Part II

Recent Advancements in Biotechnology

Chapter 11

The Many Faces of Gene Expression Profiling: Transcriptome Analyses Applied Towards Elucidating Marine Organismal Interactions and Metabolism

JOSE LOPEZ, BENOIT MOUZON, PETER J. McCARTHY and RUSSELL KERR

Learning Objectives

This chapter intends to exhibit and summarize current molecular biotechnologies that focus on characterizing gene expression at the mRNA level. Specific focus will be on recent research regarding marine ecosystems and organisms. Learning objectives for the advanced biology student are the following

- To understand the similarities and differences between genomics (DNA-based) and transcriptomics (RNA-based) laboratory methods for data generation and analysis.
- To learn the strengths and weaknesses of various state-of-the-art transcriptomic methods used to study gene expression and mRNA transcripts, and when they can be most appropriately applied.
- To rate and apply the most cost-effective means for transcriptome analyses based on the model system and hypotheses posed.

1. INTRODUCTION

The prepared student of biology in this new millennium understands that we are now in the era of genomics, which is the study of the complete heredity information or blueprint in the biochemical form of deoxyribonucleic acid (DNA) of any given organism. A wealth of this macromolecular (DNA, RNA and protein) sequence data now floods our databases due to the improving technology of data generation, such as pyrosequencing, single-molecule DNA

[1*]Oceanographic Center, Nova Southeastern University, Dania Beach FL 33004, USA,
[2] Florida Atlantic University at Harbor Branch Oceanographic Institute, Fort Pierce, FL 34946, USA;
[3]Department of Chemistry, and Department of Biomedical Sciences, University of Prince Edward Island, 550 University Drive, Charlottetown, PEI, C1A 4P3, Canada
*Email: jochrome@gmail.com

sequencing, mass spectrophotometric sequencing and other methods on the horizon. To date, based on traditional Sanger sequencing, more than 800 genomes have been fully sequenced. Most of the sequenced genomes are microbial although various eukaryotic genomes include humans, chimpanzee, sea urchin and platypus is also included (http://www.genomesonline.org/gold.cgi). Added to this are the growing number of metagenomes derived from environmental samples, and even with our still crude technology, the growth of DNA sequence data becomes unequivocally exponential (e.g. see oceanic sampling by www.venterinstitute.org). There is a growing demand for faster and less expensive DNA sequencing machines (http://genomics.xprize.org/genomics/archon-x-prize-for-genomics), but for now sequencing a species genome can still run into the tens to several thousand dollars.

However, knowing the complete nucleotide sequence of any given organism still does not elucidate the full picture of biological knowledge of that organism. In real time (as opposed to evolutionary time), DNA sequences are relatively static. With the exception of specific loci, such as eukaryotic immunoglobulin genes, telomere and mobile elements, DNA sequences rarely change during the lifetime of the individual. In addition, many biologists also want to know: "what do genes actually accomplish with respect to the function of the cell and what is their effect?" "Which genes are active (i.e. being "expressed" in functional proteins) or inactive?" "How is gene activity regulated?" "Does the array of active and inactive genes vary according to cell or tissue type, or does this arise from external or environmental cues?"

To answer some of these questions, the field of functional genomics has arisen, with one of its major objectives being to discover the actual function of anonymous genes. A necessary focus of functional genomics is the transcriptome, which is defined as the total population of transcribed RNA (ribonucleic acid) molecules whose biological information is required by the cell at a particular time. This RNA population includes messenger (mRNA), transfer (tRNA) and ribosomal RNA (rRNA) molecules, which are all involved in the essential function of protein translation in all cells. Although similar in structure, RNA molecules differ from DNA in the following: ribose is the sugar present in the 5′ to 3′ phosphodiester backbone instead of deoxyribose; uracil substitutes for thymine in RNA; most RNA molecules are single-stranded; and generally RNA is more labile (unstable and prone to degradation).

Based on these foundations of molecular biology, modern transcriptomic studies focus on the second tier of genetic information coding: the "messenger" RNA (mRNA). mRNA molecules are transcribed from the primary genetic code of DNA, but it is the mRNA, after introns are spliced out of eukaryotic genes, that codes for the actual peptide sequence that comprises protein building blocks and enzymes. Bacterial genomes are very economical, as introns are absent, and so splicing is unnecessary. Therefore most of the bacterial genome is encoded into mRNA. In contrast, only ~ 4% of eukaryotic genomes are converted or "transcribed" into cellular RNA.

At least one caveat that should be remembered throughout this discussion is that levels of mRNA and their associated translated peptides, do not always correspond 1:1. Because mRNA and protein molecules each have varying half-lives in a particular cell or tissue type, degradation of each may be unequal. Bacterial mRNA only prevails in a cell for about 1 1/2 minutes on average, while in growing vertebrate cells, some mRNAs may have half lives approaching 3 hours. Proteins, because of their structural and enzymatic activities, may have a much longer

duration, spanning several hours to days. Therefore, the absence of any one mRNA transcript does not necessarily mean that the corresponding protein is also absent. The converse can also be true and so the content of the transcriptome does not always equal that of the proteome.

Bioinformatics and functional genomics research continues to expand the knowledge of various organismal metabolic regulatory networks. For example, a new regulon was discovered by Rigali *et al*. (2004), who used *in silico* approaches to find new transcription factors to the HutC regulon that comprises genes of the phosphotransferase system (PTS) from anonymous gene sequences. It is now possible to identify the up or down-regulated biosynthetic enzymes and transcription factors (TFs) using a comparative approach and molecular biology (transcriptome) methods.

In general, eukaryotic and prokaryotic cells can express a wide range of mRNAs with the number of transcripts ranging from from several thousand to only a few hundred copies per cell respectively. However, most cellular transcripts likely effect "housekeeping" or primary metabolic functions. Indeed, analyses of microbial genomes suggests that these highly expressed housekeeping genes probably evolve more slowly, and it is most likely that the rare transcripts may have more divergent sequences. This is reasonable, if it is assumed that the cell has to "test" the novel sequence before committing a large amount of metabolic energy into coding and translation of a potentially useless protein.

Many functional genomicists are interested in identifying and characterizing new gene sequences as they represent the many gaps in the sequence databases (e.g. GenBank, EMBL) often designated as "unknown" or "hypothetical" sequences after BLAST database searches. To accomplish these identifications, it is more efficient to enrich (or synonymously to subtract) for specific target mRNAs prior to full scale, expensive, sequencing and informatics analyses. To accomplish the enrichment and facilitate the identification of cDNAs that may be involved in more exotic metabolic processes such as secondary metabolism, defensive mechanisms or symbiosis, innovative molecular methods such as serial analysis of gene expression (SAGE), suppression subtractive hybridization (SSH) and Differential Display (DD) have been developed.

As an example of this technology, a recent study with malarial sporozoites successfully used SSH to identify twelve novel genes encoding predicted proteins containing signal peptides, indicating their involvement in a sporozoites' secretory pathway relative to merozoite developmental stage. Another study characterized differentially expressed *Vibrio cholerae* genes to find those related to toxin production, metabolism, signal transduction, luminescence, mobile elements, stress resistance, and virulence. In a third study, transcriptome mining of *Plasmodium falciparum* revealed complex patterns of regulatory transcripts involved in the pentose phosphate pathway (PPP). Several transcripts of PPP enzyme-coding genes displayed a biphasic pattern of transcription unlike most transcripts that peaked only once during the parasite cycle. The physiological meaning of this pattern remains to be elucidated.

Overall, the use of differentially enriched transcriptome characterization circumvents the need to carry out more expensive large scale genomic DNA sequencing projects. Since most genomes span at least a few hundred thousand to several million base pairs (for prokaryotes) and over 600 billion bp in unicellular eukaryotes (table 1), concentrating efforts on the active transcriptome of an organism eliminates 50-95% of unnecessary (non-transcribed) DNA sequencing effort. Below is a more detailed listing of various current methods for studying mRNA expression.

Table 1. Representative genome sizes

Species	*Approx. Genome size (kilobase pairs)**
E. coli	4,700,000
Saccharomyces cerevisiae (yeast)	12,000,000
Drosophila melanogaster (fruit fly)	180,000,000
Canis familiaris (dog)	2,900,000,000
Homo sapiens (human)	3,600,000,000
Paramecium caudatum (ciliate)	8,600,000,000
Amphiuma means (2 toed salamander)	84,000,000,000
Amoeba dubia (amoeba)	690,000,000,000

* for exact sizes visit
http://www.ncbi.nlm.nih.gov/sites/entrez?db=Genome&itool=toolbar

2. A WIDE SELECTION OF LABORATORY METHODS FOR MRNA ANALYSIS

2.1 cDNA Synthesis and Cloning

In the first step of this process, fragile mRNAs must be converted to the DNA structure that is more stable. cDNAs are "complementary" to the mRNA coding strand (the sense strand of DNA), and are synthesized by well known "reverse transcriptase" (RT) enzymes derived from RNA retroviruses. cDNAs are synthesized one strand at a time. To prime the reaction, an iterative stretch of 20-25 thymine nucleotides (i.e. 5'dTTTTTTTTTTTTTTTTTTTT 3'), commonly called "oligo dT" primers are applied to eukaryotic mRNAs. The basic scheme of cDNA first and second strand synthesis reactions are shown in Fig. 1, but also documented in more detail in standard molecular genetics laboratory manuals such as Sambrook *et al* (2001).

Since cDNAs stem from a "mixture" of different mRNA transcripts within a cell, it is necessary to separate each unique cDNA molecule for specific, individual analyses. For this reason, cDNA "libraries" are generated using molecular cloning and genetic engineering. These libraries are recombinant versions of the original mRNA pool of a cell, which are then cloned into synthetic genetic vectors such as plasmids, phage, cosmids or fosmids. Afterwards, each clone represents one specific cDNA. Many cDNA methods have been developed through the years, mainly focusing on the efficiency of cDNA synthesis, extending reverse transcription using enzymes such as SuperRT in order to obtain longer cDNA molecule lengths. The overall goal is to try to capture as many of the mRNA molecules from a particular tissue or organism into renewable and sustainable high quality "archive" of molecules that can be amplified as needed. Although very important in its goal of isolating pure forms of any specific gene or DNA sequence, the "cloning" of DNA molecules and PCR products has become a routine task in molecular biology, with many various modalities.

2.2 Micro arrays for Gene Expression Profiling

The quest to "see" or witness actual cellular processes as they occur represents a major theme of this chapter, and indeed, represents a holy grail for many cell biologists. The technique of gene

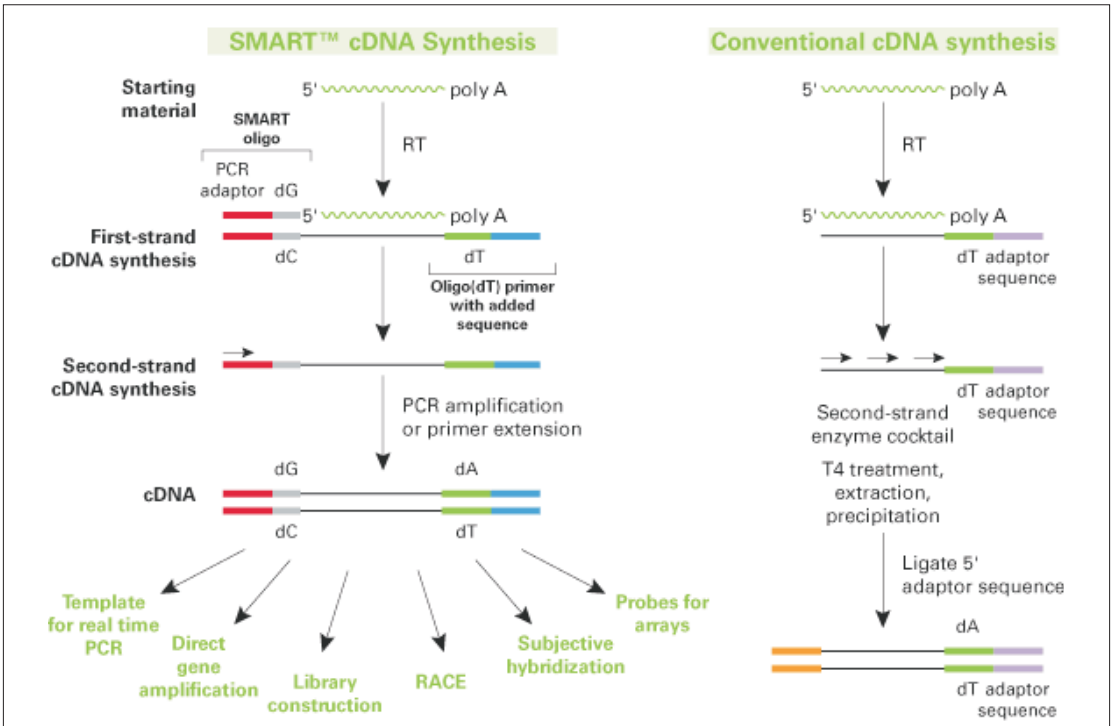

Fig. 1 Comparison of novel and standard cDNA synthesis reactions in eukaryotes. PolyA tail is exploited in both. Courtesy of Clontech Laboratories, Inc., a Takara Bio Co., Clontech Laboratories, Inc.

microarrays in essence permits this glimpse, albeit not totally *in situ*, and only after the live cell itself has been sacrificed. To date, no imagers or microscopes that can view all of the activity in an individual cell in real time exist, and therefore researchers must rely on indirect methods for viewing cellular activity at this time.

DNA micro arrays were first developed in the late 1990s, interestingly with the help of the same researcher, Sir Edwin M. Southern, who gave his name to the "Southern blot", one of the most widely used methods and cited papers in the 1970's and 80's for probing gene sequences immobilized on nitrocellulose and nylon filters after electrophoretic separation. The same principles of hybridizing complementary gene "target" and "probe" molecules together on a fixed surface were applied for the newer microarrays, though on a smaller surface area, such as the glass surface of a slide. Technical details will be bypassed here, but can obtained from several primary references. Micro arrays were first applied to determine changes in gene expression based on glucose or ethanol addition to growing cultures of the model organism *Saccharomyces.*

Since then, microarrays have been one of the most heavily used technologies to study differential gene expression. It can be applied to differences between tissues in the same organism, between different species, and even between different organismal communities. One key unifying principle in microarrays is that the labeled nucleic acid molecules (DNA or RNA "probes")

hybridize in solution with high sensitivity and specificity to complementary sequences immobilized on a solid substrate. Hybridization allows simultaneous and parallel quantitative measurement of many different target sequences on the array and in the complex hybridizing mixture. The main strength of the method and one of the main benefits of microarrays is their ability to study thousands of genes at the same time.

Two major micro array approaches have been documented by Cummings and Relman (2000):

2.2.1 Spotted microarrays

DNA fragments, such as library clones or PCR products (amplicons), are physically attached or spotted onto a solid substrate (using a robot arrayer or capillary printing tips), and can reach up 23,000 /microscope slide. The array is then probed with a second or different cDNA library. The advantages of this approach include the relatively low cost and flexibility; and primary sequence information is not necessary to print the original array.

2.2.2. DNA "Chips" or Printed arrays

Spotting synthetic single stranded "oligonucleotides" in situ by use of photolithography. These are probed with a labeled cDNA library, with the advantage of very high information density (>280,000 elements or DNA sequences on a 1.28x1.28 cm array), eliminating the need to collect and store cloned DNA or PCR products. This approach now has many variations and has been most popularized by Affymetrix Inc, (www.affymetrix.com) which has mass produced "gene chips" and also moved several steps further by generating specific organismal genomes or transcriptomes on customized chips for probing.

Overall, the raw microarray data appears as a grid of multi-colored fluorescent spots, as shown in fig. 2. The appearance can appear as a constellation of genes, though representing only a portion of one small galaxy among a seemingly infinite universe of genes and genomes (remember that there are at least 1.8 million catalogued species on earth, and probably up to more than 10 million, and each species has on average 1000 – 10,000 genes, then there may be 10^{10} different genes in the world.). Each spot represents a gene that has been hybridized with a matching mRNA sequence that has been labeled with a fluorescent dye, such as Cy3, Cy5, or fluorescein. In many dual color assays, differently labeled probes are simultaneously hybridized to the array. The relative intensities of each color can be used to measure the relative abundance of specific mRNAs. Alternatively, one color will indicate higher expression of that gene, while the second color designates lower gene expression. Typically red indicates higher expression in one treatment, while green fluorescence designates relatively lower expression. When probes are simultaneously hybridized, a yellow color appears indicating similar expression in both treatments. Hybridization kinetics of nucleic acids has been detailed for many years and the same parameters apply in microarrays, differing primarily by the miniaturized scale.

Despite the many advantages of globally querying a cell's mRNA population in a single assay, several drawbacks remain. Some paralogous genes in the same genome may share high DNA sequence identity and thus cross hybridize leading to false positives. All positively detected genes eventually require verification. This is done using techniques, which inactivate genes, such as RNA interference, anti-sense RNAs, and gene specific mutagenesis. Next, normalization of the

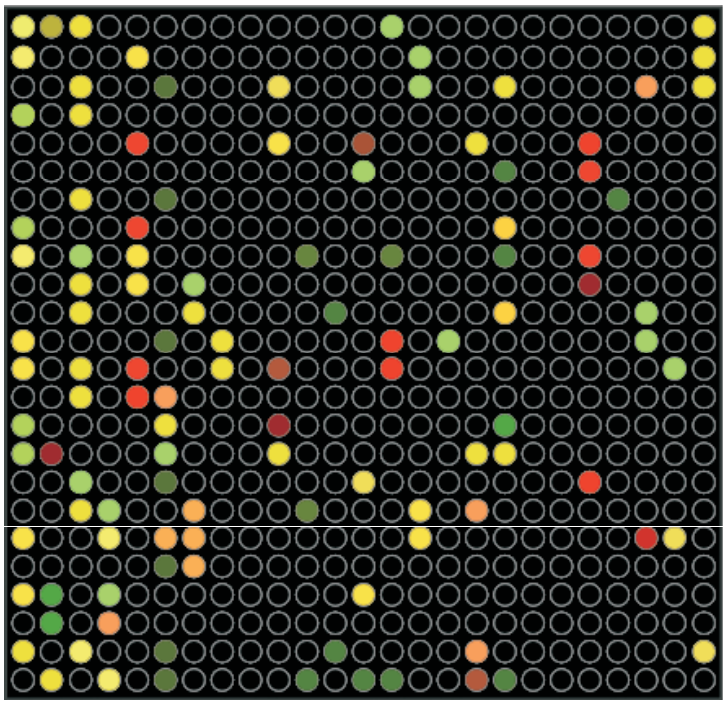

Fig. 2. Typical model of cDNA microarray results after hybridization with two different labeled probes. Courtesy of Jim Masterson (HBOI)

hybridization signals must be performed, in order to obtain more quantitative assessments of relative gene expression.

Finally, the previously unknown genes must be identified and characterized. This can be easily achieved when the microarray printing process keeps track of which genes or oligonucleotide sequences were placed on the slide. For anonymous DNA targets, more extensive bioinformatics analysis must be performed. The data from microarrays is analyzed through the use of clustering algorithms, which are applied to find patterns of expression that may show tissue or metabolism specific correlations between and among genes.

Nonetheless, the main benefit of microarrays – the ability to study numerous genes at the same time – may also be one of its biggest problems. This is because it is possible to observe cyclicity merely from slight statistical variations due to random noise and experimental variations in the assay procedure.

The cost of conducting microarray studies has to be taken into account, as the equipment and reagents can still be prohibitively expensive. Essential equipment to begin includes a microarray printer (usually robotic when printing dense arrays), hybridization fluidics platforms, microarray readers and scanners to detect and quantify the fluorescent signals, and specific software to run the clustering algorithms such as hierarchical and K-means clustering to sort and

interpret the data. Hundreds of publications focus on the microarray method, and we refer the reader to several excellent papers and reviews for validation methods.

2.2.3 Differential Display (DD)

Amplification and comparison of mRNA transcripts (cDNAs) can also be performed using mRNA differential display (DD). In a similar way to microarrays, but on a less global scale, DD enables characterization of relevant molecules that are either up or down regulated. A complete description of the DD protocol has been previously described, while Fleming *et al.* (1998) has shown successful use of the DD technique in prokaryotic systems.

The technique has been patented by the inventors, and commercial kits (GenHunter, Nashville, TN) are now available which streamline and customize the process (see www.genhunter.com).

The DD method operates by systematic RT-PCR amplification near the 3' terminus of mRNAs (cDNA). Typically, an oligo dT primer is used since eukaryotic mRNAs often have poly-adenylated 3' termini. Since most prokaryotes do not have polyadenylated mRNAs random 14mers, instead of "anchored" oligo dT primers, can be used for reverse transcription followed by PCR amplification with additional upstream primers of arbitrary sequences. This allows direct side-by-side comparison of most of the mRNAs between or among related cells or tissue samples. The DD method is extraordinary in its capacity to survey most of the expressed genes of a cell in a sequence-dependent manner, without requiring a priori sequence data. It is theoretically possible to screen all transcripts from an organism if all possible combinations of arbitrary and "anchored" DD primers are used. Moreover, even partial fragment amplifications by DD will generate sequence information that can be used to develop probes to isolate more complete cDNAs and/or corresponding genomic DNA. Sufficient amounts of high quality mRNA can be easily obtained from cultures with either a) Microbial Express Bacterial mRNA enrichment kit (Ambion) or DD kits manufactured by GenHunter. An example of a DD study is included in the case studies section below.

2.2.4 SAGE (Serial Analyses of Gene Expression)

A related but a bit more complicated mRNA analyses method is called SAGE, developed by Velculescu *et al* (1995). This method relies on the generation of a short sequence "tag", that corresponds to a unique identifier of each cDNA molecule. The approach stems from the calculation that a 10-bp sequence can code for approximately 4^{10} (1 048 576) different sequences – unique indeed. These tags are generated enzymatically, concatenated, cloned in a "series" and then sequenced. For practical reasons, we do not explore SAGE in more depth, but refer to more detailed descriptions by Patino *et al*. (2002), and the website http://www.sagenet.org/pubs/index.html.

2.2.5 Real Time (Quantitative) qPCR

Real-time PCR methods were originally developed to quantify the number of specific DNA template copies in a sample. However, when applied to cDNA templates, they can also help determine relative quantities of expressed genes in an organism.

There are various ways to apply qPCR, but most revolve around the use of fluorescent dyes that bind or "intercalate" between the two strands of a nucleotide or ribonucleotide double helix. Standard qPCR using only a single fluorescent dye, such as SYBR Green, will show a proportional increase in fluorescence of all double stranded PCR products as the number of PCR reaction cycles increase. A single-step amplification/detection variation of qPCR is the Taqman® reaction (licensed by Applied Biosystems Inc or ABI). Instead of SYBR green, Taqman utilizes a target-specific fluorescent probe located between the two typical unique PCR primer sequences. Taq polymerase moves along the template strand, releasing the fluorophore from the proximity of a quencher dye thus generating a fluorescent signal. For every round of amplification, each DNA molecule generates a fluorescent signal that quantitatively represents the initial amount of DNA template within a given sample. Taqman PCR rapidly generates results and also provides increased target specificity due to the presence of a unique probe oligonucleotide in addition to the primers. In contrast to conventional electrophoretic resolution methods many real time systems such as the SmartCycler (Cepheid) and iCycler (BioRad), allow the data to be viewed as it is generated. For analysis of complex microbial communities, Taqman PCR offers increased target specificity over standard SYBR Green qPCR and the ability to quantitatively determine the relative abundance of both culturable and unculturable organisms within the population.

It is important to normalize qPCR reactions. This is generally achieved through the use of a baseline standard and the inclusion of a "control" template. This control is usually a gene that is expressed at constant levels within the cell, an example for eukaryotic cells is actin, a cytoskeleton protein. Alternatively, standard curves based on known template concentrations can be drawn, and then levels of expression from unknown physiological samples are compared and extrapolated to these curves. An example of this is shown in fig. 3, which shows standard curves of Ct generated for several genes and cytokines (IL-1β, IL-6, tumor necrosis factor alpha (TNF-α), matrix metalloproteinase 1 (MMP-1) and GAPDH) using quantitative PCR on activated peripheral blood mononuclear cells. The lower the Ct value corresponds roughly to the PCR reaction cycle, and so lower Ct's indicate a more efficient PCR amplification.

2.2.6 Subtractive Suppression Hybridization (SSH) PCR

Another molecular transcriptomic approach for enrichment and detection of secondary metabolite transcripts involves the normalization of transcripts through subtractive suppression hybridization (SSH) PCR, whereby pools of housekeeping (e.g. primary metabolite) genomic DNA or cDNA fragments from an organism of interest (target or "tester") are, in effect, depleted by combined hybridization and PCR amplification of sequences that are also in a reference strain (elicitor or "driver"). The remaining cDNA fragments, highly enriched for tester-specific (secondary metabolite) sequences, are then cloned for further analysis. Driver cDNAs should represent mostly unwanted, common "housekeeping" genes. Differences in mRNA abundance are alleviated by the hybridization step that normalizes (equalizes) sequence abundance during the course of subtraction by standard hybridization kinetics (Fig. 4). This technique is fully described in SuperSmart PCR cDNA Synthesis and PCR-Select^{TM} cDNA Subtraction Kits provided by Clontech (San Jose, CA), which we have employed, and can also be applied to prokaryotes.

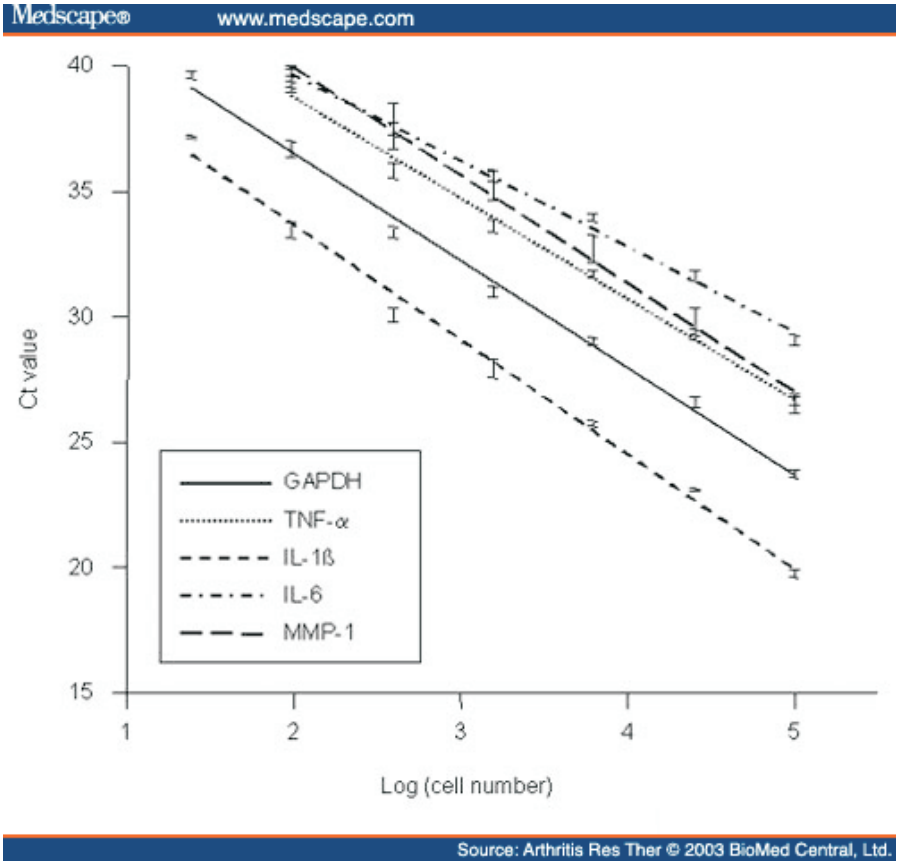

Fig. 3. Standard curve expression of IL-1β, IL-6, tumor necrosis factor alpha (TNF-α), matrix metalloproteinase 1 (MMP-1) and GAPDH. The graph shows the threshold cycle, C (t), where detectable PCR product was observed, versus the template from a known number of cells. Note that the slopes of each target gene differ, reflecting different amplification efficiency. Unless the slopes are identical, the housekeeping gene cannot be used directly to normalize for RNA content.

Arthritis Res. Ther. 5, R352-R360, 2003. © 2003 BioMed Central, Ltd.

The products resulting from the hybridization include (a) single stranded tester transcripts that are not present in the driver sample, (b) double stranded tester transcripts formed by re-annealed tester cDNA, (c) hybrids of tester and driver found in both samples, and (d) double-stranded driver transcripts. Normalization of transcript abundance is achieved because abundant differentially expressed transcripts tend to form (b) molecules than rare transcripts, based on hybridization kinetics. A second hybridization is then performed from a mixture of the

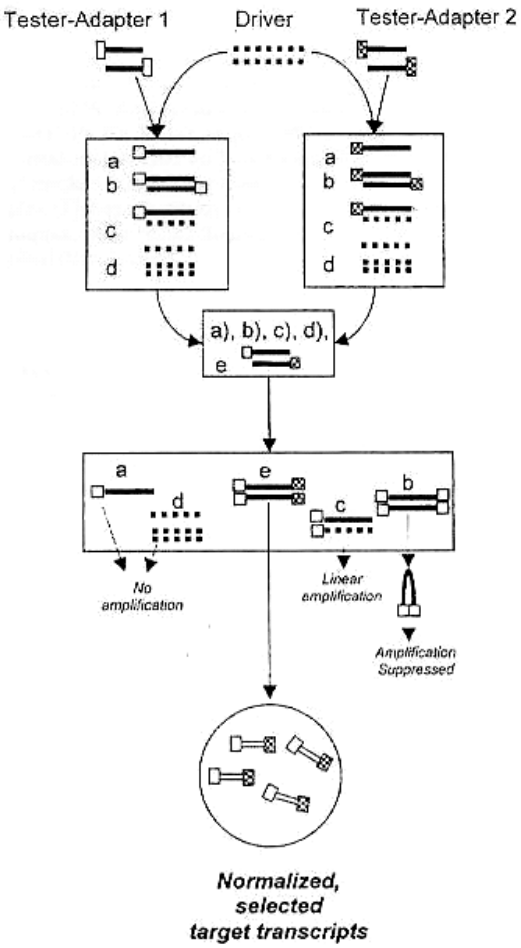

Fig. 4. Basic scheme for suppression subtractive hybridization (SSH) based on ClonTech's SuperSmart PCR cDNA Synthesis and PCR-Select™ cDNA Subtraction Kits. Small white or speckled boxers represent synthetic adapters discussed in the text.

original two hybridization products, plus additional denatured driver cDNA. The resulting products are the same as the above, except that the (a) products from each pool now anneal to each other to form (e) double stranded, differentially expressed transcripts. Enrichment of (e) molecules can then be performed by PCR amplification, since only (e) molecules have two different adapters at each terminus. Again, the primary goal of SSH is that unique genes can be separated from more mundane housekeeping genes. Examples of this and other expression profiling applications are discussed below.

3. CASE STUDIES

3.1 Coral Transcriptomics I – "Cnidarian-Algal Symbiosis"

As an example of marine transcriptomics, Rodriguez-Lanetty *et al* (2006) performed a marine subtraction study in order to identify genes that may be involved in maintaining the symbioses between the Pacific coast temperate anemone *Anthopleura elegantissima* and its dinoflagellate endosymbionts, usually from the genus *Symbiodinium*.

It is well known that this type of intracellular association is centered around nutrient exchange, since both partners live in oligotrophic waters of coral reefs. Because some anemones live in either a symbiotic or aposymbiotic state (Fig. 5), it was possible to compare the proteomic and transcriptomic profiles of each state, and therefore identify genes potentially associated with symbiosis. Allowing for a 5% type I error for the whole set of significant genes, only 189 features, or 1.82% of the original 10,368 features on a cDNA microarray, showed significant differences in expression between the aposymbiotic and symbiotic anemones ($P < 0.05$). After DNA sequencing and sequence analyses, these 189 features resolved into 91 unigenes (Fig. 6).

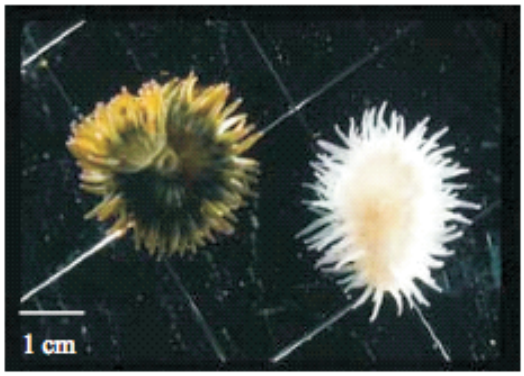

Fig. 5. Photograph of a symbiotic (left) and aposymbiotic (right) *Anthopleura elegantissima*. (Rodriguez-Lanetty *et al*., 2006)

The final analyses indicated several metabolic classes and pathways implicated in symbioses, such as lipid metabolism, apoptosis suppression, and down-regulation of a copper/zinc superoxide dismutase (CuZnSOD). Phytanoyl-CoA hydroxylase, an enzyme involved in lipid degradation, which catalyzes the conversion of phytanoyl-CoA to 2-hydroxyphytanoyl-CoA in beta-methylated fatty acid metabolism, was more highly expressed in symbiotic compared to aposymbiotic anemones (1.54 fold from array data and 1.59 fold from qPCR data). In contrast, the enzyme medium-chain S-acyl fatty acid synthase, involved in lipid synthesis, was down-regulated in symbiosis (1.56^{-1} fold from array data and 1.32^{-1} fold from qPCR. Overall, these differences suggested that the symbiotic state has a profound effect on the rates of lipid metabolism.

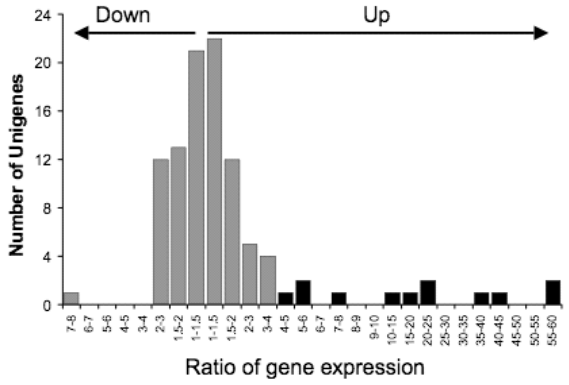

Fig. 6. Distribution of unigenes as a function of the ratio of expression between symbiotic and aposymbiotic anemones. Arrows divide those down- and up-regulated in symbiosis. Black bars denote unigenes confirmed by PCR to be contaminating algal unigenes (see text for details). (Rodriguez-Lanetty *et al.*, 2006)

3.2 Coral Transcriptomics II – "Coral SSH"

In a separate study of marine invertebrates, the author's and collaborating laboratories have validated the use of SSH PCR to enrich for potential secondary metabolite (SM)-related transcripts in the soft coral *Erythropodium caribaeorum* (EC), the source of eleutherobin, a potent chemotherapeutic agent with high activity against breast, renal and lung cancer. A DD analysis was completed and an SSH study was performed hybridizing *E. caribaeorum* cDNAs, as the "tester" organism, to the common sea fan *Gorgonia vetulina*, which served as the non-SM "driver" template. After sequencing about 200 SSH enriched clones, some of resulting sequences share similarity with monooxygenases, sporulation and quorum sensing (a bacterial response to increasing threshold numbers of cells and accumulated signal molecules such as homoserine lactone molecules. Other previously un-described molecules suggesting novel or non-housekeeping functions (table 2) were also found as were several "Transcription-factor like" sequences, including a zinc finger motif.

Table 2. Representative cDNA sequence identities of DD or SSH clones from pilot studies

Clone	*Method*	*Sequence length (bp)*	*First BLASTX match*
140-40	SSH	>156	Transcription factor, 77/156 (49%), cysteinyl-tRNA synthetase (25%),
140-46	SSH	470	Transcription factor, 77/156 (49%), unnamed protein product= (48%)
140-15	SSH	705	Putative zinc finger , 81/232 (34%),
156-05	DD	>180	Sporulation protein, *Streptomyces avermitilis* 42/60 (70%)
156-33	DD	100	two-component sensor [*Pseudomonas aeruginosa* PA01], 25/26 (96%)
156-25	DD	>150	cobyrinic acid a,c-diamide synthase, 40/57 (70%),

156-57	DD	290	hypothetical protein SC0249 [*Salmonella*] 61/94 (64%),
1-G1	SSH	434	putative fatty acid desaturase [*Pseudomonas aeruginosa*], 13/37 (35%),
1-G7	SSH	554	hypothetical protein [*Escherichia coli*] 37/61 (60%),
1-G8	SSH	751	SocE [*Myxococcus xanthus*], 59/117 (50%),
1-D3	SSH	397	bzip transcription factor C/EBP [*Podocoryne carnea*], 35/70 (50%),
1-B11	SSH	458	Probable cyclase hisF [*Mycobacterium tuberculosis* H37Rv], 21/76 (27%),
1-B5	SSH	741	hypothetical protein PC403819.00.0 [*Plasmodium chabaudi*], 69/73 (94%),
1-B11	SSH	458	Signal transduction Histidine Kinase [*Thiobacillus denitrificans* ATCC 25259], 28/84 (33%),

Although verification of gene function requires further analyses and qPCR experiments, the high frequency of "hypothetical" (unknown) sequences shows the success of both DD and SSH in excluding most housekeeping, tRNA and rRNA sequences, while enriching for interesting content (novel or regulatory sequences). Both DD and SSH methods permit the screening of other non-housekeeping metabolic molecules identified in culture, since arbitrary (10-14 bp) primers can be used, obviating the need for a priori sequence data of specific metabolic loci.

3.3 Differential display (DD) analysis of marine sponge-associated microbes

Microbial metabolic diversity has been acknowledged to greatly exceed the metabolic capabilities of eukaryotes, and thus many microbes probably contain a large untapped reservoir of genetic, biochemical and metabolic capabilities. Many microbes exist in symbiosis with marine invertebrates, and some of these symbioses may be the source for chemical defense, toxin and bioactive secondary metabolite production in sponges and soft corals.

A bacterium (HBMMCC #H756) was isolated from a deep water marine sponge by the Harbor Branch Oceanographic Institution's microbiology laboratory, it is 99% similar to *Pseudomonas aeruginosa* and produces pyoluteorin, a pyrrole antibiotic pyoluteorin.

It appears that the expression of SM biosynthesis is more dependent on complex environmental stimuli, such as nutrient deprivation, responses to stress or competition, and quorum sensing than is primary metabolism. As a part of this process it is hypothesized that co-culturing multiple bacteria will result in the expression of multiple pathways and the production of antibacterial agents. Since the levels of pyoluteorin production had been found to be variable in a series of batch cultures we undertook a project to determine the effects of co-culture on pyoluteorin expression and to determine which genes are differentially expressed in reponse to co-culture., A secondary goal of this research was to optimize and define reproducible conditions to recover novel mRNAs from co-culture systems using the DD technique. The differential display strategy as illustrated in Fig. 7 can be subdivided in three parts. The first stage was to generate a collection of cDNAs fragments from a population of mRNAs using the enzyme reverse transcriptase (RT) and an arbitrary primer. The second stage utilized the subset of cDNAs, which is then employed as a template for PCR amplification with another arbitrary primer. By using

different primer combinations (36 possible) the differential transcripts were detected in the first part by side comparison of cDNAs on a polyacrylamide gel electrophoresis. Comparison between mono- and co-cultures would then be expected to identify novel differentially expressed bands, which could then excise from the gel, cloned and sequenced (fig. 8).

Using this strategy, we (the authors) identified and characterized more than 20 differentially expressed bands that might affect the genome of the bacterial co-culture H-756-H720 and H-756-D727. This complex pattern of bands has a size range of 100bp to 3,000bp. Cloning and subsequent sequence analysis of these excised bands revealed the genes or transcription factors that encode and affect the metabolism of this antibiotic.

A technical problem forced a modification of the polyacrylamide gel method. Initially, silver staining was used to observe the banding pattern, however these bands could not be reamplified

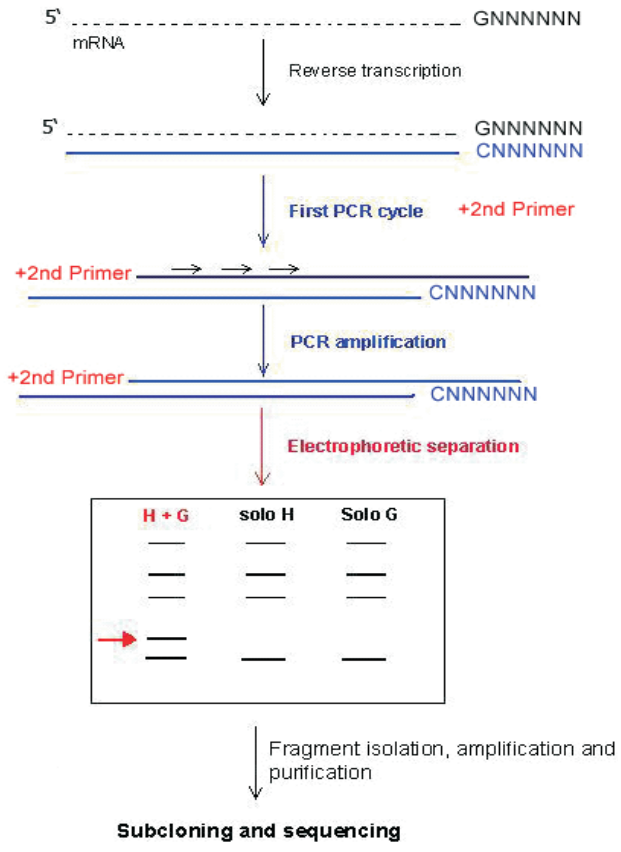

Fig. 7. Schematic representation of arbitray differential display on bacteria. Reverse transcription of mRNAs is performed with MMLV reverse transcriptase and arbitrary primer (5' NNNNNNNC 3'). PCR amplification of the cDNAs uses the same arbitrary primer plus another random primer. After electrophoresis the products are displayed on denatured polyacrylamide gel. The different set of primers chosen randomly in the RAImage® kit (GenHunter Inc) allows the visualization of most of the gene expressed in the bacteria.

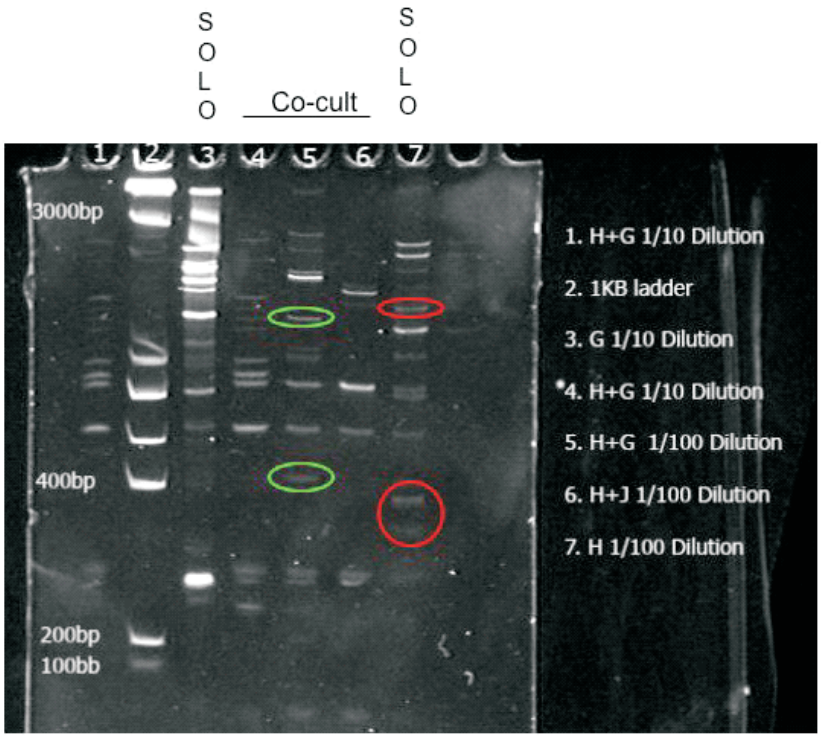

Fig. 8. Representative DD polyacrylamide gel stained with SYBR Gold. Green and red circled bands indicate possible up-regulated (in co-cultures) and down-regulated transcripts, respectively. Note that possible PKS repressor TFs (red circled bands) can only be detected in the solo culture of the PKS+ microbe (e.g. H756). Positive regulators (green) may stem from either microbe in co-culture. H= H756, G = H720 and J = D727. Some of the differential bands were cut and sequenced (Table 3). For example, the top red band matched to a hypothetical protein (156-57). One of the un-circled common bands found in all lanes matched a housekeeping sequence, pseudouridylate synthase. Work was performed by summer intern B. Mouzon.

and this led the authors to switch to the SYBR® Gold staining method. The sensitivity of SYBR® Gold is comparable to that of silver staining and resulted in significant savings in both time (20minutes to stain a gel as opposed to 3 hours for the silver staining method) and cost ($0.50 for SYBR® Gold vs. $5.00 for silver staining).

After cloning and sequencing several differentially expressed bands, the nucleotide sequences of the clones were determined and analyzed. Overall none appeared to match the pyoluteorin gene cluster sequences, however these proteins were comparable to the enzymes associated with the antibiotic theorical biosynthetic pathway. One clone, X4, encoded a protein referred to as 2-isopropylmalate synthase with a significant homology (56% identity). This enzyme has a transferase activity, which converts acyl into alkyl groups (leucine production). This clone also has a match with a putative methyl-accepting chemotaxis protein (56% identity), and a choline

dehydrogenase (64% identity). This last enzyme catalyzes the conversion of exogenously supplied choline into the intermediate glycine betaine aldehyde. The role of the putative methyl-accepting chemotaxis protein has not yet been defined.

4. CONCLUSIONS

Overall, this chapter has attempted to portray the versatility of molecular techniques for studying gene expression at the mRNA level. Despite the difficulty of working with some marine species, and the emergence of techniques in primarily human and terrestrial organismal systems, we have presented evidence that these methods can be successfully applied to marine systems. Transcriptome analysis gives a single snapshot into the metabolic and genetic activity of a particular cell or organism and therefore has more immediate physiological implications beyond mere genomic DNA sequence data. As in any other quantitative procedure, normalization measures must be implemented, such as in qPCR and microarray experiments.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from Florida Sea Grant (Grant Number SI-06-33 to JVL and R/LR-MB-14 to RK) and the National Science Foundation (Grant Number 0119011 to RK). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF). We also thank Clontech Inc, for use of some of their education materials and figures. Several portions of this chapter were derived from previous work, available in Open Access Format (http://creativecommons.org/licenses/by/2.0).

Further reading - websites

National Center for Biotechnology Information: www.ncbi.nlm.nih.gov/

Univ. of South Carolina marine transcriptomics: http://www.marinegenomics.org/
ESTs: http://www.ncbi.nlm.nih.gov/dbEST/index.html

Microarray studies:
http://cmgm.stanford.edu/pbrown/mguide/
http://www.affymetrix.com/technology/ge_analysis/index.affx
http://www.gene-chips.com/
http://genome-www.stanford.edu/cgi-bin/lung_cancer/adeno/gx?n=gxfig
http://smd.stanford.edu/cgi-bin/search/QuerySetup.pl

Realtime PCR: http://pathmicro.med.sc.edu/pcr/realtime-home.htm

Sage: http://www.sagenet.org/pubs/index.html

FURTHER READING

Brown, T.A **2007**, Genomes, 3rd Ed. Garland Science Publ., Oxon UK

Bull, A.T., Ward, A.C., Goodfellow , M. **2000**, Search and discovery strategies for biotechnology, the paradigm shift. *Micro. Mol. Biol. Rev*. 64, 573-606

Chadwick, D.J., Whelan, J. **1992** (eds), *Secondary Metabolites, Their Function and Evolution.* Ciba Fndn., Wiley, Chichester

Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S. **2003**, A vision for the future of genomics research. Nature. 422, 835-847

Cummings, C.A., Relman, D.A. **2000**, Using DNA microarrays to study host-microbe interactions. *Emerg. Infect. Dis*. 6, 513-525

Do, J.H., Choi, D.K. **2008**, Clustering approaches to identifying gene expression patterns from DNA microarray data. 2008. *Mol. Cells* 25, 279-288

Fleming, J.T., Yao, W.-H., Sayler, G.S. **1998**, Optimization of differential display to prokaryotic mRNA, application to pure culture and soil microorganisms. *Appl. Environ. Microbiol*. 64, 3698-3706

Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M., Pääbo, S. **2006**, Analysis of one million base pairs of Neanderthal DNA. *Nature* 444, 330-336

Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., Bork, P. **2008**, eggNOG, automated construction and annotation of orthologous groups of genes. *Nucl. Acids Res*. 36, (Database issue), D250-D254

Kirchman, D.L. **2000** (ed), *Microbial Ecology of the Oceans.* Wiley Liss, New York, pp 1 – 90

Mardis, E.R. **2008**, Next-generation DNA sequencing methods. *Annu. Rev. Genom. Hum. Genet*. In Press

Margulies, M. *et al*. **2005**, Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376-380

Patino, W.D., Mian, O.Y., Hwang, P.M. **2002**, Serial analysis of gene expression, technical considerations and applications to cardiovascular biology. *Circ. Res*. 91, 565-569

Rigali, S., Schlicht, M., Hoskisson, P., Nothaft, H., Merzbacher, M., Joris, B., Titgemeyer, F. **2004**, Extending the classification of bacterial transcription factors beyond the helix–turn–helix motif as an alternative approach to discover new cis/trans relationships. Nucl. Acids Res. 32, 3418-3426

Rodriguez-Lanetty, M., Phillips, W.S., Weis, V.M. **2006**, Transcriptome analysis of a cnidarian-dinoflagellate mutualism reveals complex modulation of host gene expression. *BMC Genomics*. 7, 23

Sambrook, J., Russell, D.W. **2001**, *Molecular cloning*, a *laboratory manual*. Cold Spring Harbor Laboratory Press, New York

Stormo,G.D., Tan, K. **2002**, Mining genome databases to identify and understand new gene regulatory systems. *Curr. Opin. Microbiol*. 5, 149–153

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W. **1995**, Serial analysis of gene expression. *Science* 270, 484-487

Watson, J.D., Baker, T.A., Bell, S.P., Gann, A. **2007**, *Molecular biology of the gene* (6th edition). Benjamin Cummings

Wilson, E.O., Wright, A.B., Ward, G. **2001**, *The diversity of life*. Penguin Books Ltd

Ziebolz, B., Droege, M. **2007**, Toward a new era in sequencing. *Biotech. Annu. Rev*. 13, 1-26