

Chapter 20

Transcriptome Characterization Through the Analysis of Expressed Sequence Tags

Zhanjiang Liu

A good discussion of transcriptome analysis requires us first to review the genetic central dogma and define our terms. The entire genetic material of an organism is defined as its genome. The DNA of an organism is transcribed into RNA and then translated into proteins as the final biologically active molecules. The entire RNA composition of an organism thus is defined as its transcriptome while the complete protein component of the organism is defined as its proteome. Although the genome is relatively stable for a given organism, the transcriptome is dynamic depending on changes in development, physiological conditions, and the environment. The RNA expressed in a cell at any given moment, both the classes of genes expressed and their level of expression, depends on the biological state of the cell.

Sequencing of expressed sequence tags (EST) has been a primary approach in the characterization of the transcriptome. In addition, several approaches such as expression profiling using microarrays, as described in Chapters 21–23, serial analysis of gene expression (SAGE), and massively parallel signature sequencing (MPSS), as described in Chapter 22, have also been used for the characterization of transcriptomes. ESTs are single pass sequences of random cDNA clones. They are partially complementary DNA (cDNA) sequences corresponding to messenger RNAs (mRNA) generated from randomly selected cDNA library clones. Therefore, knowledge and skill in cDNA library construction, library normalization, and analysis of generated EST sequences are essential for a good understanding of the transcriptome. In this chapter, I intend to provide information on construction of normalized cDNA libraries for efficient generation and analysis of ESTs, and the significance of EST analysis for the genome research of species lacking a sequenced genome. A successful EST project can provide the sequence tools needed for not only gene discovery, but also for physical, linkage, and comparative mapping, analysis of alternative splicing and gene duplication, and microarray development.

Construction of Normalized cDNA Libraries for Efficient EST Analysis

EST analysis has traditionally been conducted by sequencing random cDNA clones from cDNA libraries. Such an approach is efficient at initial stages of gene discovery, but has proven to be inefficient in the gene discovery of rarely expressed genes. Theoretically, a typical fish or shellfish species expresses perhaps no more than 30,000

Table 20.1. The top 100 expressed genes account for 33.9% of the transcriptome of the channel catfish head kidney.

No. of Genes	Clones	% sequenced
10	100	9.4
20	302	14.3
30	389	18.5
40	462	21.9
50	520	24.7
60	570	27.1
70	615	29.2
80	655	31.1
90	685	32.5
100	715	33.9

genes, while the human genome contains only some 25,000 genes. At initial stages of EST analysis, the gene discovery rate is almost linear to the EST sequences generated. Thus, it appears, at first glance, that even a small laboratory could complete gene discovery of the entire transcriptome of a species in less than a year. For instance, if one can sequence 100 ESTs a day, it only takes 300 days to complete 30,000 ESTs. However, the rate of gene discovery usually drops precipitously soon after reaching a level of several hundred ESTs. This phenomenon can be easily explained when one considers that several hundred of the most abundantly expressed genes make up the vast majority of the mRNA mass in cells. The chances of encountering cDNAs representing rarely expressed genes, therefore, are small without normalizing the cDNA library. For example, only 100 genes accounted for 33.9% of the transcriptome in the head kidney of channel catfish (Table 20.1). Only 23 genes, sequenced five or more times, accounted for 250 clones of the 1,093 clones (22.8%) sequenced from the catfish liver (Figure 20.1). It is estimated that levels of mRNA can range from 200,000 copies to 1 or fewer copies per cell (Galau et al. 1977). By using regular cDNA libraries, the most abundantly expressed genes would have been sequenced 200,000 times before the most rarely expressed genes are sequenced just once. Clearly, EST sequencing from non-normalized libraries is inefficient for gene discoveries of rarely expressed genes. Normalization decreases the prevalence of clones representing abundant transcripts and dramatically increases the efficiency of random sequencing and rare gene discovery.

Normalized cDNA libraries are cDNA libraries that have been equalized in representation to reduce the representation of abundantly expressed genes and to increase the representation of rarely expressed genes. This concept was initially proposed by Dr. Soares and others (1994) and further modified by Bonaldo and others (1996). In the original protocol, partial extension products of the cDNA library were used as drivers. Single-stranded (SS) libraries were prepared. The SS-library was then extended from poly A for 20–200 bases. The partially extended products were used to hybridize to the cDNA. After hybridization, the double-stranded (DS) fraction is removed by using hydroxyapatite columns, and the remaining SS fraction is converted to DS cDNA and is included in the normalized cDNA library (Soares et al. 1994). The protocol was then

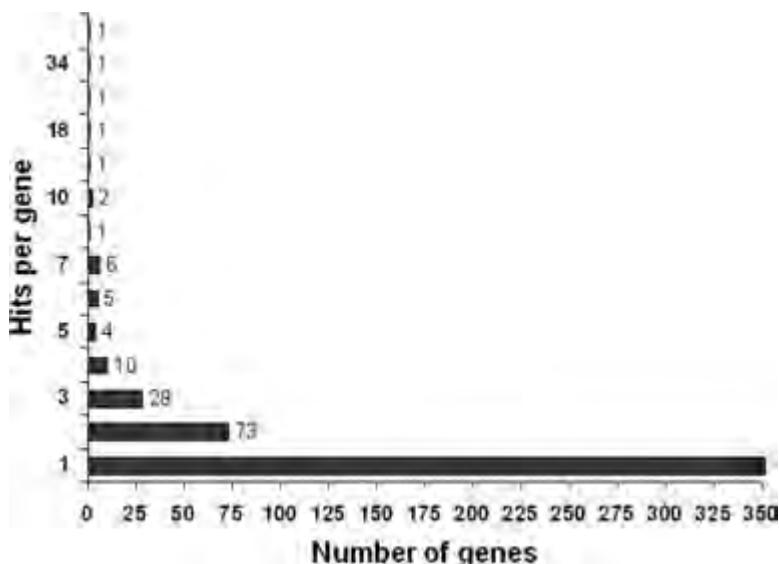


Figure 20.1. Expression profiles and sequencing redundancy among known genes in the analysis of ESTs from the channel catfish liver tissue, adopted from Liu (2006).

modified to use polymerase chain reaction (PCR) amplified cDNA inserts as drivers (Bonaldo et al. 1996). In some other cases, mRNA has also been used as drivers. In these cases, the corresponding mRNAs used for cDNA synthesis were biotinylated with photobiotin and used in excess for hybridization with the cDNA. The hybridization was then treated with streptavidin, followed by phenol extraction to remove the cDNAs that hybridized to the biotinylated mRNA. The cDNA-mRNA hybrids and mRNA are removed by phenol extractions. Only the unhybridized cDNA remain in the aqueous phase for use in construction of normalized libraries. While the details of how the subtraction is conducted may differ greatly, the basic principles behind normalization are the same (i.e., they all depend on the faster hybridization kinetics of abundantly expressed genes to form DS complexes that can be removed by various means, whereas it takes a long time for the rarely expressed genes to reassociate).

We have used the Creator™ SMART™ cDNA Library Construction Kit for the synthesis of cDNA (Zhu et al. 2001). Use of this kit is purely based on the experience of the author whose intention is not to provide a thorough list of methodologies, but to provide a procedure with personal experience. (For alternative methods, readers are referred to Chapter 22.) In our experience, this kit provides high-quality, full-length, directionally cloned cDNA libraries from nanograms of total or poly A⁺ RNA. This system has two unique characteristics. The first is provided by the SMART (Switching Mechanism At 5' end of RNA Transcript) that offers the ability to synthesize full-length cDNA (Figure 20.2). Most commonly used cDNA synthesis methods rely on the ability of reverse transcriptase (RT) to transcribe mRNA into SS DNA in the first strand reaction. In some cases, RT terminates before transcribing the complete mRNA sequence. This is particularly true for long mRNAs, especially if the first strand synthesis is primed with oligo(dT) primers only or if the mRNA contains abundant secondary

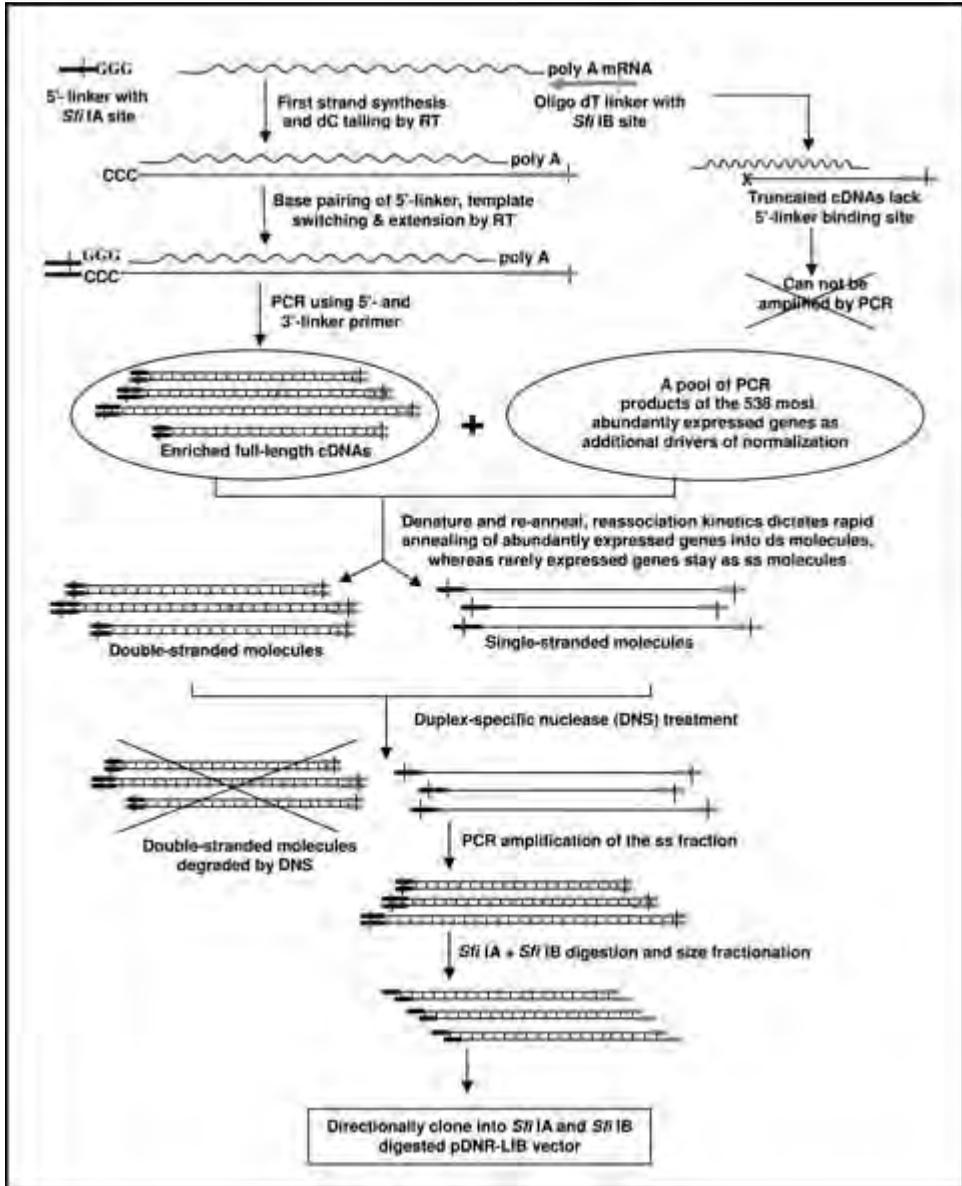


Figure 20.2. Schematic presentation of the procedures for the construction of normalized cDNA libraries, adopted from Liu (2006).

structures. The SMART system is designed to preferentially enrich for full-length cDNAs, while eliminating adaptor ligation. The mechanism for the enrichment of full-length cDNA is the use of a 5'-linker with 3'-GGG tails. RT has terminal transferase activity that preferentially adds three additional Cs at the end of first-strand cDNA. As a result, the first strand cDNA is able to base pair with the 5'-linker with 3'-GGG tails. Once base paired, the RT would switch the template and extend into the linker

sequences allowing PCR amplification of full-length cDNA. Truncated cDNAs are not able to base pair with the 5'-linker, and therefore, get lost in the PCR amplification of the full-length cDNA. The second feature is provided by the Creator system. The Creator system allows the transfer of a target gene from a single donor vector directly into multiple acceptor expression vectors using Cre-loxP recombination. Using this method, any gene cloned into a specialized cloning vector (such as pDNR-LIB) can be transferred into any acceptor vector for functional analysis without the need for subcloning. Since this feature is less relevant to EST analysis, interested readers are referred to the Web site of BD Biosciences (<http://www.clontech.com/clontech/techinfo/manuals/PDF/PT3577-1.pdf>).

cDNA Synthesis

Depending on the availability of biological samples, different protocols should be adopted. If biological samples are limiting (e.g., 50 nanograms [ng] of total RNA), cDNA synthesis can be accomplished by long-distance PCR. In this protocol, a modified oligo(dT) primer (CDS III/3' PCR Primer) primes the first-strand synthesis reaction, and the 5'-linker serves as a short, extended template at the 5' end of the mRNA (Figure 20.2). When the RT reaches the 5' end, the enzyme's terminal transferase activity adds a few additional nucleotides, primarily deoxycytidine, to the 3' end of the cDNA. The 5'-linker, which has an oligo (G) sequence at its 3' end, base pairs with the C stretch, creating an extended template. RT then switches templates and continues DNA synthesis to the end of the 5'-linker. The resulting full-length ss cDNA contains the complete 5' end of the mRNA, as well as the sequence complementary to the 5'-linker. Now long distance PCR is used to amplify the cDNAs using the 3' PCR primer and the 5'-primer. Only those ss cDNAs having a 5'-linker sequence at the 5' end and the 3'-linker sequence at the 3' end can serve as templates for PCR. Incomplete cDNAs lacking the linker sequences will not be amplified, thus allowing construction of cDNA libraries with a high percentage of full-length cDNAs. If biological samples are not limiting (e.g., 1 μ g or more poly A⁺ RNA is available), after the first cDNA is made, direct primer extension is conducted using the 5'-linker to generate the second-strand cDNA.

The quality of mRNA or total RNA (in case of small amounts of starting material) is the key to the construction of high quality cDNA libraries or the normalized libraries thereafter. We generally check the quality of RNA by running a gel. Typically, the 28S and 18S RNA should form tight bands and the mRNA appears as a smear (Figure 20.3). The ratio of the two bands should be roughly 2:1 (28S:18S). If the prominence of the 28S RNA is decreased, it is a reflection of partially degraded RNA.

Normalization and Subtraction

Several strategies have been developed for the normalization of cDNA libraries. The fundamental principles behind all the normalization procedures are the same, and they all depend on the differential hybridization of abundant molecules over rare molecules. We have used a strategy using the Evrogen TRIMMER DIRECT Kit (http://www.evrogen.com/p3_2.shtml). This system is specially developed to normalize

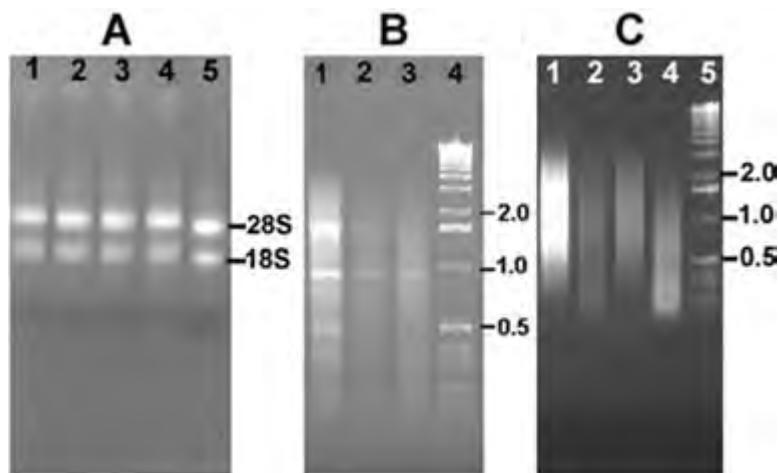


Figure 20.3. Quality checking for RNA (A), cDNA synthesis (B), and normalized cDNA (C). Lanes 1–5 of (A) were RNA from head kidney, liver, skin, muscle, and gill; lane 1–3 of (B) were cDNAs made from muscle, liver, and head kidney. Note the prominent bands over cDNA smear. Lane 4 of (B) is 1 kb molecular weight standard; Lanes 1–4 of (C) were normalized cDNAs from head kidney, liver, and gill with lane 5 of (C) as 1 kb molecular weight standard. Figure was adopted from Liu (2006).

cDNA enriched with full-length sequences (Zhulidov et al. 2004). The method involves denaturation-reassociation of cDNA, degradation of DS-fraction formed by abundant transcripts, and PCR amplification of the equalized SS-DNA fraction. The key element of this method is degradation of the DS-fraction formed during reassociation of cDNA using Duplex-Specific Nuclease (DSN) enzyme (Shagin et al. 2002). A number of specific features of DSN make it ideal for removing DS-DNA from complex mixtures of nucleic acids. DSN displays a strong preference for cleaving DS-DNA in both DNA-DNA and DNA-RNA hybrids, compared to SS-DNA and RNA, regardless of the sequence length. Moreover, the enzyme remains stable over a wide range of temperatures and displays optimal activity at 55–65°C. Consequently, degradation of the DS DNA-containing fraction by this enzyme occurs at elevated temperatures, thereby decreasing loss of transcripts due to the formation of secondary structures and nonspecific hybridization involving adapter sequences.

In addition to the normalization procedures, we have also used additional drivers for subtraction. The subtraction drivers were determined from our previous EST analysis of 40,000 ESTs of catfish. Cluster analysis of the catfish ESTs allowed the identification of abundantly expressed genes. We decided to subtract any message whose representation was more than 2 out of 10,000. In other words, if the transcript was found more than 8 times in the 40,000 ESTs, they were defined as abundantly expressed. This led to the identification of 538 genes for subtraction. The inserts of the EST clones containing these 538 genes were amplified and used as additional drivers for subtraction.

The remaining procedures for the construction of normalized cDNA libraries are straightforward. However, in our experience with the CloneTech's vector pDNR-Lib, a control of "vector only ligation" appeared to be important. The reason is that the

cloning vector is prepared by isolation of the vector backbone after restriction digestion using *Sfi* I endonuclease. In the vector, two *Sfi* I sites exist with different sequences that are referred to as *Sfi* IA and *Sfi* IB. If the vector is digested at only one site, then the vector self ligation can lead to the very high rate of insertless clones. As the quality of the vector preparation may vary from batch to batch, a control ligation of vector DNA only should reveal if the background is a problem immediately.

Significance of EST Analysis, Its By-products, and Applications of EST Resources

The significance of EST analysis has been recognized ever since the first EST analysis experiment was conducted (Adams et al. 1991). However, most scientists who are not familiar with EST analysis may only partially recognize its value. In this chapter, I will provide information concerning objectives of genome research that can be reached by EST analysis.

EST Analysis Is One of the Most Rapid Methods for Gene Discovery and Identification

EST analysis is one of the most rapid approaches for gene discovery. A small collection of ESTs in a species without any genome information can result in the rapid identification of a large number of genes. Gene discovery and identification is, therefore, the primary function of EST analysis.

Back in the 1980s, to clone a gene or a cDNA was very difficult. Ph.D. students were put to the hard task of gene cloning for 3 to 6 months to clone a cDNA. In most cases, the cloned cDNAs were quite highly expressed. Adams and others (1991) put forth a procedure defined as the analysis of ESTs that technically involves direct sequencing of random cDNA clones. However, the outcome was extremely significant. Rather than conducting cDNA cloning through traditional screening, rescreening, purification, and sequencing, direct sequencing of cDNA clones allows rapid gene discovery that often times includes the cDNAs for the genes of interest. At the same time, many other cDNAs are identified that are also of high scientific interest. For instance, during our EST work back in 1997 using manual sequencing, analysis of just 100 clones from the pituitary cDNA library led to the identification of growth hormone, gonadotropins, prolactin, and proopiomelanocortin cDNAs, all of which were of great interest, and the cloning of which would have otherwise taken us a much longer period of time and greater effort and resources (Karsi et al. 1998). Sequencing of 2,228 EST clones from the head kidney tissue allowed identification of 753 distinct known genes plus 739 unknown gene clones (Cao et al. 2001). Sequencing of 1,201 clones from the brain led to the identification of 330 known genes plus 330 unknown genes (Ju et al. 2000). In 2001, the catfish EST collection reached a historical 10,000 clones that allowed the identification of 5,905 genes. These ESTs were the basis for the first aquaculture species to be listed under the TIGR Gene Index in 2002. As of 2005, our catfish EST collection reached 44,000 that represented 25,000 unique gene

sequences. Clearly, it is the high gene discovery rate of EST analysis that has allowed the identification of such a large number of genes. To the best of my knowledge, there is no other method that can provide an equal gene discovery rate as EST analysis while also providing long-term genome resources for many other applications.

Because of the exceptionally high gene discovery rate of the EST approach, EST analysis has been extremely popular. The EST database dbEST has been one of the fastest growing databases at the National Center for Biotechnology Information (NCBI). As of January 20, 2006, there are 32,889,225 entries in the NCBI's public EST database dbEST (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

ESTs Provide a First Glance at Gene Expression Profiles

EST sequencing from non-normalized cDNA libraries should reveal the true representation of the mRNAs in the tissues or cells from which the cDNA libraries were made. Large-scale EST analysis is, therefore, a direct way to conduct expression profiling (Franco et al. 1995, Azam et al. 1996, Lee et al. 2000). It offers a rapid and valuable first look at genes expressed in specific tissue types, under specific physiological conditions, or during specific developmental stages (Ju et al. 2002, Cao et al. 2001, Karsi et al. 2002, Kocabas et al. 2002). To demonstrate this point, once again, let me use the example of our EST work from the pituitary of catfish. Sequencing of just 100 EST clones revealed that the majority of clones sequenced from the pituitary were in the category of hormones (Karsi et al. 1998). As we all know, the pituitary is the organ whose main function is production of hormones involved in various physiological regulations. When large numbers of ESTs are sequenced, a relatively accurate picture can be obtained for gene expression profiling, as has been demonstrated in various species (Kurobe et al. 2005, Song et al. 2004, Kimura et al. 2004, Lo et al. 2003).

When EST analysis is conducted using cDNA libraries constructed from various tissues, the generated ESTs can provide a comprehensive comparison of gene expression in different tissues. In most recent cases, such tissue expression profiling has been mostly accomplished by microarray analysis. However, direct sequencing of ESTs from tissue libraries can provide very similar information. In addition, EST analysis using cDNA libraries from various tissues provides a greater control to investigators as to how deep each library is sequenced. Alternatively, some investigators prefer to attach a molecular tag on each of the tissue cDNAs by using different adaptors attached to the oligo dT primers during the first-strand cDNA synthesis. Such molecular tags allow tracking of the cDNAs for the tissues from which they were derived. This value of EST analysis is undoubtedly signified by the greater power of emerging sequencing technologies (see Chapter 25).

ESTs Provide a Robust Approach for the Study of Alternative Splicing and Differential Polyadenylation

ESTs provide a good deal of information about alternatively spliced and polyadenylated transcripts. Although the number of genes now seems to be quite smaller than

we once thought, the number of distinct transcripts can be much larger. As summarized in a recently published special issue of *Science*, the total number of distinct transcripts can be one order of magnitude larger than the number of genes. (A series of reviews on RNA can be found in the September 2, 2005 issue of *Science*, Vol. 309, No. 5740). Alternative splicing and differential polyadenylation are probably widespread. Different transcripts probably exist in nature for many, if not most, genes. However, using traditional molecular biology approaches, one would have no way of knowing what other transcripts are expressed in the cells, other than the few that have been accidentally identified. In contrast, a wide variety of transcripts are sequenced in EST projects. If primary cDNA libraries are used, the chances of finding various transcripts are proportional to the representation of the transcripts in the mRNA pool. However, if normalized libraries are used, the chances of finding rare types of transcripts are greatly increased.

EST Analysis Is the Most Efficient Way for the Identification of Type I Polymorphic Markers

ESTs provide great opportunities for the identification of Type I polymorphic markers. Markers can be divided into Type I markers and Type II markers. Type I markers are markers associated with genes of known functions, whereas Type II markers are markers developed from anonymous genomic regions. Obviously, when polymorphism is the major interest, sequence variations within gene coding regions are lower as restricted by functional constraints, and thus development of Type I markers is relatively more difficult. However, Type I markers are of greater value because genes are highly conserved across a wide spectrum of evolution. In addition to their value for linkage mapping, Type I markers are also useful for comparative mapping. EST analysis can provide two types of Type I markers: single nucleotide polymorphism within transcripts, and microsatellites associated with ESTs. As a matter of fact, these approaches to developing Type I markers are among the most effective approaches. For instance, a single nucleotide polymorphisms (SNP) rate of 1.32% was found during analysis of 161 genes in catfish, making comparative EST analysis one of the most efficient approaches for the identification of SNP markers within genes (He et al. 2003). Analysis of 43,033 catfish ESTs led to the identification of 4,855 EST clones containing microsatellites (11.2%). Further cluster analysis revealed that the majority (4,103) clones represent unique genes (Serapion et al. 2004). Obviously, not all microsatellites identified in EST analysis are directly useful as markers. Several situations may be encountered. First, the microsatellites may exist at the very upstream of ESTs; second, the microsatellites may exist in the downstream immediately before poly A; and third, the microsatellites may be flanked by simple sequences. In all of these three cases, the identified microsatellites cannot be directly used until further sequences are obtained by genomic sequencing. In addition, flanking sequences used for primer design may amplify introns that are not known from EST projects, making allele prediction difficult. Nonetheless, as a by-product of EST sequencing, EST-associated microsatellites are by far the most efficient way for the development of Type I microsatellite markers.

Related ESTs Are Useful for the Identification of Duplicated Genes

EST analysis provides one of the most efficient ways for the identification of duplicated genes. Gene duplication is a widespread phenomenon in fish. Despite the great debate about the origins of duplicated genes, whether through entire genome duplication followed by differential gene retention and gene loss, or through duplication that does not involve duplication of the entire genome, studies of duplicated genes are still limited by technical difficulties. Even for entirely sequenced genomes, genome assembly can even be hindered by duplicated segments. EST analysis should produce sequences of related transcripts that can be analyzed through phylogenetic approaches. For instance, if two related transcripts are found from channel catfish and one of them is more closely related to a transcript from blue catfish than related to the other transcript from channel catfish, these two transcripts are likely encoded by two duplicated genes, rather than products of allelic variation. The rationale is that allelic variation of the same species should be smaller than the variation between species. In this case, if the two transcripts are more closely related between the two species, the transcripts are likely orthologs. In contrast, if the transcripts are more distantly related, even though it is encoded by the same species, it is likely a paralog (gene duplication products in evolution). Therefore, through large-scale EST analysis followed by phylogenetic analysis, large-scale differentiation of orthologs and paralogs becomes possible.

ESTs Serve as the Basis for Comparative Mapping

The basic concept of comparative mapping is based on the assumption that the genomes of closely related species are highly conserved in the organization of their genes (see Chapter 19). Thus, ESTs can be used as a great resource for comparative mapping. First, EST resources from aquaculture species can be analyzed to determine on which chromosomes their corresponding genes reside in completely sequenced genomes such as zebrafish and *Tetraodon nigroviridis*. We have found that the *Tetraodon* genome sequence is of good quality and works well for comparative genome analysis. For instance, a large number of catfish ESTs have been identified to correspond to genes from all 21 chromosomes of *Tetraodon* (see Table 20.2 for an example). The idea is that if the genome organization is highly conserved between the catfish and *Tetraodon*, then many of the chromosome-specific ESTs should also be located on the same chromosomes in catfish. Testing of this hypothesis should be much easier than mapping blindly without any information. Second, once conserved synteny is identified, the genes between the conserved genes can be easily verified by using the EST resources. For instance, in a recent study involving BAC-end sequencing, more than 100 mate-paired genes were identified on both ends of BAC-end sequences (Xu et al. 2006). Direct Basic Local Alignment Search Tool (BLAST) searches allowed identification of conserved synteny with the zebrafish and *Tetraodon* genomes. Once the conserved synteny is identified, the internal genes between the mate-paired genes on the BAC-ends can be inferred by determining the genes between these genes in the zebrafish or *Tetraodon* genomes. Once the gene identities are determined, sequencing primers or hybridization overgo probes can be designed using the EST resource. Direct BAC

Table 20.2. An example of the utility of ESTs for comparative mapping. Here, *Tetraodon* chromosome-specific catfish EST hits are shown along with their microsatellite repeats for future mapping.

<i>Tetraodon</i> Chromosome	<i>Tetraodon</i> Query	Catfish Hit	E-value	Repeat Location (bp)	Motif	# of repeats
1	CAG04525	BM028034	2E-20	1	gt	11
1	CAF91213	BM425454	6E-19	461	tg	8
1	CAG07953	CB936804	8E-27	527	ca	23
1	CAG07953	CB936804	8E-27	575	ca	9
1	CAG01395	CB938863	3E-74	635	tc	10
1	CAG08047	CF262362	3E-20	18	ac	14
1	CAG05652	CF971555	2E-31	429	tga	15
1	CAF93405	CK403365	1E-37	403	aata	5
1	CAG01184	CK410393	1E-73	587	gt	14
1	CAG06269	CK411617	2E-55	152	gt	16
1	CAG01390	CK412611	2E-11	108	gatg	5
1	CAG01390	CK412611	2E-11	425	gttt	5
1	CAG08382	CK414367	2E-83	833	ac	14
1	CAG11623	CK414895	2E-29	43	ag	9
1	CAG11623	CK414895	2E-29	61	aaag	5
1	CAG10993	CK415834	9E-17	616	tg	9
1	CAG09843	CK417836	4E-64	575	atg	10
1	CAG10133	CK422325	4E-15	491	at	10
1	CAF99087	CV987962	3E-15	430	agc	6
1	CAG01185	CV988882	3E-63	574	gac	6
1	CAG01185	CV988882	3E-63	595	gac	6

sequencing or overgo hybridization should verify if the genes existing between the mate-paired genes in zebrafish or *Tetraodon* indeed also exist in the same location in catfish. Such an approach has been demonstrated to be highly efficient for comparative mapping (Xu et al. 2006).

ESTs Provide the Basis for Integration of Genetic and Physical Maps

For efficient genome research, linkage maps constructed by genotyping of a resource family or families using polymorphic DNA markers need to be integrated with the physical maps. In most cases these days, particularly with aquaculture species, physical maps are constructed using BAC-based contigs. Mapping of Type I markers derived from EST analysis on both the linkage map and the physical map would effectively integrate the two maps. Although mapping of Type I microsatellites and SNP on linkage maps is straightforward by traditional linkage mapping, mapping of the same set of Type I markers to BACs requires hybridization. Traditionally, isolation of a large number of probes from cDNA clones and hybridization would involve a large amount of work. However, use of overgo probes (Figure 20.4) can significantly reduce the workload for the isolation of purified probe fragments. In this strategy, an oligonucleotide primer can be selected in the coding region of the cDNA, and an antisense primer is

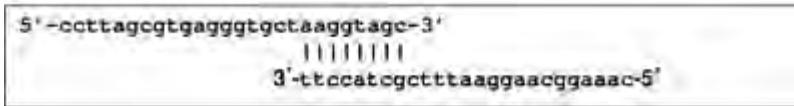


Figure 20.4. The heteroduplex formed between overgo probe primers. This structure serves as template for filling in reactions of polymerase, during which the newly synthesized probes can be labeled. Figure was adopted from Liu (2006).

then designed to partially overlap with the sense primer by 8 base pairs (bp). The primers would form a duplex upon annealing to each other, forming a structure that would be perfectly primed for “filling in” reactions by polymerase. Radioactive or fluorescent nucleotides can be used during “filling in” reactions to label the overgo probes.

ESTs Provide the Basis for the Development of Microarray Technology

Lastly, ESTs provide the material basis for the development of cDNA microarrays. cDNA microarrays are constructed by amplification of the inserts from a pool of unique ESTs. Therefore, in most cases, after EST sequencing, ESTs are analyzed through clustering analysis using various software packages. Once the cluster analysis is completed, the investigator should know how many unique genes the ESTs represent. The inserts of the distinct and unique EST clones are then amplified using PCR, and the PCR products printed on microarray slides. As a note, ESTs are also very important for the verification of gene models for completely sequenced genomes. As we all know, the Human Genome Project was initially announced to be complete in 2000. Politicians such as President Bill Clinton and Prime Minister Tony Blair declared the historical completion of the project at this date, the Human Genome Project was reannounced to be complete in 2002, and once again was announced to be truly completed in 2004. Between 2000 and 2004, the total number of genes included in the human genome was reduced from 40,000 to approximately 25,000. Many wondered what could account for the large reduction of 15,000 genes because the world’s most intelligent and highest caliber scientists were involved in the mapping and assembly of the human genome. These scientists had learned that, without the hard evidence of gene products such as ESTs, it was not possible to have an accurate picture of how many genes humans encode. The 40,000 gene number came from computational predictions in 2000, 15,000 (37.5%) of which were later found to be unsupported by EST evidence and other expression data.

Criteria for Successful EST Projects

The factors one needs to consider for a successful EST project depend on the objectives of the project. However, a set of criteria should be established at the inception of an EST project and used to measure its subsequent level of success. These criteria include a high gene discovery rate, a high gene identification rate, a high capacity for the identification of gene-associated Type I markers, a high rate of full length cDNAs, the ability to differentiate orthologs and paralogs, the ability to identify alternatively processed

transcripts, and an organized inventory so that the ESTs can be used to develop cDNA-based microarrays or be distributed to other researchers upon request.

Before the start of an EST project, the researchers should evaluate the entire purpose of the project in relation to their genome program. For instance, if interspecific hybrid systems are used for linkage mapping, then development of ESTs from both species in parallel may be of interest because many Type I markers developed from the EST project should be directly useful for mapping, as we have done with channel catfish and blue catfish (He et al. 2003).

The number of individuals used for construction of cDNA libraries is important for the identification of SNPs within the cDNAs. For the most part, the majority of aquaculture species are either diploid organisms or tetraploid organisms. Therefore, there should be some level of allelic variation in cDNA sequences. However, including multiple individuals in the library should increase the possibility of SNP identification. In the construction of the catfish cDNA libraries, we used 10 individuals that were required for the collection of sufficient biological samples, but also for the consideration of SNP discoveries.

If EST analysis is a part of broader transcriptome analysis, inclusion of biological samples containing various genetic backgrounds, tissues, developmental stages, and physiological conditions should be considered. For instance, if disease defense genes are of interest, it is best to include tissue samples after infection because some of the defense-related genes may be expressed only after infection. Inclusion of various tissues should allow the capture of various tissue-specific transcripts in sequenced ESTs. Many genes are developmental-stage specific. For instance, genes expressed at early stages of life can only be captured by using embryonic samples at various stages. Similarly, certain genes are expressed only in certain physiological conditions. For instance, genes related to reproductive processes likely are expressed when the animals are sexually mature.

The most important factor for a successful EST project is the quality of cDNA libraries. Of the important criteria for EST projects, most of them are correlated with the quality of cDNA libraries, for example, gene discovery rate, gene identification rate, percentage of known genes with complete open reading frames, etc. The most significant consideration is the quality of starting RNA samples, the production of full-length cDNAs, and efficient normalization.

Several quality standards pertain to EST sequencing or postsequencing analysis. Of these, important factors to consider include average EST length, quality scores of ESTs, usefulness and accessibility of ESTs, timely GenBank submission, and proper level of annotation. EST length and sequence quality is often related to plasmid quality. Standard plasmid kits such as those from Qiagen provide high quality plasmid DNA that allows long, high quality sequencing reads.

After completion of EST sequencing, proper annotation is crucial to make the ESTs useful. If the ESTs are properly annotated, they will be searchable in the public databases. Simply “dumping” the raw EST sequences into public databases makes the EST sequences much less useful not only for other researchers, but also for the “dumpers” themselves. Finally, a good clone inventory is required for successful EST projects because the investigators need to be able to provide any clone upon request. Scientists are excellent in discovery and innovation, but are relatively poor when it comes to inventory. A telling comparison can be made between a scientist’s ability to find proper inventories of EST clones with that of a supermarket to find retail items in the warehouse. It

likely takes far less time for Wal-Mart workers to find any of the tens of thousands of items in their warehouse than for a scientist to find a cDNA clone from his/her freezer. ESTs are valuable long-term resources that require careful inventory practices to ensure their availability in the future. In fact, EST resources often grow in value to the researcher over time. For example, as the genome resources of catfish have grown, ESTs initially used for simple gene discovery and expression analysis are now useful in marker development, comparative mapping, and gene duplication studies.

Acknowledgments

Our catfish genome research is currently supported by a grant from the USDA NRI Animal Genome Tools and Resources Program (Award No. 2006-35616-16685). We are grateful for an equipment grant from the National Research Initiative Competitive Grant No. 2005-35206-15274 from the USDA Cooperative State Research, Education, and Extension Service. I appreciated critical reading of the manuscript by Eric Peatman. The author is grateful to his staff and hard working graduate students.

References

- Adams MD, JM Kelley, JD Gocayne, M Dubnick, MH Polymeropoulos, H Xiao, CR Merrill, A Wu, B Olde, RF Moreno, AR Kerlavage, WR McCombie, and JC Venter. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252, pp. 1651–1656.
- Azam A, J Paul, D Sehgal, J Prasad, S Bhattacharya, and A Bhattacharya. 1996. Identification of novel genes from *Entamoeba histolytica* by expressed sequence tag analysis. *Gene*, 181(1–2), pp. 113–116.
- Bonaldo MF, G Lennon, and MB Soares. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res*, 6(9), pp. 791–806.
- Cao D, A Kocabas, Z Ju, A Karsi, P Li, A Patterson, and Z Liu. 2001. Transcriptome of channel catfish (*Ictalurus punctatus*): initial analysis of genes and expression profiles of the head kidney. *Anim Genet*, 32(4), pp. 169–188.
- Franco GR, MD Adams, MB Soares, AJ Simpson, JC Venter, and SD Pena. 1995. Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. *Gene*, 152(2), pp. 141–147.
- Galau GA, WH Klein, RJ Britten, and EH Davidson. 1977. Significance of rare mRNA sequences in liver. *Arch Biochem Biophys*, 179, pp. 584–599.
- He C, L Chen, M Simmons, P Li, S Kim, and ZL Liu. 2003. Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. *Anim Genet*, 34, pp. 445–448.
- Ju Z, A Karsi, A Kocabas, A Patterson, P Li, D Cao, R Dunham, and Z Liu. 2002. Transcriptome analysis of channel catfish (*Ictalurus punctatus*): genes and expression profile from the brain. *Gene*, 261(2), pp. 373–382.
- Karsi A, D Cao, P Li, A Patterson, A Kocabas, J Feng, Z Ju, KD Mickett, and Z Liu. 2002. Transcriptome analysis of channel catfish (*Ictalurus punctatus*): initial analysis of gene expression and microsatellite-containing cDNAs in the skin. *Gene*, 285(1–2), pp. 157–168.
- Karsi A, P Li, RA Dunham, and ZL Liu. 1998. Transcriptional activities in the pituitaries of channel catfish before and after induced ovulation by injection of carp pituitary extract as revealed by expressed sequence tag analysis. *J Mol Endocrinol*, 21(2), pp. 121–129.

- Kimura T, T Jindo, T Narita, K Naruse, D Kobayashi, T Shin-i, T Kitagawa, T Sakaguchi, H Mitani, A Shima, Y Kohara, and H Takeda. 2004. Large-scale isolation of ESTs from medaka embryos and its application to medaka developmental genetics. *Mech Dev*, 121(7–8), pp. 915–932.
- Kocabas AM, P Li, D Cao, A Karsi, C He, A Patterson, Z Ju, RA Dunham, and Z Liu. 2002. Expression profile of the channel catfish spleen: analysis of genes involved in immune functions. *Mar Biotechnol* (NY), 4, pp. 526–536.
- Kurobe T, M Yasuike, T Kimura, I Hirono, and T Aoki. 2005. Expression profiling of immune-related genes from Japanese flounder *Paralichthys olivaceus* kidney cells using cDNA microarrays. *Dev Comp Immunol*, 29(6), pp. 515–523.
- Lee EK, SB Seo, TH Kim, SK Sung, G An, CH Lee, and YJ Kim. 2000. Analysis of expressed sequence tags of *Porphyra yezoensis*. *Mol Cells*, 10(3), pp. 338–342.
- Liu ZJ. 2006. Transcriptome characterization through the generation and analysis of expressed sequence tags: Factors to consider for a successful EST project. *Israel Journal of Aquaculture-Bamidgeh*, 58, 328–341.
- Lo J, S Lee, M Xu, F Liu, H Ruan, A Eun, Y He, W Ma, W Wang, Z Wen, and J Peng. 2003. 15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Res*, 13, pp. 455–466.
- Serapion J, H Kucuktas, J Feng, and Z Liu. 2004. Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). *Mar Biotechnol* (NY), 6(4), pp. 364–377.
- Shagin DA, DV Rebrikov, VB Kozhemyako, IM Altshuler, AS Shcheglov, PA Zhulidov, EA Bogdanova, DB Staroverov, VA Rasskazov, and S Lukyanov. 2002. A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Res*, 12, pp. 1935–1942.
- Soares MB, MF Bonaldo, P Jelene, L Su, L Lawton, and A Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci*, 91, pp. 9228–9232.
- Song HD, XJ Sun, M Deng, GW Zhang, Y Zhou, XY Wu, Y Sheng, Y Chen, Z Ruan, CL Jiang, HY Fan, LI Zon, JP Kanki, TX Liu, AT Look, and Z Chen. 2004. Hematopoietic gene expression profile in zebrafish kidney marrow. *Proc Natl Acad Sci*, 101, pp. 16240–16245.
- Xu P, S Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and Z Liu. 2006. Generation of channel catfish BAC end sequences for marker development and assessment of syntenic conservation with fish model species. *Anim Genet*, 37, pp. 321–326.
- Zhu YY, EM Machleder, A Chenchik, R Li, and PD Siebert. 2001. Reverse transcriptase template switching: a SMART approach for full length cDNA library construction. *Biotechniques*, 30, pp. 892–897.
- Zhulidov PA, EA Bogdanova, AS Shcheglov, LL Vagner, GL Khaspekov, VB Kozhemyako, MV Matz, E Meleshkevitch, LL Moroz, SA Lukyanov, and DA Shagin. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acid Res*, 32, p. e37.