductive system.

69. The method of claim 60 wherein the tissue is normal or neoplastic tissue that comprises cells taken or derived from the group consisting of epithelia, endothelia, mucosa, glands, blood, lymph, connective tissue, cartilage, bone, smooth muscle, skeletal muscle, cardiac muscle, neurons, glial cells, spleen, thymus, pituitary, thyroid, parathyroid, adrenal cortex, adrenal medulla, adrenal cortex, pineal, skin, hair, nails, teeth, liver, pancreas, lung, kidney, bladder, ureter, breast, ovary, uterus, vagina, testes, prostate, penis, eye and ear.

70. A method of detecting a difference in action of a drug to be screened and a known compound comprising the steps of-

(a) obtaining a first sample of tissue from an organism treated with a compound of known physiological function;

(b) isolating an mRNA population from the first sample;

(c) determining the pattern of mRNA expression in the first sample of the tissue by performing steps (a)-(j) of claim 1 to generate a first display of sequence-specific products representing the 3'-ends of mRNAs present in the first sample;

(d) obtaining a second sample of tissue from an organism treated with a drug to be screened for a difference in action of the drug and the known compound;

(e) isolating an mRNA population from the first sample;

(f) determining the pattern of mRNA expression in the second sample of the tissue by performing steps (a)-(j) of claim 1 to generate a second display of sequence-specific products representing the 3'-ends of mRNAs present in the second sample; and

(g) comparing the first and second displays in order to detect the presence of

mRNA species whose expression is not affected by the known compound but is affected by the drug to be screened, thereby indicating a difference in action of the drug to be screened and the known compound.

71. The method of claim 70 wherein the drug to be screened is selected from the group consisting of antidepressants, neuroleptics, tranquilizers, anticonvulsants, monoamine oxidase inhibitors, and stimulants.

72. The method of claim 70 wherein the drug to be screened is selected from the group consisting of anti-parkinsonism agents, skeletal muscle relaxants, analgesics, local anesthetics, cholinergics, antiviral agents, antispasmodics, steroids, and non-steroidal anti-inflammatory drugs.

73. A database comprising the data produced by the quantitation of the display of sequence-specific products of claim 1.

74. The database of claim 1 further comprising data concerning sequence relationships, gene mapping and cellular distributions.

75. A method for recognizing sequence identities and similarities between the sequence of 3'-ends of mRNA molecules present in a sample and a database of sequences, comprising the steps of :

(a) preparing a double-stranded cDNA population from an mRNA population using a mixture of anchor primers, each anchor primer having a 5' terminus and a 3' terminus and including: (i) a tract of from 7 to 40 T residues; (ii) a site for cleavage by a first restriction endonuclease that recognizes more than six bases, the site for cleavage being located towards the 5'-terminus relative to the tract of T residues; (iii) a first stuffer segment of from 4 to 40 nucleotides, the first stuffer segment being located towards the 5'-terminus relative to the site for cleavage by the first restriction endonuclease; (iv) a second stuffer segment interposed between the site for cleavage by a first restriction endonuclease that recognizes more than six bases and the tract of T residues, and (v) phasing residues located at the 3' terminus of each of the anchor primers selected from the group consisting of -V, -V-N, and -V-N-N, wherein V is a deoxyribonucleotide selected from the group consisting of A, C, and G; and N is a deoxyribonucleotide selected from the group consisting of A, C, G, and T, the mixture including anchor primers containing all possibilities for V and N;

(b) cleaving the double-stranded cDNA population with the first restriction endonuclease and a second restriction endonuclease, the second restriction

This Sheet Has Been Replaced

endonuclease recognizing a four-nucleotide sequence, to form a population of double-stranded cDNA molecules having first and second termini, respectively;

(c) inserting each double-stranded cDNA molecule from step (b) into a vector in an orientation that is antisense with respect to a bacteriophage-specific promoter within the vector to form a population of constructs containing the inserted cDNA molecules, thereby defining 5' and 3' flanking vector sequences adjacent to the 5' terminus of the sense strand of the inserted cDNA and the 3' terminus of the sense strand respectively, and said constructs having a 3' flanking vector sequence at least 15 nucleotides in length between said first restriction endonuclease site and a site defining transcription initiation in said promoter;

(d) transforming a host cell with the vector into which the cleaved cDNA has been inserted to produce vectors containing cloned inserts;

(e) generating linearized fragments containing the inserted cDNA molecules by digestion of the constructs produced in step (c) with at least one restriction endonuclease that does not recognize sequences in either the inserted cDNA molecules or in the bacteriophage-specific promoter, but does recognize sequences in the vector, such that the resulting linearized fragments have a 5' flanking vector sequence of at least 15 nucleotides into the vector 5' to the double-stranded cDNA molecule's second terminus;

(f) generating a cRNA preparation of antisense cRNA transcripts by incubating the linearized fragments with a bacteriophage-specific RNA polymerase capable of initiating transcription from the bacteriophage-specific promoter;

(g) generating first-strand cDNA by transcribing the cRNA using a reverse transcriptase and a 5' RT primer being 15 to 30 nucleotides in length and comprising a nucleotide sequence that is complementary to the 5' flanking vector sequence;

(h) generating a first set of PCR products by dividing the first-strand cDNA into a first series of subpools and using the first-strand cDNA as templates for a first polymerase chain reaction with a first 3' PCR-primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a first 5' PCR-primer defined as having a 3'-terminus consisting of $-N_1$, wherein "N" is one of the four deoxyribonucleotides A, C, G, or T, the first 5' PCR-primer being 15 to 30 nucleotides in length and

This Sheet Has Been Replaced

complementary to the 5' flanking vector sequence with the first 5' PCR-primer's complementarity extending into one nucleotide of the insert-specific nucleotides of the cRNA, wherein a different one of the first 5' PCR primers is used in each of four different subpools;

(i) generating a second set of PCR products by further dividing the first set of PCR products in each of the first series of subpools into a second series of subpools and using the first set of PCR products as templates for a second polymerase chain reaction with a second 3' PCR primer of 15 to 30 nucleotides in length that is complementary to 3' flanking vector sequences between the first restriction endonuclease site and the site defining transcription initiation by the bacteriophage-specific promoter and a second 5' PCR primer defined as having a 3'-terminus consisting of $-N_1-N_x$, wherein N_1 is identical to the N_1 used in the first polymerase chain reaction for that subpool, "N" is as is step h, and "x" is an integer from 1 to 5, the primer being 15 to 30 nucleotides in length and complementary to the 5' flanking vector sequence with the primer's complementarity extending across into the insert-specific nucleotides of the cRNA in a number of nucleotides equal to "x" + 1, wherein a different one of the second 5' PCR primers is used in different subpools of the second series of subpools and wherein there are 4^X subpools in the second series of subpools for each of the subpools in the first set of subpools;

(j) resolving the second set of PCR products to generate a display of sequence-specific products representing the 3'-ends of mRNAs present in the mRNA population.

(k) eluting at least one cDNA corresponding to a mRNA from an electropherogram in which bands representing the 3'-ends of mRNAs present in the sample are displayed;

(l) amplifying the eluted cDNA in a polymerase chain reaction;

(m) cloning the amplified cDNA into a plasmid;

(n) producing DNA corresponding to the cloned DNA from the plasmid;

(o) determining the sequence of the cloned cDNA;

(p) determining corresponding nucleotide sequences from a database of nucleotide sequences, said corresponding nucleotide sequences being delimited by the most distal recognition site for the second endonuclease and the beginning of the poly(A) tail; and

(q) comparing the sequence of the cloned cDNA to the corresponding nucleotide sequences, thereby recognizing sequence identities and similarities between the sequence of 3'-ends of mRNA molecules present in a sample and a database of sequences.

76. The method of claim 76 further comprising the step of

(r) comparing the length and amount of the PCR products in a two dimensional graphical display.

77. The method of claim 76 further comprising the steps of

(s) determining the expected length of the corresponding nucleotide sequence, which is equal to the sum of the lengths of the corresponding nucleotide sequence determined from the database, the length of the 5'PCR sequence hybridizable to vector sequence, the length of the remaining anchor primer sequence, an intervening segment of vector sequence and the length of the 3'PCR sequence hybridizable to vector sequence; and

(t) comparing the length of the PCR product to the determined expected length of the corresponding nucleotide sequence, wherein the expected length of corresponding nucleotide sequence is indicated in the two dimensional graphical display by the use of a graphical symbol or text character.

78. A method for recognizing sequence identities and similarities between the sequence of a cDNA fragment corresponding to a mRNA molecule present in a sample and a database of sequences, comprising the steps of:

eluting a cDNA fragment corresponding to a mRNA molecule present in a sample;

amplifying the eluted cDNA fragment in a polymerase chain reaction to produce an amplified cDNA fragment;

cloning the amplified cDNA fragment into a plasmid;

producing a DNA molecule corresponding to the cloned cDNA fragment;

sequencing the produced DNA molecule, thereby determining the sequence of the eluted cDNA fragment; and

comparing the sequence of the eluted cDNA fragment to the sequences in a database thereby recognizing sequence identities and similarities.

79. The method of claim 78 wherein the step of comparing the sequence of the eluted cDNA fragment to the sequences in a database is performed using a

This Sheet Has Been Replaced

computer.

80. The method of claim 78 comprising the additional step of displaying the results of the comparison graphically.

81. A method for recognizing sequence identities and similarities between the sequence of a cDNA fragment corresponding to a mRNA molecule present in a sample and a database of sequences, comprising the steps of :

eluting a cDNA fragment corresponding to a mRNA molecule present in a sample, where the cDNA fragment has a length determined by the position of a restriction endonuclease recognition site and a poly(A) tail of the mRNA molecule;

determining a partial sequence of the cDNA fragment by performing a polymerase chain reaction with a 5' PCR primer corresponding to the sequence of the restriction endonuclease recognition site and

comparing the determined partial sequence of the eluted cDNA fragment and the length of the cDNA fragment to the sequences in a database thereby recognizing sequence identities and similarities.

82. A method of producing a transformed polynucleotide sequence database entry, comprising the steps of:

choosing a source sequence from a polynucleotide sequence database entry;

locating a poly(A) tail sequence within the source sequence;

locating an endonuclease recognition site sequence within the source sequence that is closest to the first recognition site;

determining an index sequence consisting of about two to about six nucleotides adjacent to the endonuclease recognition site;

determining a correlate sequence within the source sequence, said correlate sequence including the sequence bounded by the poly(A) tail and the endonuclease recognition site and including at least part of the endonuclease recognition site;

determining the length of the correlate sequence; and

storing information concerning the location and sequence of the poly(A) tail, the location and sequence of the endonuclease recognition site, and the length of the correlate sequence in relation to the source sequence, thereby producing a transformed database entry.

83. The method of claim 82 further comprising the step of:

displaying graphically the length of the correlate sequence in relation to the

This Sheet Has Been Replaced

index sequence.

84. The method of claim 83 wherein the restriction endonuclease is chosen from the group consisting of MspI, TaqI and HinP1I.

85. A method of improving resolution of the length and amount of PCR products by diminishing background that is due to amplification of untargeted cDNAs comprising the steps of:

selecting a sample of a cRNA population, wherein each cRNA molecule comprises insert sequence and vector-derived sequence;

performing reverse transcription using a reverse transcription primer that hybridizes to the vector-derived sequence and that extends about five nucleotides to about six nucleotides into the insert sequence to produce a cDNA reverse transcription product;

subdividing the cDNA reverse transcription product;

performing at least one polymerase chain reaction using the subdivided cDNA reverse transcription product, a 3'PCR primer and a 5' PCR primer that hybridizes to the vector-derived sequence and extends about seven nucleotides to about nine nucleotides into the insert sequence to produce a PCR product, thereby diminishing background that is due to amplification of untargeted cDNAs.

86. The method of claim 85 wherein there are sixteen pools of reverse transcription reactions and there are sixteen different reverse transcription primers.

87. The method of claim 86 wherein there are 4^X subpools of polymerase chain reactions, where X is the difference between the number of nucleotides that the 5' PCR primer extends into the insert sequence and the number of nucleotides that the reverse transcription primer extends into the insert sequence.

This Sheet Has Been Replaced