

Utility of Natural Populations for Microarray Analyses: Isolation of Genes Necessary for Functional Genomic Studies

Marjorie F. Oleksiak, Kevin J. Kolell, and Douglas L. Crawford*

Division of Molecular Biology & Biochemistry, University of Missouri-Kansas City, Kansas City, MO 64110, USA

Abstract: How much variation is there in gene expression? How is this variation partitioned within and among populations? How much variation is biologically important? That is, how much of this variation affects longevity, reproductive fitness, or probability of survival? Microarray analyses can be used to accurately quantify the expression of most, if not all, genes expressed in a tissue and thus address the first question. The latter questions can be investigated by examining the patterns of variation within and among natural populations of *Fundulus*. These populations are large and affected by historical, demographic, and selective constraints, providing a framework for the partition of variation in gene expression within and among populations. Additionally, the well established, phylogenetic relationship among *Fundulus* species can be used to discern adaptive change. A phylogenetic perspective reveals changes that are produced by natural selection and therefore indicates whether this variation affects longevity, reproductive fitness, or probability of survival, i.e., whether the variation is biologically important. However, a *Fundulus* microarray requires DNAs encoding specific *Fundulus* genes. This paper provides information on the production, isolation, and characterization of 4440 *Fundulus* cDNAs used in microarrays. Our approach was to pick random colonies from a normalized cDNA library and then PCR amplify and sequence these genes in a 96-well format. Periodically, the isolated and sequenced cDNAs were subtracted from the normalized library. Normalization reduced the number of redundant genes from 33% to 11%, increasing the effectiveness of this screening process. From 4440 sequenced cDNAs, 49% (2173) had a match in GenBank using BlastX searches. Of these, 53% were nonredundant, yielding 1149 identified genes. These data suggest that cDNAs necessary for microarray analyses can be produced effectively from most organisms.

Key words: *Fundulus*, teleost fish, functional genomics, protocols.

INTRODUCTION

There is a significant amount of genetic variation within and among populations. It is estimated that greater than

three million nucleotide differences occur between any two humans (Li and Sadler, 1991). Some of this sequence variation will be in promoter regions that could affect gene expression. For example, among populations of *Fundulus heteroclitus*, a few nucleotide differences in the heart-type lactate dehydrogenase B (*Ldh-B*) proximal promoter affect

transcription (Crawford et al., 1999b; Segal et al., 1999). A single mutation in the long terminal repeat of the Moloney murine sarcoma virus creates a perfect Sp1 site that allows efficient transcription in embryonic carcinoma cells that does not occur otherwise (Prince and Rigby, 1991). A single nucleotide substitution in the proximal promoter of the low density lipoprotein receptor reduces promoter activity by approximately 50% and contributes to familial hypercholesterolaemia (Koivisto et al., 1994; Sun et al., 1995). A point mutation in the TATA box is thought to be responsible for some cases of β -thalassemia (Antonarakis et al., 1984). These studies indicate that naturally occurring modifications in promoters may be an important mechanism for changing mRNA expression. However, the variation in gene expression and the resulting variation in protein concentration do not necessarily produce a phenotypic change (Crawford et al., 1999a; Flint et al., 1981; Kacser et al., 1973; Middleton and Kacser, 1983; Pierce and Crawford, 1997). Similar to our understanding of protein polymorphisms, there is likely to be considerable variation in mRNA expression, but most of this will be phenotypically unimportant or evolutionarily neutral (Li, 1997). One needs to be able to distinguish between the variations in mRNA expression that produce important phenotypic changes as opposed to those that merely reflect random genetic variation. With so much sequence variation, how will this affect gene expression and what is the expected variation in gene expression among individuals?

It is now possible to address questions concerning genome-wide variation in gene expression by using microarrays. Microarrays are thousands of 150- to 250- μm spots of DNA bound to microscope slides in a precise and known pattern (Ramsay, 1998; Schena et al., 1998). Each DNA spot quantitatively hybridizes to a specific mRNA, so that expression of thousands of individual genes can be measured simultaneously. Importantly, microarray techniques are sensitive: twofold differences in mRNA concentrations are typically determined, and each gene/DNA spot has a sensitivity of 15 attomoles (amol) (Schena et al., 1995). Microarray studies of mRNA expression ignore other processes that could affect protein expression. Variation in protein activity due to allosteric activators, phosphorylation, or other posttranslational modifications is not measured, and thus, its influence on physiological processes is not ascertained. Although enzyme activities or protein concentrations in a cell are not regulated solely by mRNA levels, cell type, developmental morphology, or physiological state often influence mRNA levels. For example, glucose metabo-

lism in humans (Granner and Pilkis, 1990) and yeast (Moore et al., 1991) is increased by changes in the expression of mRNAs for glycolytic enzymes. Among *F. heteroclitus*, the adaptive variation in *Ldh-B* enzyme concentration results from changes in *Ldh-B* mRNA expression (Crawford and Powers, 1989, 1992; Segal and Crawford, 1994). Microarray analyses of yeast mRNA expression in cultures depleted of glucose demonstrate that 1740 genes had at least a twofold change in expression (DeRisi et al., 1996). Many of these changes involve mRNAs that encode catabolic enzymes. Similar changes in mRNA expression also were observed after yeast had evolved on limited amounts of glucose (Ferea et al., 1999). These studies on yeast exposed to limited glucose demonstrate that metabolic adjustments, by evolutionary or physiological mechanisms, are accurately represented by the pattern of mRNA expression (Ferea et al., 1999). Thus, although microarray analyses quantitatively measure only mRNA, the quantification of thousands of genes provides a broad understanding of one of the factors affecting the control of cellular physiological processes.

Microarray analysis can provide data on the patterns of mRNA expression for all genes expressed in a cell. The problem with this vast amount of data is how to determine which changes are significant. This problem is exacerbated because we have little information of the variation of mRNA expression among individuals or populations. For example, the differential expression of *erk-2* in tumor cells treated with estrogen was considered important (Hilsenbeck et al., 1999), yet this observation also is consistent with the hypothesis that differential *erk-2* expression shows greater variation than other genes (Wittes and Friedman, 1999). Without knowledge of the variation in mRNA expression, it is difficult to know if differential expression represents a functionally important change or an underlying difference in variability (Wittes and Friedman, 1999).

One solution is to examine the pattern of variation among natural populations to determine how the variation in expression is partitioned. Is there more variation within or among populations? Additionally, by examining closely related species, it is possible to determine if the patterns of variation are explained most parsimoniously by neutral evolutionary processes or if the patterns more likely are due to evolution by natural selection, [e.g., (Crawford et al., 1999a; Pierce and Crawford, 1997)]. If the variation in expression is neutral, the pattern of expression will be random and thus will be correlated with evolutionary distance among taxa rather than significantly correlated with important environmental parameters (Crawford et al., 1999a;

Pierce and Crawford, 1997). Alternatively, if the pattern of expression is nonrandom and most likely due to evolution by natural selection, it will correlate significantly with important environmental parameters and be independent of evolutionary distance between taxa. Thus, natural populations offer an advantage over most common research organisms because they are subject to natural selection, and biologically important changes in gene expression can be identified using evolutionary analyses. This solution is based on the postulate that variation in the expression of a gene that has evolved by natural selection is biologically important because natural selection can only act on variation that causes phenotypic changes that affect the longevity, reproductive fitness, or probability of survival.

Analyses of the expression of most, if not all, genes in natural populations will require considerable molecular resources. Specifically, one must have the DNA to print on the array. Unfortunately, many believe that microarrays are possible only for model species or, more accurately, species that are well-defined genetically and that have had their genomes sequenced (Brown and Botstein, 1999). Many biologist who do not work on yeast, *C. elegans*, *Drosophila*, mice, or humans believe that this approach is beyond their reach (i.e., their ability to acquire the necessary molecular tools). This report provides evidence that this is not necessarily so. We have isolated and sequenced >4000 cDNAs that can be arrayed to examine patterns of gene expression. Our methods and results are reported here.

MATERIALS AND METHODS

The strategies used to isolate and sequence thousands of *Fundulus* cDNAs are: (1) generate a high-quality unidirectional cDNA library, (2) normalize the library, (3) randomly pick colonies and amplify by PCR the cDNA within the vector, (4) sequence and identify PCR products, and (5) after approximately every 1000 clones, subtract these from the normalized library and repeat steps 3–5. Details for all protocols are provided at <http://sgi.bls.umkc.edu/funnylab/index.html> and have been used in the Comparative Functional Genomic course at Mount Desert Island, 2000.

cDNA Library

To effectively isolate and sequence thousands of cDNAs for the production of microarrays, a unidirectional cDNA library with few nonrecombinants is required. For the *F.*

heteroclitus cardiac library provided by Drs. S. Karchner and M. Hahn, Woods Hole Oceanographic Institute (WHOI) (Karchner et al., 1999), this was accomplished using the UniZap λ cDNA Gigapack Gold cloning kit (Stragene). This library was produced from 27 fish hearts (both sexes) sampled from Scorton Creek in Sandwich, MA, USA. The cDNAs in this library are oriented such that the 5' end is ligated to *EcoRI* and 3' poly A is ligated to *XhoI*. This library has less than 1% nonrecombinants, i.e., 2 of 300 random clones from a nonnormalized library had no inserts.

Normalization

Normalization of cDNA libraries reduces the differences among expressed genes to less than 10-fold among rare and abundant mRNAs (Bonaldo et al., 1996; Hillier et al., 1996). Normalized libraries were produced by isolating cDNAs from approximately 10^{12} plasmids. cDNAs were isolated using PCR amplification with vector specific primers immediately 5' and 3' to the insertion site (*EcoRI* and *XhoI* sites). These PCR products (PCR cDNAs) were denatured and hybridized to single-stranded plasmids from the cardiac cDNA library. Taking advantage of Cot values, the most abundant cDNAs were annealed to the more abundant PCR products and were removed selectively by hydroxyapatite column chromatography. The single-stranded plasmids in the flow-through were converted to double strands using the DNA polymerase Sequenase (Amersham). DH10s *E. coli* (BRL) were transformed with these double-stranded plasmids by electroporation. The number of recovered plasmids and the resulting complexity of the normalized library depended on the duration of hybridization or Cot values. Two normalized libraries were made using 12- and 24-hour hybridizations. The library from the 12-hour hybridization yielded 250,000 plasmids. The 24-hour hybridization yielded 3000 plasmids and had a greater representation of rare mRNAs and greater frequency of nonrecombinants.

Isolation and Sequencing of cDNAs

Characterization of cDNAs (growth of individual bacterial colonies containing plasmids, PCRs, purification of PCR products, sequencing reactions) used 96-well plates and octopipets. This greatly increased productivity and reduced the time required for large-scale characterization of expressed genes. To characterize cDNAs, 96-individual bacterial colonies from the normalized library were randomly chosen, and each was grown in 1.25 ml of Superbroth in 2-ml, 96-well plates. After 18 hours of growth, two 250- μ l

bacterial glycerol stocks were made and stored in 96-well plates at -80°C . One microliter of these bacterial growths was used for PCR reactions using forward and reverse plasmid specific primers: (PucF⁺ = CGC · CAG · GGT · TTT · CCC · AGT · CAC · G, PucR⁺ = GAG · CGG · ATA · ACA · ATT · TCA · CAC · AGG · AAA. PCR reactions had 0.2 mM dNTPs, 10 pmols of each primer, 1 U of Promega Taq (0.2 μl), and reaction buffer with detergents and DMSO [final concentrations: 50 mM Tris HCl, pH 9.2 (25 $^{\circ}\text{C}$), 16 mM (NH₄)₂SO₄, 2.25 mM MgCl₂, 2% (v/v) DMSO, 0.1% (v/v) Tween 20]. Two-step thermal cycle conditions were used (94 $^{\circ}\text{C}$ for 10 seconds; then 32 cycles of 94 $^{\circ}\text{C}$ for 30 seconds followed by 70 $^{\circ}\text{C}$ for 5 minutes; after the 32 cycles, 72 $^{\circ}\text{C}$ for 15 minutes). PCR products were purified in a 96-well format using Sephadex G-50 in a deep well plate with a 0.2- μm filter (Millipore).

PCR products were sequenced from the 5' end (relative to the mRNA) on an ABI 373 sequencer using ABI Big Dye reaction mix. We typically used 1/10 the amount of reaction mix yielding 300–400 unambiguous bases. To increase relative signal to background fluorescence, sequencing reactions were performed using a biotin-tagged primer and purified with streptavidin-coated magnetic beads. To decrease manipulations and minimize the use of pipet tips, magnetic beads were washed and moved to a clean plate with the use of a "bed of nails": 96 nails embedded in a polycarbonate 96-well plate and held in place with epoxy. Streptavidin beads were collected by magnetizing the nails and then rinsing and placing the beads in a new plate by removing the magnetic force. Prior to loading, the sequencing reactions were denatured and the magnetic streptavidin beads removed with the bed of nails.

Characterization of cDNA Sequences

The proteins encoded by the cDNAs were first identified using the putative amino acid sequence and their similarity to proteins (BlastX search) and then compared to nucleotide (BlastN) sequences in the NCBI database. These searches were accomplished by an unsupervised program written in PERL script for the LINUX operating system on a laboratory computer that had the complete GenBank database downloaded weekly. These unsupervised searches produced three outputs: (1) alignments of BlastX, (2) alignments of BlastN, and (3) a summary description that included the sequence, top three BlastX and N identities, and their probability scores. The summary (item 3) output for 96 sequences is approximately 12 pages long and thus allows

information to be easily printed and stored as a hard copy. A cDNA was assumed to code for a homologous protein if the probability score was less than 10^{-6} for BlastX or 10^{-7} for BlastN or if the combined probability of BlastX and BlastN was less than 10^{-8} .

Validation

We use three procedures to verify that the correct sequence is associated with each cDNA: (1) Each 96-well plate has three wells with a marker cDNA (*F. heteroclitus* actin with unique vector sequences). Two wells (40 and 67) always contain the marker cDNA, and thus any misloading or mislabeling of sequencing lanes is identifiable. The third marker cDNA is placed in a well that corresponds to the plate number (e.g., plate 2 has the marker in well 2, Figure 1). (2) After the production of 12 plates, one row (8 wells) from each plate is resequenced. Thus, 8/96 or $\approx 8\%$ of all sequences and their locations are confirmed. (3) These cDNAs will be resequenced when they are arrayed again to produce the template to print microarrays (i.e., 100% of all printed cDNAs will be resequenced). These measures are important to ensure that the correct and known cDNAs are printed.

Subtraction

The complexity of the normalized library was reduced by subtracting the characterized cDNAs previously isolated from the normalized library. Subtraction greatly reduced the probability of isolating the same cDNA and thus improved the efficiency of screening the library for unique clones. Subtraction used a 100-fold molar excess of biotin-labeled, antisense cDNAs produced by PCR using all the characterized cDNAs as substrates and vector-specific primers in which the 3' primer was labeled with biotin. These PCR products were hybridized to the single-stranded normalized library in the presence of oligo-dA and vector-specific oligos (which prevent nonspecific hybridization to oligo-dT or vector sequences). After a 24-hour hybridization, genes in the library bound to these biotin-labeled PCR products were removed with the use of magnetized, streptavidin-coated beads. The single-stranded plasmids in the flow-through were converted to double strands using the DNA polymerase Sequenase (Amersham). DH10s *E. coli* were transformed with these double-stranded plasmids by electroporation.

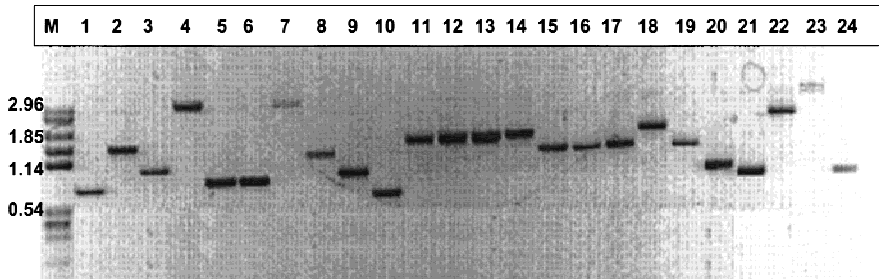


Figure 1. PCR products amplified from cDNA library, plate 2. Products: 1, ATP synthetase alpha chain; 2, actin control; 3, unidentified; 4, ras-related protein; 5, cardiac LIM protein; 6, GAPDH; 7, ?; 8, actin depolymerization factor; 9, α -actin; 10, 5S ribosomal protein; 11, myosin; 12, ?; 13, poly-A binding protein; 14, endopeptidase; 15, peripheral benzodiazepine receptor associ-

ated protein; 16, NADH-ubiquinone oxidase; 17, laminin; 18, NIPSNAP; 19, synaptic protein; 20, α -hemoglobin; 21, cyt *c* oxidase polypep. I; 22, aspartate aminotransferase; 23, unidentified; 24, ETS/Elf/E74-like transcription factor. M = molecular markers of 2.96, 2.51, 1.85, 1.44, 1.14, 0.54, 0.45, 0.41, 0.25, and 0.17 kb.

RESULTS

PCR

Polymerase chain reactions were robust: greater than 90% of the time these conditions produced single PCR products (Figure 1). Importantly, these PCR conditions produced products whose sizes exceeded 5 kb, and the size of many of the PCR products exceeded 3 kb. Large-size inserts did not seem to contribute to null amplification, although larger fragments often had less product. Instead, null amplifications most frequently were associated with lack of insert.

Normalization Effectiveness

The first 300 isolated cDNAs were from a nonnormalized library. Among these sequences, 28% were found only once (nonredundant) among all other sequences (4440). For the first 300 cDNAs from the normalized library, 62% were nonredundant among all other sequences. These percentages were determined by BlastN searches of the *Fundulus* cardiac cDNAs. This strongly suggested that normalization effectively reduced the number of cDNAs that needed to be isolated because many more of the normalized cDNAs were likely to be nonredundant. However, a cDNA may be nonredundant for two reasons: (1) the gene encoded by the cDNA was isolated only once or (2) there was more than one cDNA isolated for a gene, but the sequences for these cDNAs were nonoverlapping (remember, only partial sequences were determined).

The effectiveness of normalization also was evaluated by comparing identified genes. This approach does not de-

pend on acquisition of the same DNA sequence for each gene. Instead, cDNAs that had low BlastX *P* values were identified as the same protein and thus classified as redundant. Limiting the comparison to the first 300 nonnormalized cDNAs revealed that 33% of these cDNAs were redundant. Among the first 500 cDNAs from the normalized library, 11% were redundant. Thus, normalization decreased the probability of sampling the same cDNA by threefold, and was an effective method for increasing productivity.

Genes

In all, 5376 colonies were isolated, grown, and PCR amplified. Some of these reactions were unproductive: rarely, a whole plate would fail, and occasionally a few PCR products from a plate would not be produced. Thus, 4440 of these 5376 isolates (83%) produced sequences. Of these genes, 49% (2173) had a match in GenBank using BlastX searches. Of these 2173 cDNAs, 53% (1149) were nonredundant. From all the cDNAs sequenced, 26% (1149/4440) were identified (with $P < 10^{-6}$) and were nonredundant. For details of these nonredundant genes, please refer to my web site (<http://sgi.bls.umkc.edu/funnylab/index.html>).

Forty-nine percent of the *Fundulus* cDNAs sequenced by single pass for approximately 400 bp were identified by BlastX. This is similar to the yeast genome, where 50% of the genes are identified among all possible open reading frames >100 bp in yeast (<http://www.urmc.rochester.edu/smd/biochem/yeast/5.html>). Our results also are similar to other characterizations of randomly chosen cDNAs. For chicken and mice cDNA characterizations, 71% and 49% of

the cDNAs, respectively, were identified (Li et al., 1998; Sasaki et al., 1998). However, these libraries were not normalized, and the most frequently found genes (e.g., ribosomes, actin) were identified more readily. From the *Fundulus* nonnormalized library, 72% were identified and 28% were nonredundant and identified. This is similar to mice where 3395 cDNAs yielded 937 nonredundant genes (28%) (Sasaki et al., 1998).

DISCUSSION

The biological sciences are entering an exciting new era in which the ability to accurately analyze genome wide patterns of gene expression is possible. Functional genomics seeks to provide functional information to the growing database of DNA sequences. Measuring patterns of mRNA expression is just the first step for functional genomics. Additional research will be required to discern how changes in mRNA expression effect a phenotypic change. If we seek to understand the functional importance of DNA sequences and how patterns of mRNA expression affect the biology of organisms, one of the more productive approaches is to follow August Krogh's principle (Krogh, 1929): for many problems there will be an animal for which it can be most conveniently studied. Functional genomics can be enhanced by using a diversity of organisms in which physiological, developmental, or biochemical traits are readily studied.

The strength of a comparative approach is the utilization of species or groups of species best suited to address specific physiological or biochemical processes. For example, Hans Krebs' research depended on muscle tissue from the common dove to elucidate the TCA cycle (Krebs and Johnson, 1937). Krebs' Nobel Prize (1953) winning research used this nonmodel species because the breast muscle was rich in mitochondria and these organelles were "tough" (Krebs, 1975). Warburg (Nobel Prize, 1931) used a wide range of species to elucidate metabolic principles (Warburg, 1908, 1923). The Nobel Prize for the fundamental work on neural conduction by Hodgkin and Huxley depended on the giant nerve fiber of the squid *Loligo* (Hodgkin, 1963; Huxley, 1963). Basic research on sodium transport was done on toad bladder (Ussing, 1952). The elucidation of the role of acetylcholinesterase in neural impulses used the electric organ of the fish *Electrophorus electricus* (Nachmansohn, 1959). Isolation of influenza virus used the ferret (Smith, 1933). Nuclear and cytoplasmic interactions were explored with the giant unicellular alga *Ac-*

etabularia (Harris, 1968). The studies of prion proteins, for which a Nobel Prize was awarded in 1997, were based on a diversity of mammals (Lee Inyoul et al., 1998; Prusiner, 1998; Scott Michael et al., 1997). The recent Nobel Prize for the biological importance of nitric oxide used endothelium from rabbits. These studies are but a few biological endeavors that relied on nonmodel species to elucidate fundamental principles of human biology.

To apply functional genomics to a diversity of organisms will require considerable molecular resources, including the DNA for printing microarrays. The data presented here suggest that many laboratories can acquire the necessary molecular tools to begin a functional genomics approach. The amount of work and the cost associated with acquiring genes for microarrays is reduced by normalizing the library and subtracting previously isolated cDNAs. Normalization reduces the bias of the library [i.e., most genes have similar concentrations versus a few genes with orders of magnitude more than all others (Bonaldo et al., 1996)]. Subtracting the cDNAs that have been isolated reduces the probability of isolating them again. This serial strategy effectively reduces the number of genes that have to be isolated. The effectiveness of this approach is supported by the data presented here that demonstrates that 53% of cDNAs will be nonredundant (i.e., nonduplicates). From the 4400 cDNAs sequenced, 2400 were nonredundant. The effectiveness of normalization is demonstrated by comparing normalized versus nonnormalized libraries. Among the first 500 sequences, only 11% of the cDNAs from the normalized library were found multiple times. In contrast, in the nonnormalized library, 33% of the cDNA was found more than once within these first 300 sequences. The overall effectiveness of our strategy is demonstrated by inquiring how often the same gene was isolated more than once. Among the 300 sequences from the nonnormalized library, only 28% were nonredundant (i.e., found only once). However, after the library was normalized and subtracted, 62% of the first 300 sequences were found only once. Clearly, normalization and subtraction increase the effectiveness of microarray development by reducing the redundancy of effort.

Of all the cDNAs that produced sequences, half (49%) were identifiable. Although it is easier to understand the utility of identified cDNAs, unidentified genes are of interest because they may have interesting patterns of expression. However, there is need for a word of warning. We subtracted all isolated cDNAs, including unidentified genes. These unidentified genes could include cDNAs that are unknown because they contain exclusively 3' UTRs, which are

highly variable between species. Subtracting these 3' cDNA ends would remove other cDNAs that had more 5' sequences—sequences that could be identified. Thus, subtracting unidentified cDNAs has the benefit of increasing the effectiveness of the screening process when these genes are unknowns but also reduces the chances of finding the 5' end of a cDNA for the same gene which may be identifiable.

Functional Genomics Among Diverse Species

Analysis of most of the genes expressed in a tissue or at different developmental stages provides a foundation to examine which genes are responsible for a diversity of biological problems. Instead of trying to establish whether the expression of one or a few genes correlates with a physiological or developmental process, one can examine all the genes and simply ask which genes have a pattern of gene expression indicative of a functionally important role. This approach is likely to identify novel and important changes and therefore to identify novel physiological mechanisms. For example, a metabolic shift in yeast involves changes in mitochondrial and cytoplasmic genes involved in protein synthesis (DeRisi et al., 1997). Acclimation response in the common carp involves the change in expression of genes involved in clotting (Andrew Cossins, personal communication).

By exploring the variation in mRNA expression among nonstandard organisms, we can gain a better understanding of which genes effect an adaptive change in physiological processes. For example, *Artemia* have virtually no metabolic rate during diapause. Which genes are involved in this process? Is the concentration of one, a few, or all genes down-regulated? Toadfish are an unusual teleost in that they are capable of producing urine. Are specific genes involved in the differential regulation of this metabolic pathway? Are there unique sets of genes expressed in this species versus other similar species not capable of producing urea? Among Antarctic species are there unique patterns of gene expression found in these extreme hypotherms? In estuarine organisms, which genes have differential expression when subjected to different salinities or different environmental pollutants? In ectothermic organisms, where differences in incubation temperatures affect sex ratios, what are the changes in gene expression associated with the establishment of one or the other sex? Among social insects, what patterns of gene expression are associated with different castes: what makes a larger soldier versus a minor worker?

These problems most likely do not represent qualitative differences in gene expression (on/off) but instead represent quantitative differences in gene expression that could be readily determined using microarrays. There is an opportunity to study the functional genomics among a wide diversity of organisms and thus learn more about the biological solutions to adaptation. These solutions are unlikely to be obtained by standard molecular techniques (e.g., knock-outs) on model systems.

Much of developmental biology uses a diversity of organisms (Martindale and Swalla, 1999) to discern the genes involved in pattern formation. Until recently, only the presence and absence of a gene were readily determined (e.g., by subtractive hybridization, differential display, or in situ hybridization). Microarray technologies will allow developmental biologists to study quantitative differences in gene expression. This approach could begin to address the importance of quantitative differences in gene expression in the establishment of, for example, dorsal–ventral axis formation, bilateral symmetry, or limb development. Evolutionary developmental biology would benefit from the ability to use microarrays in a wide diversity of organisms.

Molecular ecology is a relatively young field (the first journal devoted to this subject is less than a decade old). However, it offers insights into how ecological factors affect the biochemistry and molecular biology of an organism and how these molecular attributes affect the interaction among different organisms. Should the most powerful tools for the analysis of gene expression be denied to this and other fields because of the cost and time required to develop them? The specific problem addressed here is how to provide the necessary tools to study this diversity of organisms. This manuscript suggests that there is a simple answer: isolate the cDNAs expressed in these organisms so that they can be used to print microarrays to study patterns of gene expression.

ACKNOWLEDGEMENTS

This research was supported by University of Missouri Research Board Grant, NSF IBN 9986602 to DLC and NSF BioInformatic Post-Doctoral Fellowship NSF DBI 0074520 to MFO. Additional support was provided by the Dean of the School of Biological Sciences, Dr. M. Martinez-Carrion.

REFERENCES

Antonarakis, S.E., Irkin, S.H., Cheng, T.C., Scott, A.F., Sexton, J.P., Trusko, S.P., Charache, S. and Kazazian, H.H., Jr. (1984). Beta-

- thalassemia in American Blacks: novel mutations in the "TATA" box and an acceptor splice site. *Proceedings of the National Academy of Sciences of the United States of America* 81:1154–1158.
- Bonaldo, M.F., Lennon, G., and Soares, M.B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 6:791–806.
- Brown, P.O., and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet* 21:33–37.
- Crawford, D.L., and Powers, D.A. (1989). Molecular basis of evolutionary adaptation at the lactate dehydrogenase-B locus in the fish *Fundulus heteroclitus*. *Proc Natl Acad Sci USA* 86:9365–9369.
- Crawford, D.L., and Powers, D.A. (1992). Evolutionary adaptation to different thermal environments via transcriptional regulation. *Mol Biol Evol* 9:806–813.
- Crawford, D.L., Pierce, V.A., and Segal, J.A. (1999a). Evolutionary physiology of closely related taxa: analyses of enzyme expression. *Am Zool* 32:389–400.
- Crawford, D.L., Segal, J.A., and Barnett, J.L. (1999b). Evolutionary analysis of TATA-less proximal promoter function. *Mol Biol Evol* 16:194–207.
- DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer [see comments]. *Nat Genet* 14:457–460.
- DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–686.
- Felsenstein, J. (1985). Phylogenies and the comparative method. *Am Nat* 125:1–15.
- Ferea, T.L., Botstein, D., Brown, P.O., and Rosenzweig, R.F. (1999). Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci USA* 96:9721–9726.
- Flint, H.J., Tateson, R.W., Barthelmess, I.B., Porteous, D.J., Donachie, W.D., and Kacser, H. (1981). Control of the flux in the arginine pathway of *Neurospora crassa*. Modulations of enzyme activity and concentration. *Biochem J* 200:231–246.
- Garland, T., Jr., and Carter, P.A. (1994). Evolutionary physiology. *Annu Rev Physiol* 56:579–621.
- Granner, D., and Pilkis, S. (1990). The genes of hepatic glucose metabolism. *J Biol Chem* 265:10173–10176.
- Harris, H. (1968). *Nucleus and Cytoplasm*. Oxford: Clarendon Press.
- Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chisoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M., and et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* 6:807–828.
- Hilsenbeck, S.G., Friedrichs, W.E., Schiff, R., O'Connell, P., Hansen, R.K., Osborne, C.K., and Fuqua, S.A.W. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *JNCI* 91:453–459.
- Hodgkin, A.L. (1963). The ionic basis of nerve conduction. *Prix Nobel* 1963:224–241.
- Huxley, A.F. (1963). The quantitative analysis of excitation and conduction in nerve. *Prix Nobel* 1963:242–260.
- Inyoul L.Y., Westaway, D., Smit A.F.A., Wang, K., Seto, J., Chen, L., Acharya, C., Ankener, M., Baskin, D., Cooper, C., Yao, H., Prusiner, S.B., and Hood L.E. (1998). Complete genomic sequence and analysis of the prion protein gene region from three mammalian species. *Genome Res* 8:1022–1037.
- Kacser, H., Bulfield, G., and Wallace, M.E. (1973). Histidinaemic mutant in the mouse. *Nature* 244:77–79.
- Karchner, S.I., Powell, W.H., and Hahn, M.E. (1999). Identification and functional characterization of two highly divergent aryl hydrocarbon receptors (AHR1 and AHR2) in the teleost *Fundulus heteroclitus*. Evidence for a novel subfamily of ligand-binding basic helix loop helix-Per-ARNT-Sim (bHLH-PAS) factors. *J Biol Chem* 274:33814–33824.
- Koivisto, U.M., Palvimo, J.J., Janne, O.A., and Kontula, K. (1994). A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia. *Proc Natl Acad Sci USA* 91:10526–10530.
- Krebs, H.A. (1975). The August Krogh principle: "For many problems there is an animal which it can be most conveniently studied." *J Exp Zool* 194:221–226.
- Krebs, H.A., and Johnson, W.A. (1937). The role of citric acid in intermediate metabolism in animal tissues. *Enzymologia* IV:148–156.
- Krogh, A. (1929). Progress of physiology. *Am J Physiol* 90:243–251.
- Li, S., Liu, N., Zadworny, D., and Kuhnlein, U. (1998). Genetic variability in white leghorns revealed by chicken liver expressed sequence tags. *Poult Sci* 77:134–139.
- Li, W.-H. (1997). *Molecular Evolution*. Sunderland, MA: Sinauer Associate, Inc., 487 pp.
- Li, W.-H., and Sadler, L.A. (1991). Low nucleotide diversity in man. *Genetics* 129:513–523.

- Martindale, M.Q., and Swalla, B.J. (1999). The evolution of developmental patterns and process. *Am Zool* 38:3–12.
- Middleton, R.J., and Kacser, H. (1983). Enzyme variation, metabolic flux and fitness: alcohol dehydrogenase in *Drosophila melanogaster*. *Genetics* 105:633–650.
- Moore, P.A., Sagliocco, F.A., Wood, R.M., and Brown, A.J. (1991). Yeast glycolytic mRNAs are differentially regulated. *Mol Cell Biol* 11:5330–5337.
- Nachmansohn, D. (1959). *Chemical and Molecular Basis of Nerve Activity*. New York: Academic Press.
- Pierce, V.A., and Crawford, D.L. (1997). Phylogenetic analysis of glycolytic enzyme expression. *Science* 275:256–259.
- Podrabsky, J.E., Javillonar, C., Hand, S.C., and Crawford, D.L. (2000). Intraspecific variation in aerobic metabolism and glycolytic enzyme expression in heart ventricles from *Fundulus heteroclitus*. *Am J Physiol* 279:R2344–2348.
- Prince, V.E., and Rigby, P.W. (1991). Derivatives of Moloney murine sarcoma virus capable of being transcribed in embryonal carcinoma stem cells have gained a functional Sp1 binding site. *J Virol* 65:1803–1811.
- Prusiner, S.B. (1998). Prions. *Proc Natl Acad Sci USA* 95:13363–13383.
- Ramsay, G. (1998). DNA chips: state-of-the-art. *Nat Biotechnol* 16:40–44.
- Sasaki, N., Nagaoka, S., Itoh, M., Izawa, M., Konno, H., Carninci, P., Yoshiki, A., Kusakabe, M., Moriuchi, T., Muramatsu, M., Okazaki, Y., and Hayashizaki, Y. (1998). Characterization of gene expression in mouse blastocyst using single-pass sequencing of 3995 clones. (published erratum appears in *Genomics* 54(3):583, 1998). *Genomics* 49:167–179.
- Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–470.
- Schena, M., Heller, R.A., Thériault, T.P., Konrad, K., Lachenmeier, E., and Davis, R.W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol* 16:301–306.
- Scott Michael, R., Safar, J., Telling, G., Nguyen, O., Groth, D., Torchia, M., Koehler, R., Tremblay, P., Walther, D., Cohen Fred, E., Dearmond, S.J., and Prusiner, S.B. (1997). Identification of a prion protein epitope modulating transmission of bovine spongiform encephalopathy prions to transgenic mice. *Proc. Natl. Acad. Sci. USA* 94:14279–14284.
- Segal, J.A., and Crawford, D.L. (1994). LDH-B enzyme expression: the mechanisms of altered gene expression in acclimation and evolutionary adaptation. *Am J Physiol* 267:R1150–1153.
- Segal, J.A., Barnett, J.L., and Crawford, D.L. (1999). Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. *J. Mol. Evol.* 49:736–749.
- Smith, W.C. (1933). Virus obtained from influenza patients. *Lancet* 2:66–68.
- Sun, X.M., Neuwirth, C., Wade, D.P., Knight, B.L., and Soutar, A.K. (1995). A mutation (T-45C) in the promoter region of the low-density-lipoprotein (LDL)-receptor gene is associated with a mild clinical phenotype in a patient with heterozygous familial hypercholesterolaemia (FH). *Hum Mol Genet* 4:2125–2129.
- Ussing, H.H. (1952). Some aspect of the application of tracers in permeability studies. *Adv Enzymol* 13:21–65.
- Warburg, O.H. (1908). Beobachtungen über die Oxydationsprozesse in Seeigelen. *Hoppe-Seyler's Z Physiol Chem* 57:1–16.
- Warburg, O.H. (1923). Versuche an überledendem Carcinomgewebe. *Biochem Z* 142:317–334.
- Wittes, J., and Friedman, H.P. (1999). Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. *JNCI* 91:400–401.