Construction of a Set of Full-Length Enriched cDNA Libraries as Genomics Tools for *Xenopus Tropicalis* Research

FINAL

Jisong Peng[1], Bridget L. Riggs[1,2], Hajime Ogino[3] and Bruce Blumberg[1,*]

[1]*Department of Developmental and Cell Biology, 5205 McGaugh Hall, University of California, Irvine, California 92697-2300, USA;* [2]*Present address: Department of Pathology and Laboratory Medicine, UCLA Medical School, 10833 Le Conte Ave., LA, CA 90095-7300, USA and* [3]*Department of Biology, University of Virginia, Gilmer Hall, P.O. Box 400328 Charlottesville, VA 22904*

Abstract: A large variety of mammalian and non-mammalian animal models have been used in research designed to uncover fundamental mechanisms underlying development and disease. Genomics tools have become increasingly necessary for the molecular genetic analysis of important biological questions. However, there are few genomics resources available for the emerging vertebrate model *Xenopus tropicalis*. Here we discuss our approach towards making a collection of full-length cDNAs from *X. tropicalis* that will serve as a resource for EST sequencing, microarray development and large-scale functional genomics analysis of *Xenopus* development.

Key Words: cDNA, full-length, functional genomics, library construction, molecular interaction screen.

INTRODUCTION

In recent years, we have witnessed an explosion of DNA sequence data from human and other model organism genome projects. However, the rapid increases in nucleotide sequence data have not been matched by a correspondingly large increase in the identification of gene expression patterns or understanding of gene function. As complete genome sequences become available, it will be possible to study the regulation and expression of genes on a full-genome basis, rather than gene by gene. The various expressed sequence tag (EST) projects of human and model organisms have been a great boon to research. This is also the case for *Xenopus.* One persistent difficulty in gene identification from genomic sequences has been that EST sequences are biased toward the 3' end of mRNAs since cDNAs are rarely full length. This deficiency in 5' sequences makes it difficult to accurately predict transcriptional start sites. In large part, the 3' sequence bias occurs because the majority of cDNA libraries used in EST sequencing have been constructed for identifying individual genes, rather than large-scale sequencing. Those that have been made for large-scale sequencing are often normalized, which biases these libraries against complete sequences (see below). Our laboratory has undertaken to construct a set of full-length cDNAs from the emerging vertebrate model organism, *Xenopus tropicalis*. This brief review describes our approach to making these libraries and their potential uses in genomics and functional analyses.

*Address correspondence to this author at the Department of Developmental and Cell Biology, 5205 McGaugh Hall, University of California, Irvine, California 92697-2300, USA; Tel: 949 824 8573; Fax: 949 824 4709; E-mail: Blumberg@uci.edu

THE VALUE OF ESTs AND HIGH-QUALITY cDNA LIBRARIES

Experience has shown that animal models are important research tools for understanding development and diseases. Genomic tools, such as EST databases have led to rapid advances in the utility of model systems, particularly for gene identification and comparative genomics. However, aside from the mouse, and to a lesser extent zebrafish, the progress in developing state of the art genomics tools has been less than ideal for vertebrate model systems. Simple inspection of the EST databases shows that there are many fewer EST sequences available for most model organisms http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html, although the model organism ESTs have increased rapidly in the past two years. ESTs have become very important for genomics studies in humans as well as mouse and other model organisms. They are ideal resources with which to create microarrays and provide a rapid way to clone sequences of potential interest by sequence comparisons with databases, rather than by more laborious physical screening approaches. ESTs represent mRNAs that are expressed in the source tissues, thus they contain more "concentrated" information than genomic sequences. They are an important adjunct to genomic mapping studies, providing a rapid way to identify candidate genes in mapped regions. As genomic sequences become available from the *X. tropicalis* genome project that is underway at the US Department of Energy Joint Genome Institute (see Chapter 2, this issue), it will be advantageous to have ESTs that contain the 5' regions of mRNAs. Such 5' sequences facilitate the rapid delineation of putative promoter regions. The ability to identify promoter boundaries by sequence comparisons significantly reduces the effort required to generate suitable reporter constructs for promoter analysis. In turn, this accelerates the rate at which individual

1389-2029/03 $41.00+.00

researchers can characterize the regulatory regions of genes of interest using the powerful transgenic techniques now available in *Xenopus laevis* and *tropicalis*.

Our aim is to enhance the array of available reagents in the vertebrate model *X. tropicalis* by the production of a set of full-length cDNAs. The first step in this process is the construction of a group of cDNA libraries enriched in full-length cDNAs. This critical step will benefit the EST project by increasing the available sequence from the 5' end of transcripts as well as generating suitable templates for full-length cDNA sequencing.

The full-length cDNA collection will also facilitate *in vitro* and *in vivo* functional studies. One such application is the generation of cDNA microarrays, as has been pioneered by Pat Brown and colleagues [1, 2]. Hybridization analysis with fluorescently labeled probes from various sources permits sensitive detection of alternations in gene expression in response to experimental perturbation or mutations in *Xenopus.* Ken Cho and colleagues have developed such arrays from *X. laevis* (see Chapter 4, this issue). The *X. tropicalis* gene collection will serve as an ideal resource to produce an equivalent set of *X. tropicalis* microarrays.

For many years, the frog *X. laevis* has been favored by biologists to investigate the mechanisms of vertebrate embryonic development. Despite its numerous benefits for research in Cell and Developmental Biology, *X. laevis* is pseudotetraploid and takes nearly 2 years to reach sexual maturity, making it virtually impossible to use as a genetic system. Recently, investigators have focused on the closely-related *X. tropicalis*, which is the only diploid species in the *Xenopus* genus and therefore, better suited for genetic approaches [3]. *X. tropicalis* requires approximately 5 months to reach sexual maturity, making the generation of lines and other types of genetic studies feasible. A number of large-scale mutagenesis projects are underway in laboratories around the world. These screens should produce a large collection of *X. tropicalis* mutants for analysis. The combination of these mutants with *X. tropicalis* genomic and EST sequences will provide a valuable resource for genomics studies of this important vertebrate model animal.

Enhancing the Utility of the Libraries Constructed

In principle, most libraries of reasonable quality (i.e. complexity) are useful for EST sequencing. The *X. laevis* EST project has employed libraries from numerous investigators, which has enabled the rapid development of a valuable community resource. However, the further creation of a physical resource of full-length unique genes from the *X. laevis* ESTs is problematic, since these ESTs have been derived from libraries constructed in at least 11 different vectors. This also makes it nearly impossible to use such a collection of full-length ESTs for functional analysis. A further difficulty is that, while the extant cDNAs are all publicly available individually from the usual IMAGE consortium distributors, it would be prohibitively expensive to obtain the source plates for rearraying them into such a collection of unique genes.

In undertaking this project, we sought to create a set of libraries that would be suitable for multiple purposes including EST sequencing, full-length cDNA sequencing, large scale functional analysis (e.g. expression cloning, *in situ* hybridization) and the creation of a physical "UniGene" resource that will ultimately facilitate microarray and functional analyses. As a result, we designed a vector pCS22+ (described below) that has been optimized to maximize the utility of the libraries for multiple purposes. As EST sequences become available, the unique cDNAs that encode full-length sequences will be periodically rearrayed into a unique set that will be made available to the research community without restriction on a cost recovery basis (currently $8.00/plate). Individual cDNAs will be available from the usual IMAGE consortium distributors.

A further goal is to maximize the concordance between the cDNA sequences in the library and the genomic sequence of *X. tropicalis* that will be completed by early 2005. Therefore, we are constructing the libraries in our laboratory from animals closely related to the sixth generation inbred Nigerian strain *X. tropicalis* that is being sequenced at the U.S. Department of Energy Joint Genome Institute. This will minimize the differences between the genomic and cDNA sequences that would otherwise result from the use of outbred strains for large-scale sequencing.

OUR CURRENT APPROACH TO MAKING cDNA LIBRARIES

cDNA libraries can be constructed in different vectors, depending on the ultimate intended use. As a rule of thumb, it is much less tedious to screen lambda phage based cDNA libraries with nucleic acid probes, whereas plasmid libraries are less cumbersome to use for many functional analyses, such as expression cloning. The wide availability of laboratory automation (which has facilitated recent advances in genomics) has also been a great advantage in the production of cDNA libraries. An individual laboratory can now purchase an automated colony picker/gridder with which to transfer bacteria harboring cDNA colonies to microtiter plates, then grid these colonies onto nylon membranes at high density for subsequent screening. The availability of cDNA libraries on high-density filters greatly facilitates library screening, making it a single-step process.

Our cDNA synthesis methods have been described in detail elsewhere [4]. Therefore, this article will focus on new developments and optimizations that have arisen as we undertake the effort to construct a set of full-length *X. tropicalis* cDNAs. The libraries we are constructing will cover a range of developmental stages and adult tissues. Since our libraries will serve as a resource for a variety of genomics and functional approaches to development, we have chosen to make them in a plasmid vector.

1. Preparation of Embryos and mRNA

X. tropicalis embryos are prepared by standard methods [5-7]. Briefly, synchronized embryos are obtained by *in vitro* fertilization of eggs and allowed to develop until the appropriate stage. Embryos or adult tissues are either processed directly for RNA preparation, flash frozen in liquid nitrogen and stored at –70 ˚C, or stored in RNAlater (Ambion) at 4 ˚C.

Our experience with *X. laevis* has shown that extraction of embryos with LiCl/urea/SDS produces maximum quantities of biologically active RNA [4, 8]. This method recovers very close to the theoretical maximum of 4 ug/ embryo for *X. laevis*. Guanidine thiocyanate based methods are utilized only in the case of tissues containing high endogenous RNase levels [9]. Oligotex beads (oligo dT-coated latex beads, Qiagen) are preferred for the purification of poly A^+ RNA. Unlike conventional oligo dT cellulose beads that have dT tracts of 12-18 nucleotides, OligoTex has 30 dT nucleotides coupled to a nonporous latex support by an additional 10 nucleotide dC spacer. The greater length and accessibility of the dT tract increases the affinity of the matrix for bona fide poly A^+ tails, while reducing the interactions between the matrix and A-rich runs (which are typically less than A_{30}) in other cellular RNAs. This increases the purity of the mRNA fraction in the eluted RNA to approximately 90% rather than 50% as is common for oligo dT celluloses, making a single chromatographic step sufficient. An additional benefit of Oligotex is that the latex matrix is nonporous, which enables the mRNA to be recovered in a minimum volume.

2. cDNA Synthesis

First strand cDNA synthesis is accomplished using reverse transcriptase (RT). There are two major types of reverse transcriptases currently employed for cDNA construction. The most widely used are enzymes derived from the Moloney murine leukemia virus (MMLV). Several variants have been developed that eliminate endogenous RNAse H activity that is inherent to the enzyme as well as increasing the thermal stability of the enzyme. This is believed to increase the yields of full-length transcripts [10, 11]. The performance of these enzymes varies significantly among companies; hence, it is worthwhile to test a few different products to determine which works best in your application. An underappreciated concern is the rather egregious licensing agreement that one implicitly agrees to by using the Superscript enzymes produced by InVitrogen. This agreement effectively precludes the distribution of cDNAs produced using Superscript without the permission of InVitrogen, which is problematic for EST distribution and the type of libraries we are producing. As a result, we advocate the use of MMLV-RT enzymes (e.g. Stratascript, Stratagene) that do not have such licensing agreements. Another reasonable choice is the enzyme purified from the avian myeloblastosis virus (AMV). We have found that the AMV-RT produced by Seikagaku America is consistently the best of the commercially available AMV-RT enzymes.

1-2 μg of poly A+ RNA are sufficient for the construction of very complex lambda phage libraries (10^7-10^8 primary clones); however, 5 μg or more are required for the construction of comparable plasmid libraries in our experience. First strand synthesis is primed with an anchored primer of the sequence ($V(T)_{20}$CTCGAGAGAGAGAGAG) that contains an *XhoI* site (underlined) for cloning and is intended to reduce the overall length of the poly A tail in the resulting cDNA clones. This is important for EST sequencing since 3' cluster analysis of ESTs is a valuable classification tool and tails longer than about 30 nucleotides impair sequence analysis.

The second stand of the cDNA is made using the now classical random priming of the first strand cDNA by RNA primers produced by limited digestion of the mRNA template by *E. coli* RNAse H followed by transcription with *E. coli* DNA polymerase I and the repair of remaining nicks with *E. coli* DNA ligase [12]. The use of RNAse H in second strand synthesis largely negates the logic for its elimination in first strand synthesis.

3. Addition of Linker or Adaptors to Facilitate Cloning of Double-Stranded cDNA

After second strand synthesis, the cDNA is rendered blunt-ended by a brief treatment with T4 DNA polymerase. Although commercial kits commonly use asymmetric double stranded adaptors to create restriction sites for cloning, we have found that ligating linkers is substantially more efficient. Specific methylases are often used to protect internal restriction sites from restriction endonuclease digestion. However, methylases are often contaminated with nucleases that reduce the overall efficiency of the cDNA synthesis process. It has been reported that a variety of restriction enzymes are unable to digest hemi-methylated DNA, thus we incorporate 5-methyl dCTP into first strand cDNA to protect endogenous *PstI* and *XhoI* sites that will be used to clone the cDNA. After linker ligation and digestion, the cDNA is size separated by gel filtration chromatography on Sepharose CL-4B. cDNAs larger than 1 kb are pooled and used for library construction. This step also allows further size fractionation, should that be desirable. It is essential to accurately quantitate the relatively dilute cDNA to maximize the fraction of clones harboring inserts during the ligation step. As a result, we advocate fluorometric (e.g. DynaQuant 200, Amersham) determination of DNA concentration.

4. cDNA Cloning and *E. coli* Transformation

Our cDNA synthesis strategy produces cDNAs with a *PstI* site at the 5' end and *XhoI* at the 3' end. Since the libraries are intended for EST sequencing and functional analysis, a plasmid vector is preferred, despite the slightly higher efficiency of cloning into lambda phages, such as lambda ZAP. The cDNAs are ligated into *PstI-XhoI* digested, plasmid at a molar ratio of 1:1 then electrotransformed into DH10B cells, which appear to be universally preferred by the EST projects. Insert frequencies of greater than 90% are routinely achievable, with careful quantitation of vector and cDNA amounts.

5. Modifications Favoring the Synthesis of Full-Length cDNAs

The question of how best to make full-length cDNAs, i.e. faithful copies of the mRNA up to the cap site, has been a persistent problem over the years. It is self-evident that the first requirement is to begin with intact, capped mRNA. Next, one must convert this to a full-length cDNA copy. There has not been very much progress in this area over the years, other than the development of the RNAse H minus reverse transcriptases, which contribute relatively little to first strand synthesis in our experience. It was reported that inclusion of 2.0 M betaine and 0.6 M trehalose in the first

strand synthesis results in longer cDNA synthesis products [13]. We tested different combinations of reverse transcriptase together with betaine and/or trehalose, while optimizing the incubation temperature. A typical result is shown in Fig. (**1**). These experiments showed that trehalose addition is optimal for AMV-RT mediated first strand synthesis, while betaine is best for MMLV-RT. The combination was detrimental for both enzymes as was the use of betaine for AMV-RT or trehalose for MMLV-RT (Fig. **1**). When considering the overall yield of full-length cDNAs alone, the use of AMV-RT plus 0.6 M trehalose at 45 ˚C is clearly superior to the use of MMTV-RT (in this case Superscript II) plus 2.0 M betaine, although the yield of large cDNAs in the latter case is reasonable (Fig. **1**). It is also notable that the inclusion of 0.6 M trehalose to the "optimized" buffer we previously used [4] or 2.0 M betaine to the buffer supplied with the MMLV enzyme significantly improves the yield of the largest cDNAs (Fig. **1**). Although the first strand synthesis with AMV-RT/trehalose was superior, for unexplained reasons the length of the average double-stranded cDNAs is longer when using MMLV-RT (data not shown).

Despite the best efforts and optimization of cDNA synthesis, cDNA libraries constructed by conventional methods typically contain an excess of partial cDNAs. For example, an analysis of the libraries supplied for *Xenopus* EST sequencing from a variety of laboratories showed that the majority contained only 30-60% of demonstrably full-length sequences [14]. Conventionally, first strand cDNA is primed with oligo dT, which largely ensures that the 3' end of the cDNA represents the poly A tail of the mRNA. However, there are no specific sequences at the 5' end of mRNA that can be used to direct synthesis of the 2nd strand cDNA from the authentic transcriptional initiation site. To combat this difficulty, several methods have been developed to introduce specific sequences to the 5' end of mRNAs that contain a ^{5me}GpppG cap structure [15-20]. The two methods in widest use are the cap trapping method of Carninci and colleagues [15] and 5' oligo capping, developed by Sugano and colleagues [19]. Both methods show excellent performance in that they produce cDNA libraries with approximately 80% full-length clones, including the transcriptional start sites [18, 21-25].

It is debatable whether the increase from the routinely achievable 50% full-length clones to the achievable but difficult 80% full-length clones is worth the extra effort it entails. To get a feeling for the cost vs. actual benefit of full-length cloning, we are currently experimenting with 5' oligo capping. Briefly, the poly A^+ RNA is treated with bacterial alkaline phosphatase (BAP) to remove the 5' phosphate from any non-capped (i.e. non-full-length) RNAs. Treatment with tobacco acid pyrophosphatase (TAP) removes the cap structure, leaving a 5' PO_4 only on previously capped mRNAs (Fig. **2**) [18, 19]. These are substrates for the addition of an unphosphorylated oligoribonucleotide with T_4 RNA ligase (Fig. **2**). Thus, only previously capped mRNAs will contain this new 5' sequence. A corresponding oligodeoxynucleotide is then used to prime 2nd strand synthesis, in principle ensuring that the entire 2nd strand is synthesized. The original method uses PCR between this 5' primer and a dT-containing primer to select for full-length cDNAs [18, 19]. However, we are concerned about the potential bias towards short molecules introduced by the PCR step. Therefore, our initial efforts will focus on cloning the cDNAs without PCR amplification using the *PstI* site present in our 5' adapter and the *XhoI* site in the oligo dT primer (Fig. **2**).

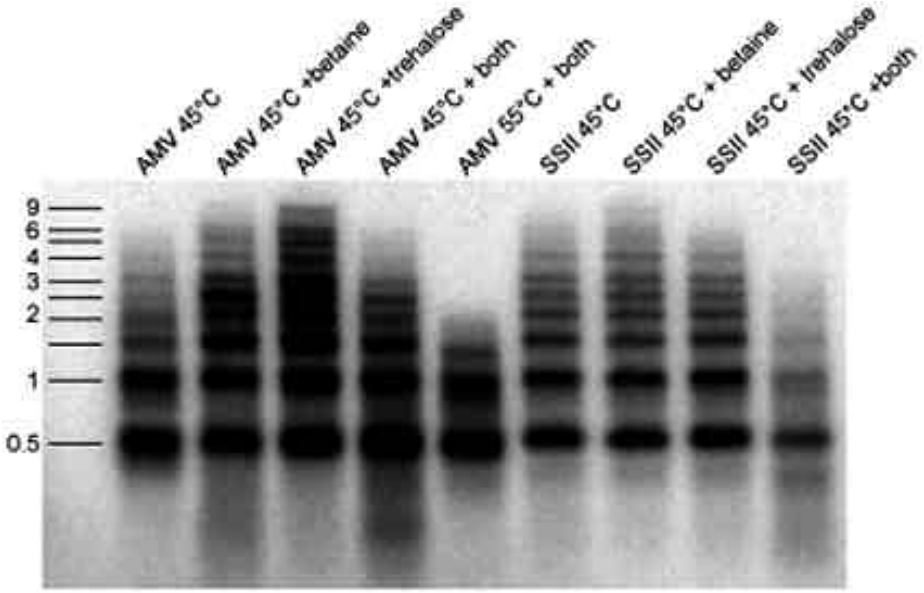

Fig. (1). Optimization of first strand cDNA synthesis.

A comparison between the optimized buffers for AMV reverse transcriptase and MMLV reverse transcriptase evaluating the effects of adding 0.6 M trehalose, 2.0 M betaine or both to the reaction.

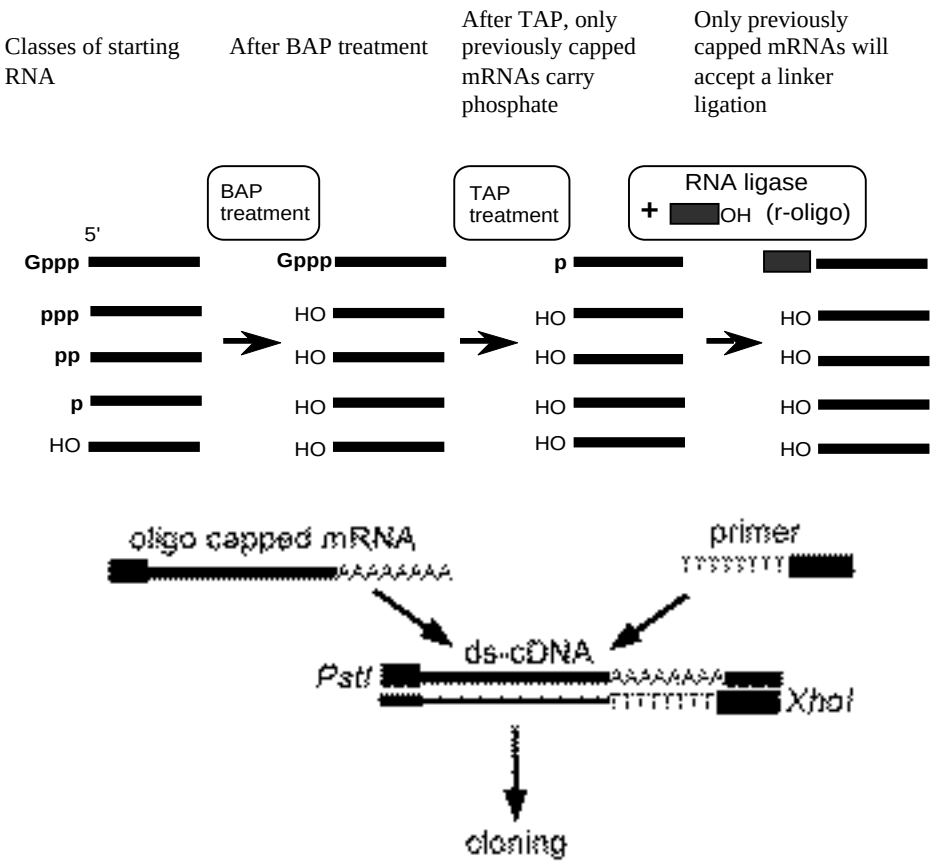

Fig. (2). The oligo capping method for 5' sequence enrichment.

THE EFFECTS OF NORMALIZATION ON LIBRARY LENGTH AND COMPLEXITY

Analyses of classical reassociation kinetics have shown that the mRNAs of a typical somatic cell are distributed in three different frequency classes: (1) abundant (consisting of about 10-15 mRNAs that altogether represent 10-20% of the total mRNA mass; (2) intermediate (1,000-2,000 mRNAs) and (3) rare (15,000-20,000 mRNAs; 40-45%) [26, 27]. Therefore, random sequencing of cDNAs will result in the overrepresentation of sequences from cDNAs in the abundant and intermediate classes. This sequence redundancy prevents the random approach from being cost effective.

The problem of sequence redundancy was addressed by Soares and colleagues [28-30]. They made a number of "normalized" libraries in which the abundance of cDNAs was equalized by a hybridization step between the cDNA library and a large excess of driver cDNA prepared from the same library. A short reassociation time allows only the most abundant sequences to hybridize. Hence, the double stranded cDNAs in the hybridization mix will be predominantly from mRNAs in the abundant and intermediate classes. The remaining single-stranded cDNAs are thereby, enriched in the lower abundance sequences. In principle, a normalized library should produce a larger fraction of unique or low-abundance sequences than a non-normalized library and this has proven to be the case.

What is not intuitive is that the process of normalization seriously biases the library against long cDNAs. We normalized an aliquot of our *X. laevis* gastrula library [31] and picked colonies from both normalized and non-normalized versions for EST sequencing. The results showed that 66.7% of the ESTs from the normalized library were unique (66.7% diverse), whereas only 50.4% of ESTs from the non-normalized library were unique, which is the expected result from normalization [14]. However, only 22.2% of the normalized cDNAs sequenced were full-length whereas 35.6% of the cDNAs in the original library were full-length [14]. Therefore, the increase in diversity comes at the price of a sharp decrease in the number of full-length cDNAs in the library. This decrease in the number of full-length clones is problematic for the resource we wish to create hence another approach is needed.

Our efforts will employ the approach of iterative normalization, which we developed for the *Hydra vulgaris* EST project. In this method, a group of cDNAs are sequenced, typically 5-10,000. These are clustered and sequences representing the most abundant cDNAs (typically 96 or a multiple thereof) are chosen for primer design. These primers are radiolabeled and hybridized to high-density filters containing the next 27,648 candidate cDNAs (the capacity of one large filter, spotted in duplicate) to be sequenced. Colonies that hybridize with the oligos are excluded from further consideration. The ESTs to be sequenced in the next round will come from the non-

hybridizing cDNAs and the process repeated until sufficient sequences are obtained. This approach enables the rapid identification of cDNAs corresponding to lower abundance mRNAs without the introduction of any size or composition bias. In turn, subsequent EST sequencing will result in the identification of a higher proportion of novel sequences than would be possible by sequencing the parent library directly in the absence of a normalization step.

NEW EXPRESSION VECTOR-pCS22+

A number of different plasmid vectors have been used for cDNA library construction. All have strengths and weaknesses for particular approaches. Since our resource is intended to serve a number of functions, a new vector was needed that could facilitate all of the types of applications that would be desired. We considered the following characteristics to be essential for the vector used for library construction:

1. It should contain both a strong eukaryotic promoter for direct expression in cultured cells and microinjected embryos and bacteriophage RNA polymerase promoters for *in vitro* and *in vivo* expression of sense and antisense RNAs.
2. It should have a short polylinker with a minimum of sequence between the bacteriophage promoters and the cloning sites for maximal translation in microinjected oocytes and embryos.
3. It should contain the restriction sites required for our cloning methodology.
4. Restriction endonuclease sites for rare-cutting enzymes should be both upstream and downstream of the cloning sites to facilitate the production of both sense and antisense RNAs for functional analysis.

The plasmid CS2 has been used extensively for experiments in *Xenopus* and zebrafish. Microinjection of pCS2+ -based constructs into *Xenopus* embryos has been demonstrated to result in a lower degree of mosaicism than if other plasmids are used [32], hence we chose pCS2+ as a starting material. pCS2+ is a phagemid derived from pBluescript that contains the simian cytomegalovirus IE94 promoter followed by a convenient polylinker and the SV40 late polyadenylation signal to ensure that the *in vivo* transcripts are polyadenylated (Fig. **3**) [33, 34]. The 5' UTR contains a bacteriophage SP6 promoter to drive sense mRNA synthesis and a T7 promoter (in reverse orientation) on the flanking side of the polylinker to generate antisense RNAs (Fig. **3**). Following the SV40 polyadenylation site is a second polylinker to allow linearization of the template for *in vitro* RNA production followed by the T3 promoter (in reverse orientation).

We modified pCS2+ in the following ways to make it more useful for our purposes. First, the T3 promoter was removed while retaining the 2nd polylinker. The entire SP6-T7 cassette was removed and replaced with a T7-T3 cassette with T7 driving sense RNA production. The logic is that the T7 and T3 polymerases give substantially higher RNA yields (~2x in our hands) than SP6 RNA polymerase and are both more stable to storage than SP6 polymerase. The T7 and T3 promoter sequences were standardized to those in

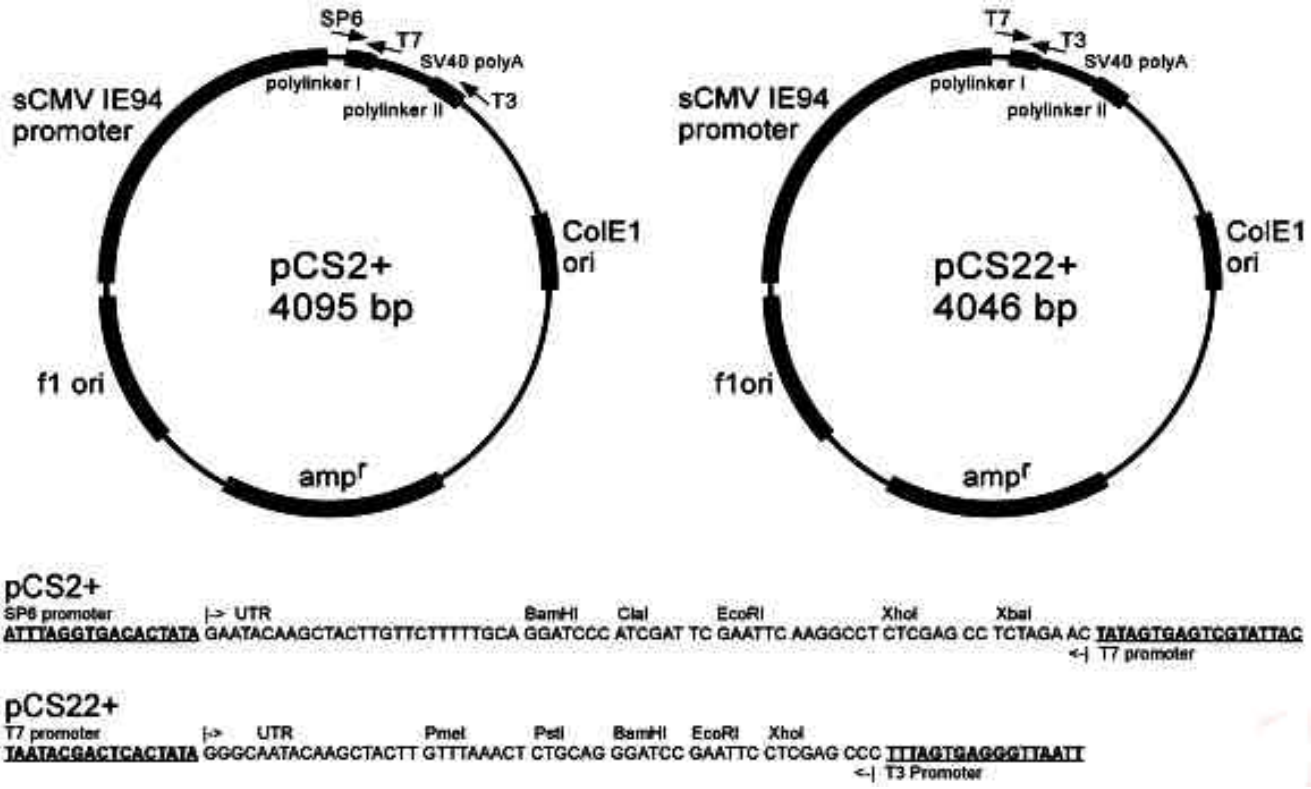

Fig. (3). Comparison of the new plasmid vector pCS22+ with pCS2+.

pBluescript. We retained much of the 5' UTR sequence that was originally between the SP6 promoter and the polylinker in pCS2, since anecdotal evidence suggested that this sequence was very favorable for translation in embryos. We introduced a restriction site for the eight-cutter PmeI (GTTTAAAC) between the T7 promoter and polylinker to allow linearization for the production of antisense RNAs. This enzyme is estimated to cut once per 100K bases in the *Mus musculus* genome and once per 70K bases in the human genome. No sites for PmeI were found in the *X. laevis* or *tropicalis* EST databases (459,281 sequences as of May 30, 2003) suggesting that the vast majority of cDNAs will not be internally digested when linearizing with PmeI. The polylinker is minimal and contains sites for *PstI, BamHI, EcoRI, and XhoI*. The resulting vector was designated pCS22+ (Fig. **3**). We have tested pCS22+ in transfection experiments and mRNA generated from it in microinjection experiments in comparison with pCS2-nuclear-β-galactosidase. In both cases, we found that the β-galactosidase activity was comparable. Thus, pCS22+ is suitable for both library construction and functional studies in *Xenopus* as summarized below.

EST Sequencing - The cDNA libraries are constructed by unidirectional cloning with *PstI* at the 5' end and *XhoI* at the 3' end. Therefore, 5' sequences are readily generated by T7 priming and 3' sequences by T3 priming.

RNA Probes for Northern Hybridization and *In Situ* Hybridization - Antisense RNA probes can be easily generated by using T3 polymerase after linearization with *PmeI* (preferred) or *HindIII*.

In Vitro Transcription – Sense RNAs can be generated by transcription with T7 polymerase after linearization with *NotI* (preferred), *AvaIII*, *ApaI*, or *Asp718*.

In Vitro Expression - The full-length cDNAs in pCS22+ can be conveniently used to produce proteins *in vitro* or *in vivo*. This facilitates approaches such as *in vitro* expression cloning (IVEC) [35, 36] and proteomics where proteins themselves are arrayed and tested for function.

In Vivo Expression - Since pCS22+ has the simian CMV-IE94 promoter 5'to the cDNA insert, the plasmid DNA can direct expression of the cloned cDNA in cultured frog and mammalian cells or in microinjected oocytes or embryos.

Normalization – In the event that normalization of the libraries is ever desired, pCS22+ can be used to generate both single stranded phagemids and driver (via PCR with T7 and T3 primers) to serve as the starting materials for normalization.

THE MOLECULAR INTERACTION SCREEN – AN EXAMPLE FUNCTIONAL GENOMICS APPLICATION FOR THE FULL-LENGTH cDNA COLLECTION

A fundamental challenge facing biologists in the post-genomics era is to determine the functions of genes identified by the genome projects and to understand how gene products interact to mediate the myriad of functions required during embryonic development and the maintenance of the adult organism. Systematic and broadly applicable approaches based on the biochemical properties or biological activities of the gene products are needed for functional characterization of the entire genome. Microarray technologies provide important information about cellular response programs but say little about gene function. Numerous methods are available to detect interactions between proteins, e.g. two-hybrid and phage display screening [37]. Many of these are capable of detecting genome-wide interactions between single proteins. However, none can detect interactions when one of the partners is a macromolecular complex – a particularly troublesome limitation because the majority of cellular proteins function as members of large multi-subunit complexes.

We developed a new technology, the "molecular interaction screen" (MIS), to overcome this limitation and allow one to detect interactions between cDNA-encoded proteins and macromolecular complexes of any size, composition, and complexity [38]. A significant advantage of our approach is that the target can be a protein, protein complex, nucleic acid, protein:nucleic acid complex, small molecule drug, or even a carbohydrate. This flexibility allows the rapid identification of interacting proteins and enables the detection of proteins that may not interact with all members of the complex. The molecular interaction screen is broadly applicable to the study of intra- and extracellular interactions to facilitate the characterization of macromolecular complexes that are responsible for important biological functions.

The MIS is based on *in vitro* expression cloning (IVEC) [35, 36]. The essential difference between IVEC and other methods of expression cloning is the use of small, rather than large, pools of cDNAs. Pools of 96 cDNAs are transcribed and translated *in vitro*, and the resulting proteins used to directly identify the corresponding cDNAs based on the protein activity. Many assays are possible, including microinjection, electrophoretic mobility shift, enzymatic activity, etc [35, 36]. Using small pools normalizes the representation of each protein and minimizes sampling bias. IVEC is a major improvement over previous methods of expression cloning, however, purification of the pools remains laborious and throughput is only low to modest.

Instead of random pools of libraries, our approach begins with collections of known cDNAs such as the *X. tropicalis* gene collection or normalized cDNA libraries. Bacterial colonies containing individual cDNAs are arranged one clone per well in 384-well microplates and, therefore, the identity and location of each cDNA is known from the outset. cDNAs are robotically combined into defined pools, the bacteria grown and plasmid DNA prepared. These DNA pools are used to generate radiolabeled protein pools using coupled *in vitro* transcription and translation. The target protein or complex is purified separately and then bound to scintillant-impregnated microtiter plates (FlashPlates, Perkin-Elmer Life Sciences). Protein pools are added to the target-containing wells in a simple salt and detergent buffer, allowed to interact and then the interaction detected using a microplate scintillation counter. It is also possible to radiolabel the bait protein complex and coat the plate with unlabeled protein pools. Components of positive pools are next tested individually, leading to the identification of an interacting protein in two automated screening steps. This

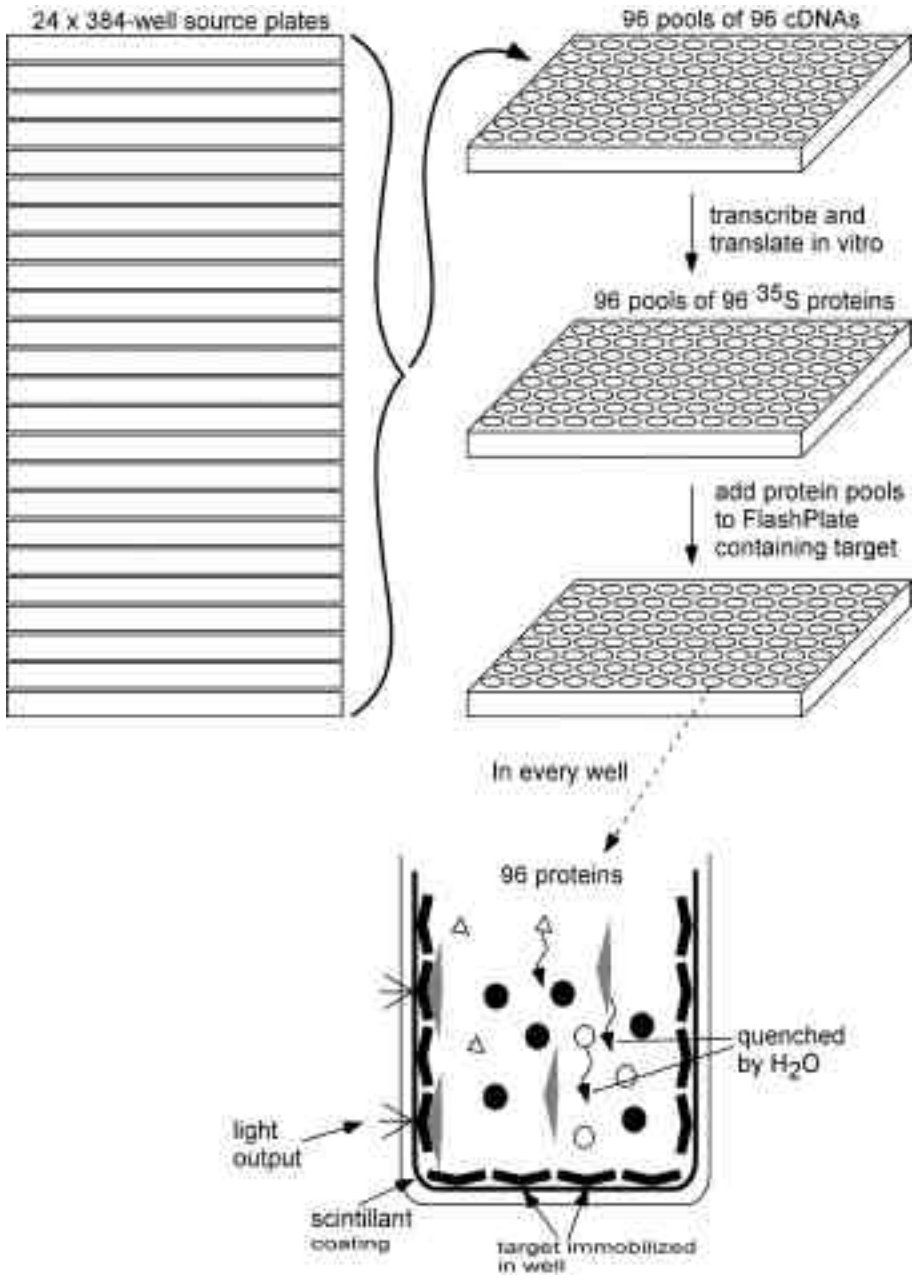

Fig. (4). Schematic illustration of the molecular interaction screen.

The screen comprises 384-well library plates that are pooled into 96 pools of 96 plasmids. These are transcribed and translated *in vitro* into radiolabeled proteins. These protein pools are added in a suitable buffer to FlashPlates coated with the target complex. Proteins that specifically interact with the target remain in the proximity of the scintillant long enough for radioactive emissions to be converted by the scintillant into light output that is detected. Radioactive decay not in the proximity of the target is quenched by the aqueous medium and not detected.

method capitalizes on the short path length of β particles and their quenching by the aqueous solution unless one of the expressed molecules binds to the target immobilized on the plastic. In the absence of specific binding, the time each molecule spends near the plastic is a function of Brownian motion. Specific binding, even if weak, results in a many fold increase in the time that a labeled protein remains in the proximity of the highly sensitive scintillant, which in turn leads to a signal that stands out above background. An elegant feature of this scintillation proximity assay (SPA) [39, 40] is that one does not need to separate bound from free protein or unincorporated label.

The ability to detect interactions with complexes is a decisive advantage of the MIS; moreover, it also has a number of significant benefits over other methods for

studying interactions between two proteins. Yeast two-hybrid screening and phage display utilize fusion proteins, which limits the number of potential interactions to those that occur between individual protein domains [37]. The molecular interaction screen utilizes full-length proteins, allowing simultaneous interactions between multiple protein domains. MIS is superior to other forms of expression cloning because it employs small, rather than large pools of cDNAs for functional screening. Using small pools normalizes the representation of each protein and minimizes sampling bias thereby, increasing sensitivity. The SPA has a very large dynamic range (~10^5 fold) that allows simultaneous detection of weak or strong interactions. Protein chip technology typically utilizes fluorescent labels, limiting the dynamic range to less than 1000 fold and making the detection of weak interactions problematic.

The advantages of the MIS are that the SPA provides quantitative results using readily optimized equilibrium binding. The SPA is rapid, sensitive and homogeneous; therefore, washing steps with the potential to disrupt weak interactions are eliminated. It is also possible to increase the stringency of the assay in a stepwise fashion by adding salt and detergent then rereading the plates, thereby reducing the number of false positives to be characterized. A defined, small number of candidate interactors are being tested, allowing direct comparison between the strength of different interactions. The assay is completely automated, enabling the screening of very large arrays of individual cDNAs. Since only small pools are tested at a time, this strategy is uniquely sensitive to a range of interacting species including weak interactions. All molecules are screened at similar concentrations initially, and ultimately one at a time. This results in the generation of binding curves for each positive, allowing one to classify the interactions hierarchically by binding affinity, which can be used to guide further analysis. The output of the screen is a collection of pure, probably full-length cDNAs that specifically interact with the target of interest.

CONCLUDING REMARKS

The generation of a set of high-quality cDNA libraries will be very beneficial in combination with the mutagenesis, mapping and genome sequencing currently underway for *X. tropicalis*. We have taken pains to ensure that the materials are as closely related as possible to the developing *X. tropicalis* genomic sequence, which should reduce the number of polymorphisms between cDNA and genomic sequences. The sequencing of ESTs and full-length cDNAs is being conducted in collaboration with Naoto Ueno and his colleagues (National Institute of Basic Biology, Okazaki, Japan) and Paul Richardson and his colleagues (U. S. Department of Energy Joint Genome Institute, Walnut Creek, CA, USA). These efforts are underway and we anticipate the first iteration of a partial collection of full-length *X. tropicalis* genes to become available early in 2004. We expect that this set of full-length cDNAs will be widely useful for genomics and functional approaches to *X. tropicalis* Developmental and Cell Biology. These libraries and the derived *X. tropicalis* unique gene set will be made available to the research community without restriction and it is hoped that other laboratories will be built on these tools to further develop *X. tropicalis* as a vertebrate model organism.

ACKNOWLEDGEMENTS

Supported by a grant from the National Center for Research Resources (RR 15088) to BB.

ABBREVIATIONS

EST	=	Expressed sequence tag
DNA	=	Deoxyribonucleic acid
RNA	=	Ribonucleic acid
cDNA	=	Complementary DNA
mRNA	=	Messenger RNA
BAP	=	Bacterial alkaline phosphatase
TAP	=	Tobacco acid pyrophosphatase
PCR	=	Polymerase chain reaction
UTR	=	Untranslated region
IVEC	=	*In vitro* expression cloning
SPA	=	Scintillation proximity assay
MIS	=	Molecular interaction screen

REFERENCES

[1] Brown, P. O. and Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **1999**, *21*: 33-37.

[2] Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O. and Davis, R. W. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* **1997**, *94*: 13057-13062.

[3] Amaya, E., Offield, M. F. and Grainger, R. M. Frog genetics: *Xenopus tropicalis* jumps into the future. *Trends Genet.* **1998**, *14:* 253-255.

[4] Blumberg, B. and Belmonte, J. C. Subtractive hybridization and construction of cDNA libraries. *Methods Mol. Biol.* **1999**, *97*: 555-574.

[5] Wu, M. and Gerhart, J. (1991) Raising *Xenopus* in the Laboratory. in *Xenopus laevis: Practical uses in Cell and Molecular Biology*, eds. Kay, B. K. and Peng, H. B. (Academic Press, San Diego, CA), Vol. 36, pp. 3-18.

[6] Grainger, R. M., Amaya, E., DeSimone, D., Harland, R. M. and Keller, R. X. tropicalis, amphibian model for vertebrate developmental genetics. http://faculty.virginia.edu/xtropicalis/ (**2002**).

[7] Khokha, M. K., Chung, C., Bustamante, E. L., Gaw, L. W., Trott, K. A., Yeh, J., Lim, N., Lin, J. C., Taverner, N., Amaya, E., Papalopulu, N., Smith, J. C., Zorn, A. M., Harland, R. M. and Grammer, T. C. Techniques and probes for the study of *Xenopus tropicalis* development. *Dev. Dyn.* **2002**, *225*: 499-510.

[8] Auffray, C. and Rougeon, F. Purification of mouse immunoglobulin heavy-chain messenger RNAs from total myeloma tumor RNA. *Eur. J. Biochem.* **1980**, *107:* 303-314.

[9] Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. Isolation of biologically active RNA from sources enriched in ribonuclease. *Biochemistry* **1979**, *18*: 5294-5299.

[10] D'Alessio, J. M. and Gerard, G. F. Second-strand cDNA synthesis with *E. coli* DNA polymerase I and RNase H: the fate of information at the mRNA 5' terminus and the effect of *E. coli* DNA ligase. *Nucleic Acids Res*. **1988**, *16*: 1999-2014.

[11] Gerard, G. F., Fox, D. K., Nathan, M. and D'Alessio, J. M. Reverse transcriptase. The use of cloned Moloney murine leukemia virus reverse transcriptase to synthesize DNA from RNA. *Mol. Biotechnol.* **1997**, *8*: 61-77.

[12] Gubler, U. and Hoffman, B. J. A simple and very efficient method for generating cDNA libraries. *Gene* **1983**, *25*: 263-269.

[13] Spiess, A. N. and Ivell, R. A highly efficient method for long-chain cDNA synthesis using trehalose and betaine. *Anal. Biochem*. **2002**, *301*: 168-174.

[14] Klein, S. L., Strausberg, R. L., Wagner, L., Pontius, J., Clifton, S. W. and Richardson, P. Genetic and genomic tools for *Xenopus* research: The NIH *Xenopus* initiative. *Dev. Dyn.* **2002**, *225*: 384-391.

[15] Carninci, P. and Hayashizaki, Y. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **1999**, *303*:19-44.

[16] Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y. and Schneider, C. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **1996**, *37*: 327-336.

[17] Carninci, P., Westover, A., Nishiyama, Y., Ohsumi, T., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., Schneider, C. and Hayashizaki, Y. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res.* **1997**, *4*: 61-66.

[18] Maruyama, K. and Sugano, S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **1994**, *138*: 171-174.

[19] Suzuki, Y. and Sugano, S. Construction of full-length-enriched cDNA libraries. The oligo-capping method. *Methods Mol. Biol.* **2001**, *175*: 143-153.

[20] Suzuki, Y. and Sugano, S. Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.* **2003**, *221*: 73-91.

[21] Suzuki, Y., Yamashita, R., Nakai, K. and Sugano, S. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.* **2002**, *30*: 328-331.

[22] Watanabe, J., Sasaki, M., Suzuki, Y. and Sugano, S. Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. *Gene* **2002**, **291**, 105-113.

[23] Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schonbach, C., Gojobori, T., Baldarelli, R., Hill, D. P., Bult, C., Hume, D. A., Quackenbush, J., Schriml, L. M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K. W., Blake, J. A., Bradt, D., Brusic, V., Chothia, C., Corbani, L. E., Cousins, S., Dalla, E., Dragani, T. A., Fletcher, C. F., Forrest, A., Frazer, K. S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I. J., Jarvis, E. D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R. M., King, B. L., Konagaya, A., Kurochkin, I. V., Lee, Y., Lenhard, B., Lyons, P. A., Maglott, D. R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W. J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J. U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J. C., Reed, D. J., Reid, J., Ring, B. Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C. A., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M. S., Teasdale, R. D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L. G., Wynshaw-Boris, A., Yanagisawa, M., et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **2002**, *420*: 563-573.

[24] Nishikawa, T., Ota, T., Kawai, Y., Ishii, S., Saito, K., Yamamoto, J. I., Wakamatsu, A., Ozawa, M., Suzuki, Y., Sugano, S. and Isocal, T. Comparison of sequences of cDNA clones obtained from oligo-capping cDNA libraries with those from unigene. *DNA Res.* **2001**, *8*: 255-262.

[25] Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., Adachi, J., Fukuda, S., Aizawa, K., Izawa, M., Nishi, K., Kiyosawa, H., Kondo, S., Yamanaka, I., Saito, T., Okazaki, Y., Gojobori, T., Bono, H., Kasukawa, T., Saito, R., Kadota, K., Matsuda, H., Ashburner, M., Batalov, S., Casavant, T., Fleischmann, W., Gaasterland, T., Gissi, C., King, B., Kochiwa, H., Kuehl, P., Lewis, S., Matsuo, Y., Nikaido, I., Pesole, G., Quackenbush, J., Schriml, L. M., Staubli, F., Suzuki, R., Tomita, M., Wagner, L., Washio, T., Sakai, K., Okido, T., Furuno, M., Aono, H., Baldarelli, R., Barsh, G., Blake, J., Boffelli, D., Bojunga, N., Carninci, P., de Bonaldo, M. F., Brownstein, M. J., Bult, C., Fletcher, C., Fujita, M., Gariboldi, M., Gustincich, S., Hill, D., Hofmann, M., Hume, D. A., Kamiya, M., Lee, N. H., Lyons, P., Marchionni, L., Mashima, J., Mazzarelli, J., Mombaerts, P., Nordone, P., Ring, B., Ringwald, M., Rodriguez, I., Sakamoto, N., Sasaki, H., Sato, K., Schonbach, C., Seya, T., Shibata, Y., Storch, K. F., Suzuki, H., Toyo-oka, K., Wang, K. H., Weitz, C., Whittaker, C., Wilming, L., Wynshaw-Boris, A., Yoshida, K., Hasegawa, Y., Kawaji, H., Kohtsuki, S. and Hayashizaki, Y. Functional annotation of a full-length mouse cDNA collection. *Nature* **2001**, *409*: 685-690.

[26] Davidson, E. H. and Britten, R. J. Regulation of gene expression: possible role of repetitive sequences. *Science* **1979**, *204*: 1052-1059.

[27] Bishop, J. O., Morton, J. G., Rosbash, M. and Richardson, M. Three abundance classes in HeLa cell messenger RNA. *Nature* **1974**, *250*: 199-204.

[28] Soares, M. B. and Bonaldo, M. F. (**1998**) Constructing and Screening Normalized cDNA libraries. in *Detecting Genes* (Cold Spring Harbor Press, Cold Spring Harbor), Vol. 2, pp. 49-157.

[29] Soares, M. B., Bonaldo, M. F., Jelene, P., Su, L., Lawton, L. and Efstratiadis, A. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci. USA* **1994**, *91*: 9228-9232.

[30] Bonaldo, M. F., Lennon, G. and Soares, M. B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **1996**, *6*: 791-806.

[31] Cho, K. W. Y., Blumberg, B., Steinbeisser, H. and De Robertis, E. M. Molecular nature of Spemann's organizer: the role of the *Xenopus* homeobox gene goosecoid. *Cell* **1991**, *67*: 1111-1120.

[32] Kroll, K. L. and Amaya, E. Transgenic *Xenopus* embryos from sperm nuclear transplantations reveal FGF signaling requirements during gastrulation. *Development* **1996**, *122*: 3173-3183.

[33] Turner, D. L. and Weintraub, H. Expression of achaete-scute homolog 3 in *Xenopus* embryos converts ectodermal cells to a neural fate. *Genes Dev.* **1994**, *8*:1434-1447.

[34] Rupp, R. A., Snider, L. and Weintraub, H. *Xenopus* embryos regulate the nuclear localization of XMyoD. *Genes Dev*. **1994**, *8*: 1311-1323.

[35] Lustig, K. D., Stukenberg, P. T., McGarry, T. J., King, R. W., Cryns, V. L., Mead, P. E., Zon, L. I., Yuan, J. and Kirschner, M. W. Small pool expression screening: Identification of genes involved in cell cycle control, apoptosis, and early development. *Methods Enzymol.* **1997**, *283*: 83-99.

[36] King, R. W., Lustig, K. D., Stukenberg, P. T., McGarry, T. J. and Kirschner, M. W. Expression cloning in the test tube. *Science* **1997**, *277*: 973-974.

[37] Allen, J. B., Walberg, M. W., Edwards, M. C. and Elledge, S. J. Finding prospective partners in the library: the two-hybrid system and phage display find a match. *Trends Biochem. Sci.* **1995**, *20*: 511-516.

[38] Blumberg, B. High throughput functional screening of cDNAs. in US Patent *6,274,321* (Regents of the University of California, USA) (**2001**).

[39] Undenfriend, S., Gerber, L. and Nelson, N. Scintillation proximity assay: A sensitive and continuous isotopic method for monitoring ligand/receptor and antigen/antibody interactions. *Anal. Biochem.* **1987**, *161*: 494-500.

[40] Nelson, N. A novel method for the detection of receptors and membrane proteins by scintillation proximity radioassay. *Anal. Biochem.* **1987**, *165*: 287-293.

Dear Dr. Blumberg:
Kindly check the word RNAse on page 3 column 2 line No. 4.
Please verify whether it will be RNAase or RNAs.
Thanks in advance
With regards
Liquat
Proofs Reading Department