Expressed Sequence Tags (ESTs)

Generation and Analysis

Edited by

John Parkinson

Molecular Structure and Function, Hospital for Sick Children, Departments of Biochemistry & Molecular Genetics, University of Toronto, Toronto, ON, Canada

Humana Press
Preface

Expressed sequence tags (ESTs) are single-pass reads derived from randomly selected cDNA clones. As such, they provide a highly cost-effective route for the purposes of gene discovery. Over the past couple of decades, we have seen a remarkable rise in their use. To date ESTs have been generated for over 1,500 different eukaryotes with new species datasets being produced almost daily. These rich sources of sequence data provide a wealth of information that has been used to explore issues of eukaryotic diversity. Typically, EST sequencing has been used to identify novel and/or “interesting” genes associated with a species for which full genome sequencing is not currently economically viable; however, ESTs have also proved valuable for purposes of genome annotation and also as alternatives to microarray analyses, for the assessment of relative levels of gene expression.

With the advent of “Next Generation” sequencing technologies such as the Illumina Solexa, ABI SOLID, and 454 Life Sciences platforms, EST-based resources are now becoming readily accessible for labs with even limited budgets. This volume aims to introduce the reader to many of the fundamental concepts underlying the generation and analysis of ESTs. Targeted mainly at life-science researchers (clinician, zoologist, botanist, etc.) interested in incorporating ESTs into their research programs, chapters are provided that give detailed descriptions of the various methods used to generate and analyze EST datasets. While primarily aimed at scientists interested in exploiting EST technology as a means of surveying the genetic diversity of an organism of interest, this book will also be of interest to those researchers who wish to use EST technology for other purposes such as expression profiling, analysis of alternative transcripts, and phylogenomics.

Following a brief overview of ESTs and their previous and current uses (Chapter 1), the first section of this volume (Chapters 2–7) focuses on general strategies surrounding EST projects and provides details on the methods employed to prepare cDNA libraries from a range of organisms (protists and fungi, plants, and animals). Within these chapters, the reader is presented with a number of protocols that outline the construction of a variety of different types of cDNA libraries including libraries that are enriched for full-length cDNA sequences, libraries that are normalized with respect to relative abundance, and libraries that make use of splice-leaders that some eukaryotes employ particularly for splicing of polycistronic units. Chapter 8 then discusses available sequencing resources and provides protocols for the generation of sequence information from these libraries. Chapters 9–11 present state-of-the-art software tools (which can operate on a single desktop workstation) used for processing the raw sequence information generated by the sequencing platforms and extracting biologically meaningful sequence data in the form of ESTs. Protocols are given to remove low-quality sequence or vector contaminants and process the ESTs into a format that can be submitted to the public EST database resource – dbEST. Additional bioinformatic pipelines are described that (1) cluster ESTs into nonredundant sets of related
gene objects; (2) annotate these clusters with functional associations; and (3) store the results within a centralized Web-accessible database resource. **Chapters 12–13** outline statistical methods to analyze these data, both to examine differential expression and also to explore their evolutionary properties. Finally, the last two chapters (**14 and 15**) discuss how ESTs can be used to construct microarray resources and, in more global terms, how they can be usefully applied within the context of human health.

This book assumes that the reader has some knowledge of molecular biology techniques and is also familiar with basic bioinformatics concepts.

*John Parkinson*
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contributors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Expressed Sequence Tags: An Overview</td>
<td>John Parkinson and Mark Blaxter</td>
</tr>
<tr>
<td>2</td>
<td>Strategies for Undertaking Expressed Sequence Tag (EST) Projects</td>
<td>Sandra W. Clifton and Makedonka Mitreva</td>
</tr>
<tr>
<td>5</td>
<td>Preparation of Full-Length cDNA Libraries: Focus on Metazoans.</td>
<td>Masako Harada and Yoshihide Hayashizaki</td>
</tr>
<tr>
<td>6</td>
<td>Expressed Sequence Tags: Normalization and Subtraction of cDNA Libraries.</td>
<td>Marcelo Bento Soares, Maria de Fatima Bonaldo, Jeremiah D. Hackett, and Debashish Bhattacharya</td>
</tr>
<tr>
<td>7</td>
<td>Generating EST Libraries: Trans-Spliced cDNAs.</td>
<td>Cecilia Fernández and Rick M. Maizels</td>
</tr>
<tr>
<td>8</td>
<td>Large-Scale Sequencing and Analytical Processing of ESTs.</td>
<td>Makedonka Mitreva and Elaine R. Mardis</td>
</tr>
<tr>
<td>9</td>
<td>EST Processing: From Trace to Sequence.</td>
<td>Ralf Schmid and Mark Blaxter</td>
</tr>
<tr>
<td>10</td>
<td>Obtaining Accurate Translations from Expressed Sequence Tags.</td>
<td>James Wasmuth and Mark Blaxter</td>
</tr>
<tr>
<td>12</td>
<td>Phylogenomic Analysis of EST Datasets</td>
<td>José M. Peregrín-Alvarez and John Parkinson</td>
</tr>
<tr>
<td>13</td>
<td>Statistical Analysis of Expressed Sequence Tags.</td>
<td>Edward Susko and Andrew J. Roger</td>
</tr>
<tr>
<td>14</td>
<td>Application of ESTs in Microarray Analysis.</td>
<td>Weizhong Li, Lisa Olohan, Daryl Williams, Margaret Hughes, Andrew Gracey, and Andrew Cossins</td>
</tr>
<tr>
<td>15</td>
<td>Exploiting ESTs in Human Health.</td>
<td>Sandro José de Souza</td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Contributors

DEBASHISH BHATTACHARYA • Department of Biological Sciences and Roy J. Carver Center for Comparative Genomics, University of Iowa, Iowa City, IA, USA
MARK BLAXTER • Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
MARIA DE FATIMA BONALDO • Children’s Memorial Research Center, Northwestern University’s Feinberg School of Medicine, Chicago, IL, USA
EMMET O’BRIEN • Robert-Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada
GERTRAUD BURGER • Robert-Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada
PIERO CARNINCI • Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center, Yokohama, Japan; Genome Science Laboratory, RIKEN, Wako, Japan
SANDRA W. CLIFTON • Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
ANDREW COSSINS • Laboratory for Environmental Gene Regulation and Liverpool Microarray Facility, School of Biological Sciences, University of Liverpool, Liverpool, UK
CECILIA FERNÁNDEZ • Facultad de Química, Universidad de la República, Montevideo, Uruguay
LISE FORGET • Robert-Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada
ANDREW GRACEY • Marine Environmental Biology, University of Southern California, Los Angeles, CA, USA
JEREMIAH D. HACKETT • Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, AZ, USA
MASAKO HARADA • Genome Exploration Research Group, Genomic Sciences Center (GSC), RIKEN, Yokohama, Japan
YOSHIHIDE HAYASHIZAKI • Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center, Yokohama, Japan; Genome Science Laboratory, RIKEN, Wako, Japan
MARGARET HUGHES • Laboratory for Environmental Gene Regulation and Liverpool Microarray Facility, School of Biological Sciences, University of Liverpool, Liverpool, UK
ASAKO KAMIYA • Plant Functional Genomics Research Group, RIKEN Genomic Sciences Center, Yokohama, Japan
LIISA KOSKI • Robert-Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada
B. FRANZ LANG • Robert-Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada
WEIZHONG LI • Laboratory for Environmental Gene Regulation and Liverpool Microarray Facility, School of Biological Sciences, University of Liverpool, Liverpool, UK

RICK M. MAIZEELS • Institute of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK

ELAINE R. MARDIS • Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

MAKEDONKA MITREVA • Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

LISA OLOHAN • Laboratory for Environmental Gene Regulation and Liverpool Microarray Facility, School of Biological Sciences, University of Liverpool, Liverpool, UK

JOHN PARKINSON • Molecular Structure and Function, Hospital for Sick Children, Departments of Biochemistry & Molecular Genetics, University of Toronto, Toronto, ON, Canada

JOSÉ M. PEREGRÍN-ALVAREZ • Molecular Structure and Function, Hospital for Sick Children, Toronto, ON, Canada; Department of Molecular Biology and Biochemistry, University of Málaga, Málaga, Spain

NAIARA RODRÍGUEZ-EZPELETA • Robert-Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada

ANDREW J. ROGER • Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

RALF SCHMID • Department of Biochemistry, University of Leicester, Leicester, UK

MOTOAKI SEKI • Plant Functional Genomics Research Group, RIKEN Genomic Sciences Center, Yokohama, Japan; Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, Tsukuba, Japan

YAO-QING SHEN • Robert-Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada

KAZUO SHINOZAKI • CREST, Japan Science and Technology Corporation (JST), Japan; RIKEN Plant Science Center, Yokohama, Japan

MARCELO BENTO SOARES • Children’s Memorial Research Center, Northwestern University’s Feinberg School of Medicine, Chicago, IL, USA

SANDRO JOSÉ DE SOUZA • Ludwig Institute for Cancer Research, São Paulo, Brazil

EDWARD SUSKO • Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

SHONA TEJEIRO • Robert-Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada

JAMES WASMUTH • Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK; Molecular Structure and Function, Hospital for Sick Children, Toronto, ON, Canada

DARYL WILLIAMS • Laboratory for Environmental Gene Regulation and Liverpool Microarray Facility, School of Biological Sciences, University of Liverpool, Liverpool, UK
Expressed sequence tags (ESTs) are single-pass reads of approximately 200–800 base pairs (bp) generated from randomly selected cDNA clones. Since they represent the expressed portion of a genome, ESTs have proven to be extremely useful for purposes of gene identification and verification of gene predictions. They therefore represent a low-cost alternative to full genome sequencing. This is of particular value for eukaryote organisms, which tend to have larger and less gene-dense genomes than do prokaryotes. The use of expressed sequence tags dates back to the early 1980s, when Putney and colleagues sequenced inserts from 178 clones derived from a rabbit muscle cDNA library (1). Several
years later, with the advent of high-throughput DNA sequencing technology, they were proposed as a complementary resource to genome sequencing projects (2). Today ESTs are still being exploited for purposes of novel gene discovery, especially for those organisms for which economies of scale preclude complete genome sequencing. The advent of cheaper and faster sequencing technologies has generated an exponential growth in the number of ESTs that have been generated. At the time of writing, 45,660,524 ESTs from 1408 different species had been deposited in the public EST sequence database, dbEST (3). It is interesting to note that the 25 billion bp contained in these data exceeds that of the entire set of the non-redundant nucleotide database (22 billion bp). Furthermore, whereas over 500 species are associated with at least 1000 ESTs, only 66 eukaryotes have been fully sequenced to date (4), highlighting the impressive breadth and depth of coverage associated with EST datasets. Importantly, the recent release of “next-generation” sequencing platforms that can generate hundreds of thousands to millions of sequencing reads (of 35–250 bp) promises to again revolutionize the speed and depth of transcriptome data acquisition especially for the majority of neglected organisms.

In this volume, you will find protocols giving practical details on generating and processing ESTs. In addition, several chapters provide details of their downstream analysis and how they may be applied within a variety of contexts. The following provides a brief overview of some of the issues and possibilities you may wish to consider in undertaking your own EST sequencing project.

2. Generation of ESTs – Overview

The process of generating ESTs starts with the initial purification of pools of mRNAs from either a whole organism or specific tissues (Fig. 1.1). mRNAs are isolated on the basis of their 3’ poly-A tails (hence their restriction to eukaryotes) and reverse transcribed to create libraries of cDNAs cloned into an appropriate vector. Individual clones from these libraries are then selected (typically at random) and subjected to a single sequencing reaction using universal primers that can be associated with either end of the insert. Downstream of the sequencing reaction, a number of sophisticated bioinformatics pipelines have been developed that process the raw sequence read to remove low-quality sequence information and contaminating vector sequence. The length of the generated read varies according to the sequencing technology used. For example, the latest capillary-based technologies can produce reads of up to 800 base pairs of relatively high quality. The
Fig. 1.1 (continued)
resultant high-quality trimmed sequence is then deposited in the specialized dbEST database (3). At each stage of this process, a number of strategic decisions must be made according to the objectives of the EST project. These include the source and number of clone libraries to generate ESTs, the sequencing technology itself and the software to employ for sequence processing. An in-depth discussion of these decisions is presented in Chapter 2.

2.1. Generation of cDNA Libraries

Due to the random process of selecting clones for sequencing, the distribution of transcripts represented by sequenced ESTs reflects their relative abundance within the clone library. The choice of material and method for library production is dependent, to some extent, on the aims of the project. For example, projects attempting to identify new stress genes in plants would likely use material from plants that have been exposed to a particular stress. Considerations for material include life cycle stage, sex and tissue types. Even for relatively small organisms or cell- or tissue-types within larger organisms, techniques such as laser microdissection microscopy offer avenues for tissue-specific cDNA library generation (5). It should be noted that while continued sequencing from the same individual library may be useful for determining the relative abundance of transcripts, it does lead to diminishing returns in terms of novel gene discovery. Projects focused on gene discovery should therefore consider the generation and sampling of multiple libraries. From our experience, maximizing the cost effectiveness of an EST project is critically dependant on ongoing assessment of the rate of novel sequence discovery per library. Many protocols now exist for the generation of cDNA libraries and several are presented within this volume (Chapters 3–7). The first three of these chapters provide complementary techniques for the generation of libraries enriched in full-length cDNAs. While the application of each protocol is illustrated using specific sources of material (protists/fungi, plants and metazoans), it is important to note that each method may be readily adapted to other organisms. One source of inefficiency in EST projects is the repeated sequencing of highly abundant mRNAs. Consequently, a number of methods have been developed to account for the relative abundance of transcripts, i.e. to

Fig. 1.1. Schematic overview of the generation and processing of expressed sequence tags. (A) Depending upon the size of the organism, mRNA may be collected either from the whole organism or from specific tissues. (B) Pools of mRNA are extracted and purified (typically on the basis of their poly-adenylation). (C) A cDNA library is then constructed from this pool and clones are randomly picked for a single-pass sequencing read (D). The raw trace data are then processed to derive the underlying sequence (E). Further processing identifies and removes low-quality sequence and contaminating sequence associated with the cloning and/or sequencing vectors. The purified sequence data can then be submitted to the international repository of ESTs – dbEST – along with associated metadata detailing the source of ESTs and methods used in their generation. (F) Finally, ESTs may undergo postprocessing, for example, they may be clustered on the basis of sequence similarity to derive groups of sequences that putatively derive from the same gene. Consensus sequences derived from these clusters may be further annotated, e.g. via BLAST searches and used to construct queryable databases such as PartiGeneDB.
“normalize” the relative abundance of each cDNA in the library and thus to optimize the recovery of as diverse a set of sequences as possible. Two example protocols are presented in Chapters 3 and 6. Finally, Chapter 7 outlines a protocol for exploiting trans-splicing which, in addition to enriching for full-length cDNAs, also provides a mechanism for selecting a specific organism’s mRNA from mixed samples – an obvious concern in generating libraries for parasites.

2.2. EST Sequencing and Processing

While outsourcing of sequencing to commercial concerns is certainly a possibility, nowadays many research groups have access to some form of sequencing within their local institute. Over the past 5 years, we have seen significant improvements in DNA sequencing resulting in an increase in both the length and the quality of reads. Two recent breakthroughs in sequencing have been the successful deployment of “massively parallel” platforms based on reversible dye termination sequencing (exemplified by Illumina’s Solexa technology) and pyrosequencing (exemplified by Roche 454 Life Sciences GSFLX technology). Due to the generation of relatively short read lengths (currently limited to ~30 bp), the former is currently unsuited for EST projects; however, attempts are now underway to exploit the 454 platform, which offers an impressive capacity for sequencing. Protocols for generating ESTs from this technology are provided in Chapter 8 along with details of the more conventional capillary technology.

With the increasing volume of sequences now being generated, bioinformatics pipelines are now becoming essential tools for their processing. Critical steps in these pipelines include converting the raw traces to user-readable sequence data, identification and trimming of contaminating vector and low-quality sequence, submission of data to dbEST and the identification of non-redundant sets of genes. This latter step typically involves clustering of sequences sharing highly significant similarity (typically ≥ 95% sequence identity) into groups that putatively derive from the same gene. Exploiting these clusters, “consensus” sequences (also termed “contigs”) can be built, which have greater overall quality and length than the ESTs from which they derive, and thus better model the mRNA. While ideally each gene should be associated with a single contig, in practice sequencing errors, the presence of alternative transcripts and gene family expansions can result in complications in the clustering process. Numerous solutions for EST processing exist, embodying different philosophies of cluster discovery. One of the more robust is PartiGene (6), described in Chapter 9, which provides a readily portable and user-friendly software suite suitable for processing, submitting and clustering hundreds of thousands of ESTs at a time.

Having derived a set of putative genes (which we have referred to in the past as “partial genomes”), annotation pipelines are typically employed to provide the sequences with some sort of
functional annotation. The ability to robustly predict peptide sequences from their cDNAs greatly aids this process. While simple six-frame translations could be adopted, due to possible errors (such as frame shifts and base miscalls) more sophisticated techniques have been developed such as prot4EST (7), which is outlined in Chapter 10. A major goal for most EST projects is to identify functions associated with identified sequences. Several tools are now available for automated functional annotation and usually rely on existing tools such as BLAST. Chapter 11 provides an overview of the AutoFACT pipeline (8), which provides a readily customizable tool for BLAST-based annotation. With the availability of peptide sequences, it is possible to envisage such tools adopting more sophisticated functional annotation algorithms, perhaps based on hidden Markov models or other sequence profiling tools (9).

3. EST Resources

The main repository for EST sequence data is dbEST available through the National Center for Biotechnology Information (NCBI) Web portal (and the Web portals of the other members of the International Nucleotide Sequence Database Consortium). The NCBI database is readily queryable through the Entrez nucleotide or BLAST interfaces. In addition, it is now possible to upload and retrieve the raw trace (or sequencing chromatogram) information associated with individual ESTs. This latter feature allows users to perform their own analyses to detect low-quality or vector-contaminating sequence rather than relying on the submitters to correctly identify such sequence.

A number of databases have been developed that provide access to EST meta-analyses. Such resources take the common form of clustering the ESTs to derive sets of non-redundant gene objects as highlighted above (Section 2.2) and associating these data with additional annotations to provide a community resource that focuses on interesting aspects of the species concerned. The number of EST-based databases has grown considerably over the past few years and range from those featuring only one or two species to those collating hundreds of data sets. It is beyond our scope to list all the EST-based databases, but the following provides a list of the more notable. Nembase (10) and Nemagene (11) are complementary resources collating EST information for almost 40 different species of nematode. The ESTIMA package is a collection of databases featuring a diverse set of organisms. TBestDB is a taxonomically broad eukaryotic database that organizes and
integrates protist cDNA sequences generated by the pan-Canadian Protist EST program (PEP). More global databases include UniGene (12) (which covers 86 species), the TIGR Gene Indices (13) (92 species) and PartiGeneDB (14) (303 species, but soon to be upgraded to 500 species). Typically, these databases feature BLAST and annotation search tools that facilitate the identification of sequences of interest (e.g. “show me the list of gene objects annotated as kinesin”) although many feature additional tools allowing more focused analyses. For example, Nembase features a java tool, SimiTri (15), that allows the user to visualize sequence relationships between one organism with three different organism datasets.

Since EST datasets represent valuable resources that may be exploited beyond their initial purpose, we would encourage all researchers generating new ESTs to deposit sequences and traces to the public repositories. In the following section, we highlight some of the applications of ESTs.

4. Application of ESTs

4.1. Gene Discovery

ESTs have proved to be a cheap and effective tool for gene discovery across a variety of scales. Many small projects have been initiated to provide a first glimpse of genes that an organism may be expressing at critical stages of their life cycle. For example, applied to parasites such as the sheep scab mite Psoroptes ovis (16) or the cestode Echinococcus granulosus (17), the generation of even a few hundred ESTs successfully identified a number of possible antigens that are now the focus of more in-depth functional investigations. ESTs derived from plants such as peanut Arachis hypogaea (18) and salt cress Thellungiella salsuginea (19) have been used to characterize resistance and stress genes, which may be used to accelerate future breeding programs. Finally, ESTs generated for the American oyster, Crassostrea virginica (20), and the earthworm, Lumbricus rubellus (21), have been used to identify genes that could be used as bioindicators of exposure to pollutants and other toxic agents.

At larger scales, projects have also been initiated to derive large collections of ESTs from a variety of related organisms including Apicomplexa (22), Coleoptera (23) and Nematoda (24). Such studies are beginning to yield new insights into sequence conservation within and between taxa and have also highlighted sequences that are likely to be associated with novel taxon-specific innovations. For example in nematodes, over half of the consensus sequences built from 265,000 ESTs were unique to the phylum. Of these, 1262 were highly conserved across all major nematode clades suggesting their potential as targets for broad-spectrum anti-nematicides (24).
While EST projects are typically seen as a low-cost alternative to full genome sequencing, they have also found application as a complementary resource. Full genome sequencing programs now usually include extensive EST sequencing to provide training data for gene-finding algorithms and to provide experimental affirmation of gene models. Many genome browsers provide additional feature tracks indicating a sequence match between an EST and the genome. During gene-finding exercises, such matches provide additional support for confirming the existence of a gene particularly when the only other evidence is a computational prediction (25, 26). Furthermore, the mapping of ESTs onto genomes can greatly aid the identification of correct exon–intron boundaries.

4.2. Sequence Diversity and Phylogenetics

The increasing number of ESTs from a diverse set of organisms deposited in dbEST provides an unrivalled opportunity to explore eukaryotic diversity in a depth not possible using only the few fully sequenced genomes. To this end, several studies have exploited these datasets to address questions of taxonomy and diversity. For example, combining sequences from complete genomes and EST datasets, Philippe and co-workers used 146 genes from 35 species to explore eukaryotic phylogeny and provide strong support for a new animal phylogeny (27). A more recent study examining almost 1.3 million sequences from almost 400 species provided a detailed map of the origins and extent of sequence diversity, highlighting significant differences between prokaryotes and eukaryotes (28). Studies such as these demonstrate the utility of ESTs to derive meaningful phylogenetic relationships. We would therefore encourage researchers, particularly those focusing on species with little or no prior sequence data, to consider performing preliminary phylogenetic analyses that help infer taxonomic relationships. Protocols for such analyses are presented in Chapter 12.

4.3. Transcript Profiling

The study of expression patterns of genes across a range of cell types, life cycle stages, or environmental conditions has proved to be extremely useful for understanding gene functions. Traditionally, such studies are undertaken using high-throughput expression analysis technologies such as microarrays. However, ESTs can also be used to provide a snapshot of gene expression in the material from which they are derived. Providing the cDNA library has not been “normalized” or “subtracted,” the relative number of EST clones assigned to a cluster should correspond to the relative abundance of the corresponding mRNA, and thus the gene’s expression level. Statistical analyses that can be applied to identify genes which are differentially expressed between two libraries are described in Chapter 13. With the availability of
millions of sequences from hundreds of different libraries, these analyses can provide a powerful alternative to microarray-based approaches that is not biased by probe selection and hybridization intensities. Consequently, several studies have been published which make use of the vast amount of EST data available for mouse and human tissues to identify genes associated with tumors (29), neural development (30) and a variety of mouse tissues (31). It is envisioned that the continued generation of additional EST datasets will allow similar explorations across many different species.

Although ESTs can themselves be used for expression analyses, they are more commonly employed in their gene discovery role allowing the design and construction of novel microarray platforms. For example, the generation of almost 27,000 ESTs from the green peach aphid (*Myzus persicae*) led to the creation of a *M. persicae* microarray resource based on 10,000 oligonucleotides (32). This microarray will allow researchers to track adaptations that allow *M. persicae* to interact with their host. Similarly, EST projects have been used to help design microarray resources to investigate stress responses in fish (33) and toxicological responses in earthworms (34). Chapter 14 provides a series of protocols that address the specific challenges associated with designing and generating a microarray platform based on EST sampling.

**4.4. Other Applications**

The continued generation of ESTs for ever-increasing numbers of species and the development of new technologies will undoubtedly lead to the exploitation of ESTs across a range of novel applications. A major technology in the field of proteomics is tandem mass spectrometry, which attempts to match peptide fragments to known protein sequences. However, even for well-studied organisms, the limited number of such sequences in protein databases such as Swiss-Prot results in a computational bias against poorly characterized proteins. Thus, due to their coverage, ESTs are beginning to find widespread appeal and also allow the identification of peptides from alternative spliced isoforms (35). ESTs have been used to identify and characterize alternative splicing events that are otherwise difficult to predict from the raw genome sequence (36, 37). Finally, it is worth mentioning that with the recent rise of highly parallel sequencing, ESTs and EST-linked polymorphisms are beginning to be exploited for purposes of genotyping. Several tools are now available, which attempt to identify single nucleotide polymorphisms (SNPs) on the basis of EST data (38–40). Chapter 15 provides an overview of some of the current applications of ESTs within the context of human disease.
References


Chapter 2

Strategies for Undertaking Expressed Sequence Tag (EST) Projects

Sandra W. Clifton and Makedonka Mitreva

Abstract

Complementary DNA (cDNA) sequencing can be used to sample an organism’s transcriptome, and the generated EST sequences can be used for a variety of purposes. They are especially important for enhancing the utility of a genome sequence or for providing a gene catalog for a genome that has not or will not be sequenced. In planning and executing a cDNA project, several criteria must be considered. One should clearly define the project purpose, including organism tissue(s) choice, whether those tissues should be pooled, ability to acquire adequate amounts of clean and well-preserved tissue, choice of type(s) of library, and construction of a library (or libraries) that is compatible with project goals. In addition, one must possess the skills to construct the library (or libraries), keeping in mind the number of clones that will be necessary to meet the project requirements. If one is inexperienced in cDNA library construction, it might be wise to outsource the library production and/or sequence and analysis to a sequencing center or to a company that specializes in those activities. One should also be aware that new sequencing platforms are being marketed that may offer simpler protocols that can produce cDNA data in a more rapid and economical manner. Of course, the bioinformatics tools will have to be in place to de-convolute and aid in data analysis for these newer technologies. Possible funding sources for these projects include well-justified grant proposals, private funding, and/or collaborators with available funds.

Key words: Complementary DNA (cDNA), expressed sequence tags (ESTs), normalization, subtractive hybridization (subtraction), suppression subtractive hybridization, oligo-capping, SL cloning, sequencing.

1. Introduction

Complementary DNA (cDNA) end sequencing to produce expressed sequence tags (ESTs) is an effective means of sampling an organism’s transcriptome. ESTs can be used for (1) gene discovery in advance of genome sequence, (2) providing data for
2. Organizing an Efficient and a Cost-Effective Conventional cDNA Project

There are many factors to consider when launching an EST project. The first decision must be to clearly define the purpose for carrying out the project. Which or how many of the several uses listed above (or others) does one want to pursue? This information will determine which library preparation and sequencing procedures are to be used. Multipurpose use of the generated data is by far the most cost-effective approach. For example, one might like to plan to sample the transcriptome of an under-studied organism and use the generated ESTs as probes for expression profiling using cDNA-based microarrays or as probes for identifying genes of interest in a BAC or fosmid library.

The techniques described here mainly address conventional cDNA library construction (cloned libraries). It should be noted that cDNA sequencing technology is rapidly evolving and protocols for sequencing cDNAs without cloning on next generation sequencers are coming to the fore (Chapter 8). Nevertheless, many of the suggestions and techniques discussed here will be applicable to both clonal and nonclonal cDNA sequencing.

The success of any EST project depends on the quality and characteristics of the sequenced libraries. Good-quality libraries should have low levels of chimeric and no-insert clones, low levels of “contaminating” sequences (such as mitochondrial, bacterial, rRNA reads, etc), and high levels of directionally cloned inserts. Properties of ESTs that make them invaluable resource include

1. sequence length and accuracy that is adequate for homology searching (BLAST analysis);
2. redundancy that aids in resolution of potential artifacts;
3. cost effectiveness of a large-scale effort;
4. high throughput that allows rapid access to many genes (parallel collection);

5. multiple applications as noted in the Introduction.

However, the shortcomings of EST data should be acknowledged, accepted, and accounted for when undertaking a cDNA project. Since these are single-pass unedited sequence reads, there will be base-calling errors. That is why a certain level of redundancy in the EST reads is desirable. Artifacts, as reverse transcriptase errors and heterogeneous nuclear RNA (hnRNA) contamination will occur. Expression bias (cDNA abundance) could be a technical or biological factor, reflecting the levels of mRNA in the preparation; however, this can be ameliorated, if desired, by eliminating short inserts, as they clone more efficiently, or using techniques reviewed below such as normalization, subtractive hybridization, and/or suppression subtractive hybridization. Finally, the information gleaned from an EST read will be incomplete, as it represents only gene fragments with little or no information regarding “control elements.” In spite of these issues, sequence-based sampling of the transcriptome will give a view, though partial, of the gene expression occurring in an organism at a particular time. This can be invaluable data for interested researchers, especially for under-studied transcriptomes or as an aid to annotating whole shotgun assemblies.

2.1. Quality and Quantity of Biological Material

Keep in mind that a cDNA library is no better than the quality of the starting material. It is imperative to get the tissue from organism to liquid nitrogen as quickly as possible, in order to preserve the original state of the mRNA population in that material. There are commercial products (e.g., Qiagen [Valencia, CA]-RNAEasy®) (Note 1) that claim to stabilize RNA in tissues and bacteria, though we have not had experience with these types of products.

A cDNA library represents the genes expressed at the moment of sampling; therefore, it is essential to choose the right combination of biological material to identify the maximum number of genes per species. For example:

1. If the organism is very small, tissue from the whole organism typically has to be used, and for most, if not all microscopic organisms, populations of individuals will have to be used.

2. Two options can be explored to increase the number of represented mRNA species (Note 2):
   (i) use different developmental stages or growth conditions (1);
   (ii) use different construction protocols (as briefly described in Section 2.2).
Furthermore, researchers have employed an amplification step to obtain an adequate amount of material for cDNA library construction (2), or have taken advantage of the ever evolving technologies to obtain biological material, i.e., using tissue isolated by laser capture dissection microscopy (3) (Note 3).

2.1.1. Using Tissue- or Organ-Specific Material

This presupposes that the organism is large enough to be able to separate tissues for cDNA preparation and will be a good resource for a diverse pool of mRNAs. If monetary restrictions must be considered, and only a few libraries can be prepared and sequenced, then the brain (or equivalent tissues in nonvertebrates such as nerve cord in the roundworm Ascaris), in all likelihood, will give the greatest variety of cDNA material due to the complexity of this tissue. For example, an analysis of brain ESTs from the adult honeybee revealed ESTs representing a broad range of molecular functions and biological processes, with a rich representation of neurobiological gene models (4). In another instance, an EST study of the genes represented in the adult mouse chromosome 21 in the brain found 85% of the identified genes from that chromosome expressed in the brain, with only 21% of those genes represented in muscle tissue (5, 6). A pool of several tissues for a single library also can be used. Other tissues to consider, again depending upon the desired data outcome, are liver, kidney, ovary/testes, and gut. If using gut tissue, ensure that a cDNA library construction protocol is used that will eliminate the presence of nontargeted endosymbiots’ RNA (e.g., using oligo-dT primer will exclude bacterial RNA because it is nonpolyadenylated). In addition, if the tissue is one that undergoes many changes in physiology, such as mammary or gonadal tissue, several libraries from that particular tissue in differing developmental states or locations could be sampled to assess gene expression. This would lead to a larger compilation of unique genes, as genes are switched off or on as differentiation proceeds. Our empirical experience has shown that preparing a different second library from the same tissue as the original library will inevitably yield some unique ESTs not recovered from the first library (see Section 2.2.3 and Note 2).

2.2. cDNA Library Construction Systems

Choose a cDNA library construction method appropriate for the amount of your starting material. Do not hesitate to involve an expert in the field (as described in Section 2.3). If the aim is simply to identify genes and explore the transcriptome using comparative genomics, then conventional libraries without predispositions (as described below) can be used. If one expects to return to the identified genes and characterize them biochemically, then the preferred method is to generate a full-length cDNA library to avoid subsequent 5’ and 3’ rapid amplification of cDNA ends (RACE). Another option is to use cloning and expression vector systems that allow cloning of protein-coding cDNAs into multiple
destination vectors for subsequent protein expression and functional analysis (see Section 2.2.2). These libraries tend to have fewer chimeras and produce higher titers (10).

2.2.1. RNA/mRNA Isolation and cDNA Synthesis

Many protocol variations are in use, and biotechnology companies have their proprietary protocols for library construction (Note 4). The Washington University’s Genome Sequencing Center (WUGSC) has had success with the CloneMiner™ kit provided by Invitrogen (Carlsbad, CA) and the SMART cDNA library construction system (Clontech, Mountain View, CA) (7), both with slight modification of the protocols (Note 1).

Despite the differences among protocols, there are several common steps within the construction process. Cells of interest are harvested from culture or from tissue samples. They can be grown in conditions that elicit the desired gene expression, if known. Total RNA can be extracted by lysing the cells, precipitating and spinning down and discarding the proteins, precipitating the DNA/RNA in the remaining precipitate, treating with DNase, and recovering the total RNA by precipitation. The resulting product can be checked for purity by running a sample on a 0.8% agarose gel or by using an Agilent Technology2100 BioanalyzerRT.

The mRNA can be separated from the tRNA by a variety of methods, including the use of manufacturer’s kits. Originally, the polyadenylated mRNA in eukaryotic cells was recovered from total RNA by passing the total RNA over oligo(dT) or oligo(uT) columns, with subsequent elution from the column. Alternatively, biotechnology companies have kits containing proprietary solutions (as Epicentre Biotechnologies [Madison, WI]-mRNA-ONLY™, or Terminator™ 5‘-Phosphate Dependent Nuclease) that will remove the large tRNAs and leave essentially only mRNA in solution or binding elements (as Clontech [Mountain View, CA]-NucleoSpin®, or Qiagen [Valencia, CA]; Note 1) that will capture the mRNA with poly-A tails. Another choice for either eukaryotic or prokaryotic RNA isolation is the use of random hexamer priming from internal sequences. In either event, the resulting substrate is sometimes LiCl precipitated to further purify. Ideally, these isolations should be performed in a clean room with no residual RNases or DNA exo- and/or endonucleases present.

Once mRNA of sufficient quality and quantity has been isolated, first strand cDNA synthesis is initiated. Again, there are a variety of options available from commercial suppliers, including purchasing individual reagents or obtaining kits and following manufacturers’ directions (Clontech [Mountain View, CA], Epicentre [Madison, WI], Fermentas, Inc. [Burnie, MD], Invitrogen [Carlsbad, CA], Lucigen [Middleton, WI], and Marligen [Ijamsville, MD]) (Note 1). Some of these kits advertise reverse transcription from total RNA, but we suggest isolating mRNA...
first, as more undiluted mRNA substrate will be subjected to the presence of the polymerase. An RNA- and DNA-dependent transcriptase that lacks RNaseH and 3' to 5' exonuclease activity is usually used. Many types are on the market and are supplied in kits, as well. Betaine and/or DMSO can be added to improve the primer/template specificity thereby improving the tenacity of the polymerase, as it proceeds through difficult areas, such as high G/C content, to produce more full-length products (8, 9).

The template RNA can be degraded by base hydrolysis, usually NaOH, and the reaction is then neutralized with acid. The substrate is purified to remove any unincorporated reaction products. DNA polymerase I and dNTPs are added to the single stranded (ss) cDNA, and incubated to yield double stranded (ds) cDNA. Again, the substrate is purified to remove any unincorporated reaction products. The resulting cDNA fragments are size fractionated on an agarose gel or a fractionation column, recovered, and cloned into a suitable cloning vector. This creates a population of clones called a cDNA library.

2.2.2. Cloning and Expression Systems

Our high-throughput sequencing experience has shown that the library quality and yield are enhanced, if the appropriate vector and host are used for library construction. Some vectors that perform well with our high-throughput sequencing methods include pDONRTM222, pSPORT (Invitrogen, Carlsbad, CA), pSMARTR (Lucigen, Middleton, WI), pBluescript (Fermentas, Inc., Burnie, MD), and pDNR-LIB (Clonetech, Mountain View, CA) (Note 1).

The highly efficient recombinational cloning systems usually result in higher numbers of primary clones (those from the first cloning event of the cDNA inserts) compared to the standard cDNA library construction methods using restriction enzymes and ligation cloning techniques as reported in (10). Furthermore, these systems should be considered when planning to do high-throughput proteomics experiments in different vectors, as they enable highly efficient transfer of a large number of protein-coding cDNAs from multiple donor vectors to multiple destination vectors for protein expression and functional analysis. There are systems that use the lambda bacteriophage phage site-specific recombination (reviewed in (11)) sequences (attL, attR, attB, and attP), such as the Invitrogen (Carlsbad, CA) GatewayR technology. The pDONRTM222 vector is used in the Invitrogen (Carlsbad, CA) CloneMinerTM kit. Others, such as the Clontech (Mountain View, CA) CreatorTM system, use bacteriophage P1 site-specific Cre-loxP recombination (12).

Once the cDNA fragments are cloned into the original (entry or donor) vector, they can be shuttled into any vector (destination or acceptor) that contains a corresponding site compatible with the site in the donor (entry) vector. This yields an expression clone with the genes of interest in proper orientation and reading frame.
Transformation of the ligated vector/inserts into host cells is performed next. We have found that the best hosts for our purposes are GC10 cells (Clontech [Mountain View, CA]), G10 Supreme cells (Lucigen [Middleton, WI]), and MegaX cells (Invitrogen [Carlsbad, CA]) (Note 1).

2.2.3. Types of cDNA Libraries

There are several distinct types of cDNA libraries: full-length cDNA sequence libraries, normalized libraries, subtracted and suppressive subtractive hybridization (SSH) libraries. Each of these library construction systems will be discussed briefly in Sections 2.2.3.1–2.2.3.4 and in more detail in other chapters of this book (as indicated).

2.2.3.1. Full-Length cDNA Libraries-5' Capture

The oligo-capping method (13, 14) can be used to increase the capture of the true 5' end of a cDNA. This procedure uses bacterial alkaline phosphatase to remove the phosphates of most truncated mRNA with no 5' cap structure. Tobacco acid pyrophosphatase (Epicentre Biotechnologies [Madison, WI]) (Note 1) is then used to remove the caps from the remaining mRNAs, leaving a phosphate at the 5' end. Next, T4 ligase adds a 5' G-RNA oligo to the 5' end originally having a 5' cap structure. Only a full-length cDNA will have the specific RNA oligo that is recognized by the 5' RACE PCR primer specific to the RNA oligo sequence, yielding amplified cDNA that is full length. The remainder of the cDNA preparation procedure is standard (e.g., see Chapters 3, 4, and 5 for more details). The CapTrapperTM technique allows the capture of full-length cDNA by chemical oxidation of the unique diol groups on the true 5' and 3' termini of the RNA followed by biotinylation of the oxidized diol groups (15, 16). RNase I is used to digest all ss products and remove the biotin group from all mRNA molecules that were not full-length, as incompletely synthesized cDNAs will have some ss areas leading to RNase I cleavage. Streptavidin-coated magnetic beads are used to recover the first strand full-length cDNA/mRNA hybrids, and second strand synthesis is accomplished with thermostable enzymes to produce full-length cDNAs. Another effective method to capture the 5' end of a cDNA is used for organisms that have the ability to attach spliced leader (SL) RNA to the 5' end of some mRNAs. This phenomenon has been reported in only a few metazoan phyla, including Nematoda, Platyhelminthes, Chordata, and Cnidaria (17–26). If the organism studied has this characteristic, it would be well to prepare SL libraries (see Chapter 7), as this will ensure that a full-length cDNA is cloned. For all nonfree living organisms, since an organism-specific primer is used for amplification, the cDNAs cloned are extremely likely to be of that organism’s origin and not a product of host contamination. For example, we at the WUGSC have been able to successfully prepare SL1-PCR libraries from 18 nematode species to date including human, animal, and plant.
parasites (www.nematode.net; (27)). However, not all of the genes in the organism will use this mechanism, so a second conventional cDNA library from the same organism is recommended to capture genes that are not SL spliced. Our experience has shown that generating two distinct libraries per source substantially increased the diversity of transcripts obtained. For example, using serum-stimulated L3 from *Ancylostoma caninum*, libraries produced by the SL1-TOPO versus the modified SMART cDNA library construction system showed only 3.5% overlap (Mitreva, unpublished) (Note 2).

### 2.2.3.2. Normalized cDNA Libraries

Often, cDNA libraries are ‘normalized’ to reduce the frequency of abundant and moderately abundant transcripts (about 50–65% of the cell’s total mRNA) and to increase the frequency of novel transcripts, yielding a more equal distribution of genes. For instance, normalization of a muscle tissue cDNA library would decrease the over-representation of myosin and actin proteins in that library. Normalization (28) is a technique based on differing reassociation kinetics for cDNA molecules in varying relative abundance and should reduce the proportion of abundant sequences an average of 20- to 200-fold. Briefly, in the standard procedure, the ds plasmid cDNA is converted into ss circles using ExoIII, and ds plasmids are removed by hydroxyapatite (HAP) chromatography (28). Avidin-biotin binding (29) and oligo(dT)$_n$-latex beads have been used, as well (30). An aliquot of this purified cDNA is used to PCR amplify inserts, and the amplified cDNA is used to hybridize with the ss circles. The more highly abundant molecules will reassociate more quickly and the ds cDNA is removed by the use of HAP chromatography or other binding agents as described above. The remaining ss cDNA (lower abundance) molecules are converted to ds molecules by primer extension, ready for transformation into competent cells. As a general rule, this would seem to be the obvious choice, but in some circumstances one might not want to normalize. Firstly, normalization is usually technically demanding and labor intensive (28), and normalization can be ineffective for isolating rare transcripts. There are several companies that advertise good products with guaranteed results (as Evrogen [Moscow, Russia], Invitrogen [Carlsbad, CA], Clontech [Mountain View, CA], Marligen [Ijamsville, MD]; Note 1), but the library preparation is generally expensive, and this might be a consideration if cost is paramount. Secondly, the tissue type must be considered. If the tissue contains only one cell type, then normalization is warranted. However, if there are many other cell types within the tissue, normalization might increase the ratio of low abundance to higher abundance mRNAs of all tissue types, giving nonspecific results (see Chapters 3 and 6 for details). However, a simpler method for preparing normalized libraries recently has been described, using Duplex-Specific Nuclease (DSN) from Kamchatka crab. Instead of
using the aforementioned techniques, the DSN is used to degrade
the ds cDNA fractions formed by abundant transcripts followed by
PCR amplification of the remaining ss cDNA molecules (31, 32). Evrogen (Moscow, Russia) supplies the DSN and also prepares
cDNA libraries (Note 4) using this enzyme for normalization.
Clontech (Mountain View, CA) uses DSN normalization for its
custom normalized libraries (Note 4) as well.

2.2.3.3. Subtractive
Hybridization (Subtraction)
cDNA Libraries

A technique to remove genes that are present in one mRNA
preparation (target) from those in another (driver) is subtraction
(30, 33–43). Genes common to both libraries are eliminated by
hybridizing the two cDNA libraries, leaving a sample with an
increased ratio of uniquely expressed genes (see Chapter 6 for
details). This can be used for enriching for differentially expressed
genes in different cell types, tissues, and stages of differentiation in
a single cell type or for assessing gene expression in normal versus
pathogenic tissues. This technique also may be applied to remov-
ing already sequenced clones from a single sampled cDNA library,
allowing for a cost-effective identification of the more rarely
expressed transcripts within that library. As always, the nature of
the starting material will determine the success of this procedure,
which is somewhat difficult and labor intensive (43). Briefly, the
target mRNA can be prepared from either tissue or a cDNA
library. If using a cDNA library, the cDNA will have to be tran-
scribed with T4 polymerase to produce the mRNA. The same
holds true for the subtractive mRNA with the exception that
photoreactive biotin is added to the transcription cocktail if the
cDNA library is used (44–46). The target mRNA then is reverse
transcribed. Exposure of the sample to a sunlamp binds the biotin
to the target mRNA. The resulting target and driver mRNAs are
hybridized at appropriate $C_{ot}$ temperatures to remove the abun-
dant mRNAs and, enriching the subtracted library with unique
mRNAs of interest. More than one hybridization reaction can be
performed at different times and temperatures to increase the
removal of redundant clones. Streptavidin is used to remove the
biotinylated molecules from the sample. Multiple cycles of sub-
traction can be done until the desired decrease in abundant mole-
cules is realized. Some companies that provide this service include
Clontech (Mountain View, CA), Invitrogen (Carlsbad, CA), and
Marligen (Ijamsville, MD) (Note 4).

2.2.3.4. Suppression
Subtractive Hybridization
(SSH) cDNA Libraries

SSH is used when genes are differentially expressed, and cDNA
libraries are desired for these genes (31, 32). This approach will
preferentially amplify target DNA fragments and suppress non tar-
gett amplification. Essentially, the target cDNAs are present in the
tester DNA, which is divided into two samples and different adap-
ters are used to tag each sample. The driver cDNA will have little or
no differentially expressed DNA and an excess of this DNA is
added to the two tester samples and hybridization is allowed to occur. This will elevate the concentration of ss cDNAs in the tester fraction. The nontarget cDNAs will form hybrids with similar DNA fragments from the tester fraction and no longer be single stranded. Next, the two tester samples are allowed to hybridize, and only the single stranded cDNAs will reassociate to form new hybrids. The adapters allow amplification by PCR of the differentially expressed genes. Kits for these techniques are available from biological companies such as Evrogen (Moscow, Russia), Invitrogen (Madison, WI), and Clontech (Mountain View, CA) \((\text{Notes 1 and 4}).\)

2.3. Library Preparation and Sequencing Choices

There are many options for determining who should prepare and sequence cDNA libraries. We shall address several options here, though others certainly exist. If the project is one that the investigator feels should be tightly controlled, then he/she might prefer to prepare the cDNA libraries. We would suggest that one with no experience in this field enter with the full knowledge that there will be a learning curve and that patience is a necessity in this endeavor. Ample time must be planned in the project to perfect the process. It is important to introduce quality checking at each step of the protocol for easier troubleshooting. This also might be the point at which the investigator might want to outsource tissue samples to a commercial operation or to a large sequencing center that can process the tissue, isolate RNA, and prepare cDNA ligation mix. By outsourcing, investigators avoid the learning curve and problems associated with these technologies. In all probability, the cost will be less than if they performed the work in their lab, as the larger centers and companies have the advantage of buying reagents in bulk. In addition, they have optimized library construction procedures and can perform the work in a high-throughput mode. Furthermore, the data should be guaranteed to be of good quality. An alternative cost-saving method is to make cDNA libraries from several tissues separately, tag each cDNA population with a distinct nucleotide combination (e.g., tetramer), pool the tagged cDNAs, and construct one library. Bioinformatics tools can be used to separate transcripts after sequencing. This approach relieves the researcher from generating multiple cDNA libraries, and plating and picking multiple times. However, this requires that the investigator has the necessary tools/experience to perform the bioinformatics analysis, or has access to a bioinformatics group that can do this. Again, investigator preference and funds available for the project are determining factors. Either the company or the sequencing center (\(\text{Note 5}\)) can take the project through to cDNA production, transformation, plating, picking, end sequencing, analysis, annotation, and submission to the dbEST division of GenBank (if so desired). However, the investigator can intervene at any point.
after the sequencing is completed. For instance, the raw data could be transferred to the investigator as ABI or SCF trace files \((see \ Chapter \ 8)\), and all subsequent processing and analysis could be done in his/her laboratory.

EST sequence information is still increasing at a rapid rate, especially with the massive parallel sequencing platforms, as noted above, emerging and being used for cDNA sequencing. The constantly improving cDNA library construction systems coupled with the above will ensure that cDNA sequencing will continue to be of importance for many years to come. It is imperative to realize that a critical component in the characterization and understanding of these sequences is the creation and implementation of bioinformatic approaches that can be applied uniformly across the multiple growing datasets.

Whatever sequencing platform is used to generate the reads, one has to monitor the sampling efficiency. This involves assessing the probability of sampling novel genes when adding more reads to the pool of already sampled transcripts from a single library. To calculate the expected new gene discovery rate with further sampling of a cDNA library, we assemble the ESTs using Phrap ((P. Green, personal communication) and run ESTFreq (W. Gish, personal communication) – \(see \ also \ Chapter \ 13\). The generated report \((Note \ 6)\) provides an expected percent of new gene discovery rate from a second sampling of the same size. When the discovery rate drops below 10–15%, we usually abandon the library, as further sampling is no longer cost effective. However, if the tissue is difficult to obtain, and there is little hope of generating another library, we will sequence deeper into the original library.

2.3.1. Gene Discovery

If the objective is gene discovery, sequencing from the 5' end of the cDNA is probably the most propitious choice. The 5' end is more informative because with the shorter 5' UTR the sequence usually includes a larger part of the coding region. Furthermore, unless the library has been specifically constructed to eliminate a large portion of the 3' poly(A) tails, the polymerase will stutter on the long poly(A) template and give poor or erroneous reads from that end. Most cDNA libraries, even if they are constructed to be "full-length" (or enriched for full-length), often do not contain the true 5' end of the cDNA, as the reverse transcriptase has fallen off the template before reaching the 5' end. However, recent improvements in full-length protocols can improve the percentage of true full inserts within a library \((see \ Section \ 2.2.3.1 \ and \ Chapters \ 3–5 \ and \ 7)\).

2.3.2. ESTs for Sequenced Genomes

If one is collecting EST information for a genome that has been sequenced or is in the process of being sequenced, directional cDNA clones can be used by evidence-based gene calling
algorithms to aid assembly and/or annotation. However, preparation of conventional cDNA clones might be too costly and time consuming. The alternative method here might be to use new sequencing technology, such as the massively parallel pyrosequencer developed by 454 Life Sciences (http://www.454.com/; Chapter 8) to collect cDNA data. They market the Genome Sequencer FLX Titanium (GS FLX Titanium) system, capable of generating up to 400–600 Mb of data as 400 bp reads. No cloning is involved in the sample preparation; therefore, no cloning bias can be introduced. Total cytoplasmic RNA from the organism is collected, mRNA is extracted, cDNA is prepared, randomly sheared and directly sequenced. This technique should yield a preponderance of full-length cDNAs and presents the advantage that libraries can be specifically tagged (usually with a 3-bp identifier) enabling library pooling for a single run. Of course, informatics tools to reassemble the previously sheared transcripts and identify the putative translations from the partial transcripts must be in place. The disadvantage here is that there will be no cloned samples for future work, as the cDNA (not cloned) is simply sheared and run on the sequencer. However, data will exist for primer preparation and performing RT-PCR to recover any area of interest.

3. Scale of Sequencing (Hundreds or Thousands)

The size of the project, again, depends on the goals. If a sample of genes from a large number of similar and small organisms for comparative purposes or generating data for preliminary studies is desired, then a few hundred or a few thousand clones per organism should suffice. Please keep in mind that sampling these numbers of clones will identify only highly abundant transcripts from each mRNA population. If, however, the objective is to garner information for a genome that in all probability will not be sequenced, then as many tissues and/or stages as possible from that organism should be sampled, yielding a maximum number of unique transcripts. If monetary or other considerations restrict the number of cDNAs that can be sequenced, the recommendation is to do a whole organism tissue pool (or a whole organism, if small enough) and prepare a single library. This should yield the most data for the least amount of money. Sequencing costs have diminished greatly in the past few years and will continue to decline as the next-generation sequencing robots come into full production. It is a good practice to poll several companies, sequencing centers,
or university core facilities for cost estimates before project planning begins, as more data than originally thought might be obtained.

4. General Suggestions for Organisms to Use

Organisms chosen for EST projects are dependent upon the area of the investigator’s interest and the availability of organism tissue. The cost of generating draft genomes continuously declines; however, many of the neglected genomes will stay undersampled due to either genome size or inability to be cost-justified, at least for the foreseeable future. Our experience taught us that taxa for which a representative model organism has been sequenced and well studied can allow an extensive extrapolation from that model organism to the related species of interest by comparative studies. For example, as a first multicellular organism to be studied and having a genome in which every base is known, *C. elegans* has served as an excellent platform for comparative genomics. Furthermore, for the nematode project at the WUGSC, a wide and deep approach was undertaken to study the transcriptomes of parasites from the phylum Nematoda. cDNAs were sampled from 38 organisms to yield more than 375,000 ESTs (from ~1,150 ESTs for *Radolphus simulis* to ~31,400 ESTs for *Ascaris suum*; www.nematode.net; (47)). These data coupled with the *C. elegans*-related studies have been used effectively to prepare at least 14 comparative genomic papers (reviewed in (48)).

Regarding the choice of specific genus and species, that again is an investigator choice, and the purpose of the cDNA project must be paramount. For example, frequently animal infections are used as a model for human infections, since subsequent experimental work could not be easily accomplished with humans, but could be done in animals. Therefore, if one is interested only in the transcriptome of the human roundworm *Ascaris lumbricoides*, then the transcriptome of that particular species should be sequenced. However, if the major thrust of the work were a host/parasite relationship, requiring downstream experimentation and validation, the choice should be a related *Ascaris* species that infects animals, such as the swine infective species, *A. suum*. Of course, the investigator should be aware of other projects already being done that would enhance his/her work. If funds were not limiting, the ideal situation, in this instance, would be to sequence the transcriptomes of both *A. lumbricoides* and *A. suum* for comparative purposes, as mentioned above.
An example from the plant world might be an initiative to decipher the transcriptomes of several angiosperms, probing the most basal flowering plants in existence with the goal of reconstructing the ancestral gene set and allowing the understanding of the origins of the angiosperm. The selection for the project could be several phylogenetically critical angiosperms, along with a gymnosperm to aid in determination of the branch point of the angiosperm and gymnosperm lineages. The logical choices would be those that encircle the ancient angiosperm node. This project could yield specific transcriptome data, comparative data, phylogenetic data, and could be a springboard for future sequencing of the angiosperm genome determined to be the most ancient living angiosperm.

The key is to carefully determine what the specific goals are, and then to judiciously explore all viable options to achieve that goal, noting the available organisms, which species will give the desired result, and the availability of biological material.

5. Applying for Funds

The general route for obtaining funding for cDNA projects is by grant proposal submission to various granting agencies. Be sure that the project is relevant to the grant RFA and be able to justify it sufficiently well to convince the reviewers that it is a project worthy of funding. Be certain that you have the expertise to do the project, or can collaborate with others who do. Often, cDNA projects are a part of larger projects and are justified by their use as a resource in subsequent specific aims of the project. High numbers of ESTs are generated when used for annotation of already sequenced genomes, as well as helping to meld assemblies of whole genomes, or when neglected genomes are studied for which genome sequencing is not in the foreseeable future. In the latter instance, one might decide to require multiple approaches to explore the generated data to enhance the probability of funding. Possible avenues are funding by governmental agencies, private donors, and collaboration with other investigators who might have available funds.

6. Conclusions

Complementary DNA projects can yield much important information regarding an organism’s transcriptome, but will not yield the full complement. They will, nonetheless, be significant to
communities that wish to explore the transcriptome of their organism of interest for which no genomic sequence is available. In addition, they will lend information necessary for knitting together genomic assemblies of a sequencing project, as well as aiding in the building of gene models for genome annotation. Careful consideration of project goals, available quantity and quality of materials, and available expertise before the project is begun (or a grant is written) should allow avoidance of major stumbling blocks during the execution of the project. Detailed attention to organism genus and species; conventional (cloning) versus new generation sequencing (no cloning); type of cloning method to use, if cloning is the choice; type(s) of libraries to construct; scope of the project; and funding opportunities all are factors contributing to a successful cDNA sequencing project. The following chapters will allow interested researchers to explore and understand protocols that might be beneficial to investigators anticipating a cDNA sequencing project. Even so, technologies are changing rapidly, and these chapters essentially reveal a “snapshot” of techniques that are available at the time of writing.

7. Notes

1. The described protocols and kits from the various companies are not the recommended, but rather reflect the ones used in this center at the time of writing.

2. Distribution of clusters (EST clustering performed as described in (1)) by library of origin of member ESTs for *Strongyloides stercoralis* and *A. caninum* libraries from different stage material or made by different methods show a low percentage of shared sequences (see Fig. 2.1).

![Fig. 2.1. Overlap of clusters sharing ESTs for two libraries generated for both *Strongyloides stercoralis* and *Ancylostoma caninum*. For *A. caninum*, only 3.5% of clusters contain ESTs from two libraries created from the same parasite material but using different methods.](image)
3. Laser capture microdissection (LCM) enables the procurement of samples of specific cells in a heterogeneous cell population. It uses a modified microscope fitted with a low-power infrared laser. Routinely prepared stained tissue sections on a glass slide and covered with a thin, transparent, membrane of ethylene-vinyl acetate are used for cell population collection. Alternatively, a slide is commercially available (Expression Biology) that uses a proprietary energy transfer coating that replaces the plastic films. The investigator reviews the slide and selects an area to be collected. The laser can be adjusted to collect single cells or clusters of cells. The laser is activated and the cells of interest are bound to the plastic membrane and are excised and ejected into a waiting microfuge cap. Because the plastic or the proprietary energy transfer coating absorbs most of the thermal energy from the laser, the cells are relatively undamaged. The microfuge tube contains the appropriate lysis buffer (usually Trizol for total RNA extraction), and once the tube is capped and inverted, the sample is ready for molecular analysis.

Since the sample is so small, the RNA will have to be amplified before a cDNA library can be produced. A PCR cDNA kit is the usual amplification tool. General cDNA library construction techniques then can be applied.

4. These are not the only companies that provide these services, and the authors intend no endorsement of these companies. These are provided as examples for the reader.

5. Examples of large sequencing centers include Baylor Human Genome Sequencing Center, the Broad Institute, J Craig Venter Institute, Joint Genome Institute, and Washington University Genome Sequencing Center or any university core sequencing center. Companies include, but are not limited to, Clontech (Mountain View, CA), DNAFORM (Tsurami, Yokohama City, Japan), Evrogen (Moscow, Russia), Invitrogen (Carlsbad, CA), Lucigen (Middleton, WI), and Marligen (Ijamsville, MD). These lists are not exhaustive and do not indicate author endorsement. They are meant to be informational initiation points for readers. Colleagues who have done cDNA projects are an excellent resource.

6. **Table 2.1** is an example of the report generated at the GSC to evaluate the redundancy of an EST library. Cluster size indicates the bin number for the ESTs. For example, there are 2,179 unique bins (containing 1 sequence) of the total 6,253 sequenced, and 509 bins containing 2 sequences. The statistics indicate that if one sequenced another 6,253 clones, \( \sim 1,558 \) (or 24.9%) would be unique. This library should be sequenced more deeply.
Table 2.1
GSC generated report showing the distribution of cluster sizes and current redundancy status of the sampled EST library

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,719</td>
</tr>
<tr>
<td>2</td>
<td>509</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>93</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
</tr>
</tbody>
</table>

(continued)
Acknowledgments

This work was supported by the National Human Genome Research Institute (Grant U54 HG003079).

References


genes in scrapie infection. *Nucl Acids Res* 18, 2789–92.


Chapter 3

Construction of cDNA Libraries: Focus on Protists and Fungi

Naiara Rodríguez-Ezpeleta, Shona Teijeiro, Lise Forget, Gertraud Burger, and B. Franz Lang

Abstract

Sequencing of cDNA libraries is an efficient and inexpensive approach to analyze the protein-coding portion of a genome. It is frequently used for surveying the genomes of poorly studied eukaryotes, and is particularly useful for species that are not easily amenable to genome sequencing, because they are nonaxenic and/or difficult to cultivate. In this chapter, we describe protocols that have been applied successfully to construct and normalize a variety of cDNA libraries from many different species of free-living protists and fungi, and that require only small quantities of cell material.

Key words: EST, RNA purification, reverse transcriptase, template-switching, normalization, DSN.

1. Introduction

Sequencing of cDNA libraries has been extensively used to determine the expressed portion of protein-coding genes (Expressed Sequence Tags; ESTs) in model eukaryotes. It has also gained importance for eukaryotic genome projects, as the precise inference of exon-intron boundaries relies on substantial training sets of EST data from the corresponding organisms. In addition, EST sequencing can now also be applied to poorly growing organisms such as protists (for most part unicellular eukaryotes that are neither animals, fungi, nor plants) and fungi (e.g., (1–8)). In such cases, limited quantities of cell material and small amounts of RNA can be overcome by PCR-based amplification. There are two major advantages of the EST approach compared to genomics. First, bacterial contamination of eukaryotic cultures can be tolerated, as polyA tails of bacterial transcripts are too short to be primed by standard oligo-dT
primers. In fact, it is even feasible to use total RNA instead of purified mRNA for cDNA library construction, a decisive advantage in the case of nonaxenic protists or fungi. The second advantage is that some of these organisms require mechanical methods for cell breakage, because their rigid cell walls resist digestion with commercially available cell wall lysing enzymes. While genomic DNA may become too fragmented through such treatment to be useful for genome sequencing, the significantly smaller mRNAs remain sufficiently intact.

In this chapter, we describe a fast and relatively simple method to construct cDNA libraries from protists and fungi. The protocol presented here describes the RNA extraction and purification, and the cDNA synthesis, amplification, size fractionation, and normalization steps. An overview of the procedure is given in Fig. 3.1A. The protocol requires small quantities of cell material, works with both total RNA and purified (polyA)

---

Fig. 3.1. Overview of procedures. (A) cDNA library construction. Black boxes, obligatory steps; grey boxes, optional steps. Black arrows point to protocols described in detail. ‘Purification of total RNA’ stands for elimination of genomic DNA. A black asterisk indicates where a normalization step is usually introduced; a grey asterisk indicates where alternative normalization steps can be introduced. (B) First strand cDNA synthesis and amplification. First, an oligo-dT containing oligonucleotide is used to prime first strand cDNA synthesis, catalyzed by a RNase H activity-deficient MMLV reverse transcriptase (RT). When the capped 5' end of the mRNA is reached, the RT adds 2 to 5 C residues to the first-strand cDNA, permitting that the SMART IV primer anneals and that DNA synthesis continues until the end of the oligonucleotide. To our experience, noncapped 5' ends undergo the same reaction, albeit at reduced efficiency. Finally, the oligo-dT and SMART IV primers serve for PCR amplification of double-stranded cDNA.
mRNA, and enriches full-length cDNAs. Note, however, that the described protocols involve a PCR amplification step, which is prone to artifacts such as unequal amplification of cDNAs, with a tendency to more efficiently amplify shorter molecules. This potential problem is less relevant in exploratory EST projects. However, when sufficient cell material can be obtained, and when avoiding artifacts is a prime issue, we advise employing cDNA protocols without PCR amplification steps and including prior mRNA purification. In such instances, the readers may follow our procedure from cell culture to mRNA purification, and then continue with one of the protocols described elsewhere in this book.

In our standard procedure, first strand cDNA synthesis and amplification are performed using the Creator™ SMART™ cDNA library construction techniques (BD Biosciences, Palo Alto, CA), essentially following the manufacturer’s recommendations. The method is based on synthesis of the first cDNA strand with an anchored oligo-dT primer. The terminal C-addition and template switching features of the particular reverse transcriptase allow the second primer to anneal to and extend the RNA template and thus synthesize the first cDNA strand until the end of the primer (Fig. 3.1B) (9–11). Subsequently, the product is amplified by PCR using the same primers as before, cut by Sfi I (sites are introduced by the primers), and the asymmetrical restriction sites at both cDNA ends are used for directional cloning. A common problem in first-strand cDNA synthesis is premature termination of reverse transcription at mRNA secondary structures. The SMART technique enriches full-length cDNAs because template switching to the 5’ primer occurs preferentially when the reverse transcriptase has reached the (usually capped) 5’ end of the mRNA. Yet, partial cDNAs will also be produced although less efficiently. In fact, the presence of a certain proportion of 5’ end-truncated cDNAs in the library is desirable, as it allows sequences to start at various points of coding regions, thus reducing the requirement for sequencing long cDNA inserts by primer walking.

Since the abundance of the various transcripts in a cell may vary by a factor of a thousand (12), random EST sequencing typically becomes inefficient after a few thousand readings. Therefore, the detection of weakly expressed genes requires normalization of the cDNA libraries, for which we use a simple and efficient procedure involving enzymatic degradation of double stranded DNA or DNA-RNA hybrids (13, 14). We have employed this procedure successfully to construct libraries for protists and fungi. Moreover, except for the culturing and cell lysis steps, these protocols can be applied to animals and plants.
2. Materials

Enzymes, buffers, and reagents such as BSA, DTT, dNTP, and ATP should be stored frozen at -20°C.

2.1. RNA Purification

All solutions must be prepared with RNase-free water, using RNase-free chemicals, glassware and plasticware. Gloves should be worn during manipulation of samples (see Note 1).

1. *Trizol®* reagent (Invitrogen), or preferentially a homemade substitute (see Note 2): 38% stabilized phenol (to stabilize, add 1 mg of hydroxyl chinoline, 2 μL of mercaptoethanol and 0.5 mL of HPLC water per g of phenol, and incubate at 37°C shaking), 0.8 M guanidine thiocyanate, 0.4 M ammonium thiocyanate, 0.1 M sodium acetate, 5% glycerol. Note that *Trizol* contains phenol, which causes heavy skin burns and is toxic on contact or by inhalation of vapors. It should therefore be manipulated under a fume hood, using gloves. According to the manufacturer, commercial *Trizol* is stable at 4°C for at least 9 months (but see Note 2).

2. Mixture (1:1) of 150–212 and 425–600 micron-sized glass beads (Sigma); required for species with a tough cell wall (e.g., most fungi, jakobid flagellates, glaucophytes, red and green algae).

3. Chloroform (add a pinch of bicarbonate for stabilization).

4. Isopropanol.

5. ‘Wash ethanol’: Ethanol (75%).


7. Oligo-dT cellulose (Amersham Biosciences).

8. ‘1X binding buffer’ and ‘2X binding buffer’. Composition of 2X: 20 mM Tris-HCl, pH 7.5, 2 mM EDTA, pH 8, 0.1% SDS, 1 M NaCl. This buffer precipitates at room temperature; heat in a water bath at 65°C before use.

9. ‘Wash buffer’ (1X): 1 mM Tris-HCl, pH 7.5, 1 mM EDTA, pH 8, 0.05% SDS, 0.2 M NaCl. Store at room temperature or heat before use to dissolve precipitated SDS.

10. ‘Elution buffer’ (1X): 1 mM Tris, pH 7.5, 1 mM EDTA, pH 8, 0.05% SDS. Store at room temperature or heat before use to dissolve precipitated SDS.


12. ‘Ethanol-AmAc’: 95% ethanol, 0.5 M ammonium acetate.
2.2. First Strand cDNA Synthesis and PCR Amplification

1. ‘SMART IV primer’, (10 mM) (Clontech):
   5’-AAGCAGTGGTATCAACGCAGACTGGCCATTACG GCCGGG-3’.
2. ‘oligo dT-primer’: CDS III/3’ (anchored) PCR primer (10 mM) (Clontech):
   5’-ATTCTAGAGGCGAGGCGGCGACATG-d(T)_{30} N_{1}N-3’
   (N_{1}=A, G, or C; N=A, G, C, or T).
3. ‘RT-buffer’: First strand buffer (5X) (Clontech): 250 mM Tris-HCl pH 8.3, 30 mM MgCl₂, 375 mM KCl.
5. DTT (20 mM).
6. dNTP mix (10 mM).
7. RNase I (100 μM).
8. ‘10X PCR buffer’, Advantage™-2 (Clontech): 400 mM Tricine-KOH pH 8.7, 150 mM KOAc, 35 mM Mg(OAc)₂, 37 μg/mL BSA, 0.05% Tween-20, 0.05% Nonidet-P40.
9. ‘5’ PCR primer’ (10 mM) (Clontech):
   5’-AAGCAGTGGTATCAACGCAGAGT-3’.
11. ‘Proteinase K solution’, (10 mg/mL) (Boehringer).
12. ‘Gel extraction kit’, QIAquick (QIAGEN).

2.3. Normalization

1. DSN enzyme (Evrogen JSC) diluted according to the manufacturer instructions to 1 Kunitz unit/μL.
2. ‘DSN storage buffer’, (Evrogen JSC).
3. ‘5X hybridization buffer’: 0.25 M HEPES pH 7.5, 2.5 M NaCl, 1 mM EDTA. This buffer may precipitate; store at room temperature for 20 min or incubate at 37°C for about 10 min before use.
4. ‘DSN buffer’ for enzyme reaction; 2X composition: 100 mM Tris-HCl pH 8.0, 10 mM MgCl₂, 2 mM DTT.
5. EDTA (5 mM).

2.4. Restriction

2. ‘Sfi I restriction buffer’ (10X) (Clontech).
3. BSA (10 mg/mL).
4. EDTA (0.5 M, pH 8).
5. ‘Gel extraction kit’, QIAquick (QIAGEN).
2.5. DNA Fractionation
1. ‘Low melting agarose’, SeaPlaqueGTG™ ultra-pure (Mandel).
2. Formamide (highest quality, Pharmacia; stored under nitrogen or argon).
3. Electroelution chamber (see Note 3).
4. ‘TAE buffer’ (1X): 40 mM Tris-acetate, 20 mM sodium acetate, 1 mM EDTA, pH 8.0.

2.6. Cloning
1. ‘Ligation buffer’ (10X): 200 mM Tris-HCl pH 7.6, 50 mM MgCl₂, 50 mM DTT.
2. ‘pDNRlib vector’, cut by Sfi I and purified (see Note 4).
3. ATP (10 mM).
4. T4 DNA ligase (5U/μL).
5. Competent cells (DH5α), and LB agar plates containing 10 mg/mL chloramphenicol and 4 μg/mL tetracycline.

3. Methods

3.1. Cell Culture
A large variety of protists require live bacteria as food. To minimize potential RNA degradation by bacterial enzymes, it is important to keep the ratio of eukaryotic versus bacterial cells as high as possible. Bacteria can be partially removed from protist cultures through differential centrifugation, but according to our experience, this is not necessary because cultures in the late logarithmic or stationary phase of growth, which we generally use for RNA extraction, contain only few bacteria. Note that sampling from only one growth condition may restrict the effective number of cDNA sequences in a library. Ideally, cells grown under different conditions should be combined, but this is more difficult if not impossible for bacterivorous protists and for parasites.

3.2. RNA Purification
For total RNA extraction, we use a modified Trizol protocol.
1. Collect the cells by centrifugation (speed and time vary from one species to another – see Note 5), or by straining through a fine-mesh nylon coffee filter in case of filamentous fungi.
2. Remove supernatant completely, resuspend cell pellet in Trizol and mix; 1 mL of Trizol per 5 to 10 × 10⁶ cells is recommended. For filamentous fungi or species with rigid cell walls (e.g., algae), add glass beads (see Note 6) and shake by hand in a glass bottle (15). Once the cells are broken to >50% (check under light microscope), remove glass beads by repeated rinsing with small volumes of Trizol. The cells in Trizol may be stored at −80°C for at least one month.
3. Leave the cell/Trizol solution for 5 min at room temperature, then add 0.2 mL chloroform per mL of Trizol, shake vigorously for 15 sec, and let the mixture settle at room temperature for 2 to 15 min. From here on use 40 mL plastic centrifuge tubes.

4. Centrifuge at 12,000g at 4°C for 15 min, and collect the colourless aqueous top phase (about 60% of the total volume). Avoid material from the interface (if this occurs, repeat centrifugation).

5. Add 0.3 mL of isopropanol per mL of the collected aqueous phase to precipitate RNA, mix by inversion and leave 5–10 min at room temperature.

6. Centrifuge at 12,000g at 4°C for 15 min and remove the supernatant carefully. The RNA appears as a gel-like or white pellet at the side and bottom of the tube.

7. Wash by adding wash ethanol (1 mL per mL of collected phase) and by inverting the tube a few times. Centrifuge at 12,000g at 4°C for 10 min and discard supernatant. Repeat the procedure twice. Remove all traces of ethanol by using a Pasteur pipette, and air-dry briefly. The RNA pellet can be stored at –20°C for at least one year.

8. Dissolve the RNA in RNase-free water. The volume of added water will vary with the quantity of recuperated RNA. It is best to start suspending in a small volume and to continue adding until the RNA is perfectly dissolved (final RNA concentration ~1 mg/mL). From here on use 1.5 mL Eppendorf tubes.

9. Determine the quality and quantity of the RNA by agarose gel electrophoresis, together with an RNA marker of known size and concentration. Figure 3.2 (lane 1) shows a typical, high-quality RNA extracted by the described method.

10. Remove DNA from RNA preparation by purification on a MiniKit column (see Note 7). Figure 3.2 (lanes 1 and 2) shows the same material before and after this step.

### 3.2.2. Messenger RNA Purification

This step is performed using oligo-dT cellulose columns. The oligo-dT primers base-pair with the polyA tail of mRNAs, whereas non-polyadenylated RNAs will not bind and flow through the column. The protocol explained here applies to ~1 mg of total RNA; the quantity of reagents should be adapted to the actual RNA quantity. Unless otherwise specified, all centrifugations are performed in a microcentrifuge (for Eppendorf tubes) at 14,000 rpm (maximum speed) for 30 sec at room temperature. A correct pH of the solutions (7.5) is critical for high mRNA yield.

1. Fill 40 mg of oligo-dT cellulose into an Eppendorf tube. Wash the cellulose by mixing it with 600 µL elution buffer, centrifugation, and removal of the supernatant. Repeat this step another two times.
2. Equilibrate the cellulose by mixing it with 600 μL of binding buffer, centrifugation, and removal of the supernatant. Repeat another two times.

3. Adjust the RNA solution to 600 μL with RNase-free water, and heat it at 65°C for 4 min.

4. Add 600 μL of preheated (65°C) 2X binding buffer and incubate at room temperature for 15 min; constantly invert the tube.

5. Centrifuge briefly and discard supernatant.

6. Wash cellulose twice with 1X binding buffer and twice with wash buffer.

7. Add 250 μL elution buffer to cellulose, mix gently, and incubate at 37°C for 5 min.

8. Centrifuge for 1 min and transfer the supernatant to a new tube. Add another 250 μL of elution buffer to cellulose pellet, mix gently, and incubate for 5 min at 37°C.

9. Centrifuge for 1 min and combine the supernatants of Step 8 and 9.

Fig. 3.2. Agarose gel separation of various stages of a typical RNA purification experiment. 1.5 μL of total RNA were loaded before (lane 1) and after DNA removal (lane 2; Step 3.2.1.10 in protocol). PolyA mRNA was then purified from 400 μL of total RNA and recuperated in 15 μL. 3 μL of this mRNA fraction were loaded on the gel (lane 3, Step 3.2.2 in protocol). Note a carry-over of rRNAs into the mRNA fraction, which is however negligible as the amount of loaded mRNA corresponds to ~50 times more than that of the total RNA. Lane M; RNA ladder, High Range (Fermentas).
10. Using the recovered 500 µl of RNA, repeat the purification cycle at Step 3.

11. After two cycles of mRNA purification (Steps 3–10), add 40 µl of NaCl (4 M) and 1 mL of Ethanol-AmAc to the RNA and let precipitate at −20°C for 1 h, or over night.

12. Spin down at 14,000 rpm at 4°C for 20 to 30 min, and discard the supernatant.

13. Add 150 µl wash ethanol and centrifuge at 14,000 rpm at 4°C for 10 min.

14. Discard supernatant. Carefully remove ethanol, air dry pellet, and resuspend it in RNase-free water. Figure 3.2 (lane 3) shows the result of a typical mRNA purification.

### 3.3. First Strand cDNA Synthesis

1. Mix 1–3 µL of RNA solution (~25–500 ng polyA RNA, or 100–1,000 ng total RNA), 1 µL SMART IV primer, 1 µL oligo dT primer, and adjust volume to 5 µL with HPLC water. Mix well by pipetting up and down.

2. Incubate in a heat block (or PCR machine) at 72°C for 2 min; immediately chill on ice for 2 min; spin down briefly to collect droplets.

3. Add 2 µL 5X RT buffer and 1 µL reverse transcriptase; mix and spin briefly.

4. Incubate in a heat block (or a PCR machine) at 42°C for 1 h.

5. After 1 h, heat again at 72°C for 2 min, chill on ice for 2 min, spin briefly, add 1µL reverse transcriptase, and incubate in a heat block at 42°C for 1 h (see Note 8).

6. Chill on ice. At this point, the first strand cDNA synthesis step is completed.

### 3.4. cDNA Amplification

1. Remove RNA remaining from the first strand cDNA synthesis by adding 0.1 µL of RNase I 100 µM; leave at room temperature for 10 min (see Note 9).

2. Combine 2 µL first strand cDNA, 80 µL HPLC grade water, 10 µL 10X PCR buffer, 2 µL 5' PCR primer, 2 µL oligo-dT primer, and 2 µL Polymerase Mix. Mix well and PCR-amplify under the following conditions: 20 sec at 95°C followed by ~20 cycles: 10 sec at 95°C and 6 min at 68°C. The number of cycles depends on the amount of RNA starting material (see Note 10).

3. To inactivate the polymerase in the PCR reaction, add 1 µL of Proteinase K solution. Incubate at 45°C for 20 min, followed by 10 min at 65°C.

4. Purify the amplified cDNA with the Gel Extraction kit, and elute the double-stranded cDNA in a final volume of 30 µL (see Notes 11 and 12).
3.5. cDNA Normalization

If normalization is not desired, go to Step 3.6. Successful normalization requires an optimized concentration of the DNS enzyme for a given cDNA library. It is recommended to use different DNS concentrations and test the efficiency of normalization.

1. Prepare three or more dilutions of the original enzyme (1/4, 1/8, 1/16 ...) using DNS storage buffer.

2. For each sample to be normalized, mix 4 μL of amplified cDNA (~500 ng) with 1 μL of 5X hybridization buffer; heat at 98°C for 3 min and at 70°C for 4 h.

3. While keeping the samples at 70°C, add 4 μL of preheated (70°C) 2X DNS buffer and 1 μL of DNS enzyme (for the dilutions mentioned above, this makes 0.25, 0.125, and 0.0625 Kunitz units); incubate at 70°C for 20 min.

4. Inactivate DNS enzyme by adding 10 μL of 5 mM EDTA.

5. Reamplify the DNS-digested cDNA (as in Section 3.4, Steps 2 to 4), but elute in 50 μL final volume.

6. Verify the success of the normalization of each of your samples by gel electrophoresis, and choose that with the desired, even-size distribution for the following step (see Note 13).

3.6. Restriction

1. Mix 45 μL of amplified and purified cDNA, 40 μL HPLC-grade water, 10 μL Sfi I 10X restriction buffer, 1 μL BSA solution, 4 μL (= 80 U) Sfi I restriction enzyme, and incubate for 2 h at 50°C.

2. Stop reaction by addition of 1 μL 0.5 M EDTA.

3. Purify DNA with the Gel Extraction kit by eluting with a final volume of 30 μL (see Note 12).

3.7. cDNA Sizing

1. Depending on the capacity of the wells of the electrophoresis system used, the volume of the sample may have to be reduced from 30 μL to a smaller volume by evaporation in a speed-vac (see Note 12).

2. Add formamide to a final concentration of 10% and loading buffer to the cDNA, incubate for 10 min at 50°C, then chill on ice. This step will reduce aggregation of DNA.

3. Load the sample on a low-melting agarose gel (1.2%; TAE buffer) together with a size marker; start by migrating slowly (1.5 V/cm) for a few minutes, then increase to ~3 V/cm. Migrating at higher voltages may overheat and deform the gel matrix. Excise agarose blocks containing DNA fragments of desired size (e.g., 0.5 to 1 kbp; 1 to 5 kbp).

4. Electroelute each cDNA fraction (see Note 14), and check their yield and size distribution by agarose gel electrophoresis loading about 1/10 of the recuperated material (see Note 15).
3.8. Cloning

Small inserts clone more efficiently. Therefore, to avoid a bias toward small inserts in cDNA libraries, cDNAs should be size-fractioned and fractions cloned separately.

3.8.1. Ligation

The molecule-ratio of insert to vector should be ~3:1 (the vector size is 4 kbp). When calculating the size of the inserts, one needs to consider that the fragment sizes are not necessarily uniformly distributed. For example, in the 1 to 5 kbp fraction, there can be overrepresentation of fragments from 1 to 1.5 kbp. The size distribution should be assessed based on agarose gel migration and the mean size of the fragments estimated accordingly. The final DNA concentration of insert and vector DNA combined should be 5 ng/μL, and the total volume of the reaction should be 3 to 4 μL. See Table 3.1 for two examples of ligation recipes.

1. For each fraction to be cloned, mix well adequate quantities of insert and vector DNA.
2. Heat mix at 50°C for 10 min, place on ice for a few minutes, and let stand at room temperature for several minutes.
3. Add ligation buffer, ATP (final concentration 0.5–1 mM), HPLC grade water, and 0.5 U of T4 DNA ligase.
4. Place the ligation mix in a 14°C incubator overnight.

3.8.2. Transformation

The above-described ligation mix can be used for 10 transformations (200 μL of competent cells per transformation). Competent cells are prepared and transformation is conducted by standard procedures. Transformed cells should be plated onto chloramphenicol-containing agar plates (about 10 plates per transformation).

Table 3.1

<table>
<thead>
<tr>
<th>Examples for ligation reactions</th>
<th>0.5 to 1 kbp (mean: 750 bp)</th>
<th>1 to 5 kbp (mean: 1.5 kbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insert 1</td>
<td>0.5 μL (10 ng/μL)</td>
<td>0.5 μL (20 ng/μL)</td>
</tr>
<tr>
<td>Vector (10 ng/μL)</td>
<td>1 μL</td>
<td>1 μL</td>
</tr>
<tr>
<td>Ligation Buffer (10X)</td>
<td>0.3 μL</td>
<td>0.4 μL</td>
</tr>
<tr>
<td>ATP (10 mM)</td>
<td>0.27 μL</td>
<td>0.36 μL</td>
</tr>
<tr>
<td>Water</td>
<td>0.83 μL</td>
<td>1.64 μL</td>
</tr>
<tr>
<td>T4 DNA ligase (5 U/μL)</td>
<td>0.1 μL</td>
<td>0.1 μL</td>
</tr>
<tr>
<td>FINAL VOLUME</td>
<td>3 μL</td>
<td>4 μL</td>
</tr>
</tbody>
</table>

1Numbers between parentheses correspond to the initial concentration of insert.
On average, 1,000 to 2,000 colonies are expected for each transformation with ~20 ng ligation mix (when using high-quality competent cells).

4. Notes

1. A major difficulty in handling RNA is the prevention of degradation by contaminantRNases. Autoclaving glassware, tips, tubes, and solutions is often insufficient to inactivate RNases. For additional measures, glassware may be baked at 180°C overnight, and plasticware, tubes, and solutions be treated with diethylpyrocarbonate (DEPC). DEPC reacts with histidine residues of proteins and thus inactivates RNases. Add DEPC to solutions (water, buffers) at a final concentration of 0.05–0.1%, incubate for several hours, and autoclave at least 45 min (the characteristic DEPC scent should disappear). Note that DEPC also reacts with RNA; therefore, it has to be completely removed from all materials before use. Moreover, DEPC can react with chemicals containing primary amine groups, such as Tris. Therefore, these chemicals should be added to the solution only once DEPC is removed. DEPC is a suspected carcinogen; take appropriate precautions when handling it (e.g., always wear gloves and handle it under a fume hood). Water purified by a well-maintained MilliQ system is virtually RNase-free, without further treatment. To verify if MilliQ water is indeed RNase-free, dissolve high-quality RNA in this water, incubate it at 37°C for several hours, and compare the RNA before and after incubation by gel electrophoresis.

2. Highest quality RNA (and high molecular weight DNA) is regularly obtained with home-made Trizol, but not with commercial Trizol sources. An apparent reason is that the recipe for home-made Trizol calls for phenol of highest purity (supplied in light-protected glass bottles under a protective gas), and that it is protected from oxidation by additives. Advanced phenol oxidation, which is recognizable by a reddish-pink color, causes slight RNA but severe DNA degradation. Stabilized phenol or Trizol prepared by our recipe may be stored frozen at −20°C for many years. Once in use, we recommend to keep it for <1 month at 4°C, in light-protected bottles (brown glass or unstained glass wrapped with aluminum foil).

3. There are numerous techniques for electroeluting DNA from agarose, and diverse devices are commercially available that will not be described here. Relevant information specific to each
technique is easily available. The least complicated procedure is electroelution in a closed dialysis tube. Electroelution chambers are available from Schleicher and Schüll (Elutrap), Millipore (Centrilutor), EMD BioSciences (D-tube electroelution), and RPI Research Products (GeneCapsule), to mention only some of the more popular devices.

4. We have noted on several occasions that the 220-bp stuffer fragment is not removed from the commercially distributed, ‘ready-to-use’ pDNRLib cloning vector. For highest cloning efficiency, the SfiI-digested vector should be purified by electrophoresis on low-melting agarose, followed by electroelution of the 4.2 kbp fragment.

5. The centrifugation conditions for pelleting cells depend on multiple factors that are specific to each culture (density of the medium, type of cells, etc). Small cell pellets, in particular those of small and/or flagellated eukaryotes, tend to dissolve quickly and have to be decanted immediately after centrifugation, under close visual control.

6. A number of protists and fungi contain a rigid cell wall that is not (or only for a small fraction of cells) dissolved by Trizol. In such cases, cells have to be broken mechanically. We recommend cell disintegration in the presence of Trizol, as RNA will otherwise be degraded by intra-cellular (and if present, bacterial) RNases. Cells of filamentous fungi are broken by grinding together with sand or glass beads in a mortar; cells of unicellular organisms may be disintegrated by manual shaking together with glass beads in a glass bottle (e.g., (15)), or by other suitable disruption methods. Because the volume increases by the addition of glass beads, more Trizol has to be used in this case (we use ~ 10 mL of Trizol and 10 mL of glass beads for 1 g of cells). Once >50% of cells are broken (check by microscopy), decant the glass beads and collect the supernatant. Repeatedly (2 to 4 times) rinse the glass beads with small volumes of Trizol to collect a maximum of the cell lysate.

7. Total RNA extractions contain variable amounts of genomic DNA (depending on the organism and the extraction conditions), which should be eliminated to avoid undesirable PCR products.

8. The reverse transcriptase reaction is repeated once to increase cDNA length. By heating to 72°C after the first reaction cycle, secondary structures in mRNA are destabilized and elongation of the first strand may proceed in the subsequent cycle.

9. RNA should be digested after first strand synthesis to permit optimal synthesis of a second DNA strand, and to avoid interference in the following PCR amplification step.
10. The amount of RNA starting material versus the number of PCR cycles recommended by manufacturer is as follows (total RNA/mRNA/number of cycles): 1.0–2.0 μg/0.5–1.0 μg/18–20; 0.5–1.0 μg/0.25–0.5 μg/20–22; 0.25–0.5 μg/0.125–0.25 μg/22–24; 0.05–0.25 μg/0.025–0.125 μg/24–26. We recommend minimizing the number of amplification cycles to avoid PCR artifacts.

11. A proteinase K digestion prior to the gel extraction is recommended. Other PCR purification methods that efficiently remove dNTPs, salts, and long primers may be used as well (note that the longest primer used here is 59 nt long).

12. In order to maximize DNA recuperation, the elution volume may be increased to 100 μL, and subsequently reduced in a speed-vac.

13. The nonnormalized and the normalized samples generated with different DSN concentrations should be compared by agarose gel electrophoresis. For best results, the discrete bands of highly expressed mRNAs should have disappeared, and fragment sizes should be evenly distributed and not be smaller than in the nonnormalized sample.

14. For the extraction of the sized DNA fragments from the agarose gel, we discourage the use of gel extraction kits, because cloning efficiencies may be reduced by 1 to 3 orders of magnitude, compared to electroelution.

15. Size fractionation of cDNA is sometimes difficult because remaining contaminants (e.g., polysaccharides) are carried over despite purification. In such cases, DNA fragments tend to aggregate, causing contamination of the larger-size cDNA with small fragments in electrophoresis. Separation of smaller cDNA quantities (to avoid overloading of the gel) will often help to reduce, although not eliminate, the problem.

Acknowledgments

We wish to thank Jean-François Bouffard, Jung Hwa Seo, Zhang Wang, and Yun Zhu for excellent technical assistance. This research has been funded in part by grants from Génome Québec/Canada (GB; BFL) and CIHR Canada (BFL). NRE has been supported by the ‘Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación’ (Government of Basque Country). Salary and interaction support from the Canadian Institute for Advanced Research, Program in Evolutionary Biology (BFL; GB), is gratefully acknowledged.
References


Chapter 4

Generation of Full-Length cDNA Libraries: Focus on Plants

Motoaki Seki, Asako Kamiya, Piero Carninci, Yoshihide Hayashizaki, and Kazuo Shinozaki

Abstract

Full-length cDNAs are essential for the correct annotation of transcriptional units and gene products from genomic sequence data and for functional analysis of the genes. Full-length cDNA libraries are very important resources for isolation of the full-length cDNAs. The biotinylated cap trapper method using the trehalose-thermostabilized reverse transcriptase has been developed and has become an efficient method for construction of high-content full-length cDNA libraries. We have constructed full-length cDNA libraries from various plants and animals using this method. The protocol of the method is described in this chapter.

Key words: Arabidopsis thaliana, Full-length cDNAs, Full-length cDNA library, Biotinylated CAP trapper, Trehalose.

1. Introduction

Large-scale cDNA projects are in progress to determine the expressed genes in various organisms. However, most of these projects are based on cDNA libraries in which most of the inserts are not full-length. Although partial cDNAs are useful for making a catalog of the expressed genes, they are not suitable for further study of the gene function. Consequently, isolation of the full-length cDNA clones is important for the functional analysis of the genes.

Full-length cDNA libraries are invaluable resources for large-scale gene discovery projects. An efficient method for construction of full-length cDNA libraries, that is, the biotinylated cap trapper method, has been developed by Hayashizaki’s group about 10 years ago (Fig. 4.1; (1,2)). This system is based on the chemical introduction of a biotin group into the diol residue of the cap
structure of eukaryotic mRNA (Fig. 4.1). This step is followed by digestion by RNase I, a ribonuclease that can cleave single-stranded RNA at any site, followed by selection of full-length cDNA (Fig. 4.1). The library produced by this method contained a very high proportion of full-length cDNAs and produced an excellent yield without involving PCR amplification. Carninci et al. (3) also reported that introduction of a disaccharide, trehalose, to the reverse transcriptase reaction at high reaction temperature, 60°C, resulted in the synthesis of longer full-length cDNAs, and higher representation of long full-length cDNAs in the library.

Fig. 4.1. Strategy for full-length cDNA library construction by the biotinylated CAP trapper method. (A) A diol group is present on both the CAP structure and the 3'-end of the mRNA; (B) biotinylation reaction of the diol group for the CAP structure; (C) flowchart of the construction of a full-length cDNA library by biotinylated CAP trapper method.
The method using the trehalose-thermostabilized reverse transcriptase, which makes it possible to prepare longer full-length cDNAs, combined with the selection of full-length cDNAs by biotinylated cap trapper to remove nonfull-length cDNAs is ideal for construction of high-content full-length cDNA libraries. By using this method, the full-length cDNA libraries have been constructed from various plants, such as Arabidopsis (4,5), rice (6), wheat (7), poplar (8) and Physcomitrella patens(9), and animals, such as human and mouse (1,10). The detailed protocol and the principle of the full-length cDNA library construction are described in (http://www.dna-microarrays.org/content/restricted/protocols/pro01.pdf (11, 12, 13)).

We have constructed the 19 RIKEN Arabidopsis full-length (RAFL) cDNA libraries using various plant materials. In this chapter, we describe the protocol used for construction of RAFL 7-9 cDNA libraries ((5); See Note 1).
2. Materials

2.1. Preparation of Plant Materials

2.1.1. Choice of Plant Materials (See Note 2)

The following plant materials were used for construction of the RAFL 7–9 cDNA libraries: For the construction of the RAFL 7 and 8 libraries, *Arabidopsis* plants were grown on germination-medium (GM)-containing agar plates for 3 to 4 weeks with a cycle of 16-h light/8-h dark at 22°C. For the construction of the RAFL 9 library, the plants grown in soil under continuous light at 22°C and the plants...
grown on GM-containing agar plates were used. For the RAFL 7 library, the plants grown for 4 weeks were kept in an incubator for 1, 2, 5, 10, and 24 h at 4°C, and frozen. For the RAFL 8 library, plants grown for 4 weeks were removed from the agar, desiccated for 1, 2, 5, 10, and 24 h in plastic dishes under dim light, and frozen. For the RAFL 9 library, the plants harvested every 5 days after sowing until 11 weeks after sowing, the 4-week-old plants treated with dehydration (1, 2, 5, 10, and 24 h) and cold (1, 2, 5, 10, and 24 h), tissues of mature seeds, siliques, and flowers from the plants grown in soil for 6 to 11 weeks, and old, yellow leaves induced by natural senescence were used for the construction of the cDNA libraries.

2.1.2. Quantity (See Note 3)

2.2. Preparation of RNA (See Note 4)

1. GTC solution: 4 M guanidium thiocyanate (Fluka), 25 mM sodium citrate (Wako), pH 7.0, 0.5% sodium N-lauroylsarcosine, 0.1 M 2-mercaptoethanol. Prepare the solution just before use. Warm the solution to 50°C until it melts.

2. 0.1 M EDTA (pH 7.5). Store at room temperature.

3. 5.7 M cesium chloride. Cesium chloride (Iwai) is dissolved at 5.7 M in 0.1 M EDTA (pH 7.5). Store at room temperature.

4. RNA-dissolving buffer: 10 mM Tris-HCl, pH 7.5, 5 mM EDTA. Store at room temperature.

5. Equilibrated phenol: Melt phenol (Wako) at 68°C. Add 0.1 M Tris-HCl (pH 8.0) to equilibrate the phenol. Store in a light-tight bottle at 4°C for periods of up to 1 month, or at –20°C indefinitely.


8. 8 M Lithium chloride (Nacalai). Store at 4°C.

9. 70% Ethanol: Add 0.3 volume of autoclaved distilled water to ethanol. Store at –20°C.

10. Water (RNase-free). Store at room temperature.

2.3. Preparation of Poly(A)\(^+\) RNA from Total RNA

1. Oligo(dT)-Latex (OligotexTM-dT30-Super; Roche, Tokyo, Japan). Store at 4°C.

2. Elution Buffer: 10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 0.1% SDS. Store at room temperature.

3. 5 M sodium chloride. Store at room temperature.

4. 3 M sodium acetate (pH 5.6). Store at room temperature.

5. Water (RNase free). Store at room temperature.
2.4. Trehalose-Enhanced Synthesis of First-Strand cDNA

1. mRNA (10 μg). Store at –80°C.
2. Purified oligo(dT) primer-adapter (2 μg/μl). 1st-BS primers (43-mer, 5'-GAGAGAGAGAGATCCAGAGGCTTT TTGTTTTTTTTTN-3', V: A or G or C, N: A or G or C or T; See Note 5) are used for construction of the RAFL 7–9 cDNA libraries (5). Store at –20°C.
4. 5 X first-strand buffer (GIBCO-BRL). Store at –20°C.
5. 80% glycerol. Store at room temperature.
6. 100 mM dithiothreitol. Store at –20°C.
7. Solution containing dATP, dGTP, dTTP, and 5-methyl-dCTP, each at a concentration of 10 mM. Store at –20°C.
8. 2.5 mg/ml BSA. Store at –20°C.
10. Sorbitol (4.9 M) (85529, Fluka). Store at room temperature.
11. RNase H– SUPERSCRIPT II reverse transcriptase (Invitrogen) (200 U/μl). Store at –20°C.
12. [α-32P]dGTP (10 mCi/ml, 2000 Ci/mmole). This radionucleotide is used as tracer to calculate the cDNA yield in the following steps. Store at –20°C.
13. 0.5 M EDTA. Store at room temperature.
14. 10% SDS. Store at room temperature.
15. Proteinase K (10 μg/μl). Store at –20°C.
16. 0.5 M phosphate buffer (pH 7.0). Store at room temperature.
17. 10 M ammonium acetate. Store at room temperature.

2.5. Oxidation and Biotinylation of Diol Groups of mRNA

1. 1 M sodium acetate (pH 4.5). Store at room temperature.
2. 200 mM sodium periodate solution. The solution must be freshly prepared. Keep in a light-tight tube on ice until use.
3. 10% SDS. Store at room temperature.
4. 5 M sodium chloride. Store at room temperature.
5. Isopropanol. Store at room temperature.
6. 1 M sodium acetate (pH 6.1). Store at room temperature.
7. 1 M sodium citrate (pH 6.0). Store at room temperature.
8. Biotinylation buffer: 33 mM sodium citrate, pH 6.0, 0.33% SDS. Store at room temperature.
9. 10 mM biotin hydrazide long arm. For each reaction, dissolve 0.56 mg of biotin hydrazide long arm (MW = 371.51; Vector Laboratories) in 160 μl of biotinylation buffer. The solution of biotin hydrazide long arm must be freshly prepared. Keep in a light-tight tube until use.
10. Ethanol. Store at −20°C.
11. 80% ethanol. Store at −20°C.
12. 0.1 X TE: 1 mM Tris, pH 7.5, 0.1 mM EDTA. Store at room temperature.

2.6. Cap-Trapping and Release of Full-Length cDNA

1. RNase ONE (5 U/μl) (Promega). Store at −20°C.
2. RNase ONE buffer (Promega). Store at −20°C.
3. DNA-free tRNA (Sigma). Store at −20°C.
5. Binding buffer: 4.5 M sodium chloride, 50 mM EDTA, pH 8.0. Store at room temperature.
6. Magnetic stand to hold 1.5-ml tubes.
7. Washing solution 1: 0.3 M sodium chloride, 1 mM EDTA, pH 8.0. Store at room temperature.
8. Washing solution 2: 0.4% SDS, 0.5 M sodium acetate, 20 mM Tris-HCl, pH 8.5, 1 mM EDTA. Store at room temperature.
9. Washing solution 3: 0.5 M sodium acetate, 10 mM Tris-HCl, pH 8.5, 1 mM EDTA. Store at room temperature.
10. Alkaline treatment buffer: 50 mM sodium hydroxide, 5 mM EDTA. Store at room temperature.
11. 1 M Tris-HCl (pH 7.0). Store at room temperature.
13. cDNA spun column (Sephacryl S-400) kit (Pharmacia Biotech). Store at 4°C.
14. 2 mg/ml glycogen. Store at −20°C.

2.7. Homopolymeric Tailing of First-Strand cDNA with dG Residues

1. Terminal deoxynucleotidyl transferase (TdT) (Takara, 10 U/μl). Store at −20°C.
2. TdT buffer (5 X) (Toyobo): 500 mM sodium cacodylate, pH 7.2, 5 mM Cobalt dichloride, 1 mM dithiothreitol. Store at −20°C.
3. Saturated trehalose (~80% in water). Store at room temperature.
4. Synthetic oligonucleotide (20–30 mer, 10 pmol/μl). Used for control reaction. Store at −20°C.
5. [γ-32P]dATP. Store at −20°C.
7. 10 mM magnesium chloride. Store at room temperature.
2.8. Synthesis of Second-Strand cDNA

1. LA Taq polymerase (TaKaRa, 5 U/μl). Store at –20°C.
2. 2X GCI buffer (for TaKaRa LA Taq). Store at –20°C.
3. Second-strand primer (100 ng/μl). 2nd-X primer (45 mer, 5’-GAGAGAGAGATTCTCGAGTTAATTAAATTAATCCC CCCCCCCCC-3’; See Note 6) are used for construction of the RAFL 7–9 cDNA libraries (5). Store at –20°C.
4. [α-32P]dGTP (10 mCi/ml; 2000 Ci/m mole). Store at –20°C.
5. Solution of all four dNTPs, each at a concentration of 2.5 mM. Store at –20°C.
6. DE-81 chromatographic paper (Whatman). Store at room temperature.

2.9. Cleavage of the cDNA

In this section, we describe the restriction enzymes used for cloning into λ-FLC-I-B (15).

3. NEBuffer 2 (New England Biolabs): 50 mM sodium chloride, 10 mM Tris-HCl, 10 mM magnesium chloride, 1 mM dithiothreitol, pH 7.9. Store at –20°C.
4. 10 mg/ml bovine serum albumin (BSA) (New England Biolabs). Store at –20°C.
5. Shrimp alkaline phosphatase (Amersham Biosciences, 1 U/μl). Store at –20°C.
6. SizeSep 400 CL-4B spun column (Amersham Pharmacia Biotech). Store at 4°C.

2.10. Cloning of the cDNA into the Lambda Cloning Vector

1. λ-FLC-I-B cloning vector (15). Store at 4°C.
4. E. coli strain, C600 (F thi-1 thr-1 lenB6 lacY1 tonA21 supE44 λ–). Store as glycerol stocks at –80°C.

3. Methods

3.1. Preparation of Total RNA

1. Grind the plant materials in a mortar with a pestle in the presence of liquid nitrogen.
2. Mix the powder with 5 volumes of the GTC solution.
3. Centrifuge at 14,000 g for 10 min at 4°C.

4. Layer about 2.4 ml of the supernatant on top of 1.1 ml of 5.7 M cesium chloride solution to create a step gradient. Centrifuge them for 2 h in a TLA-100.3 rotor (Beckman, CA, USA) at 70,000 rpm and 20°C.

5. Dissolve the RNA pellet in the buffer (10 mM Tris-HCl, pH 7.5, 5 mM EDTA).

6. Extract the supernatant with successive, equal volumes of phenol, phenol-chloroform, and chloroform.

7. Collect the upper phase and mix with 1/3 volume of 8 M lithium chloride. Keep the tube at 4°C for more than 30 min.

8. Centrifuge at 14,000 g for 30 min at 4°C.

9. Wash the pellet with 70% ethanol and dissolve in water.

3.2. Preparation of Poly(A)+ RNA from Total RNA

Poly(A)+ RNA is prepared from the total RNA by using Oligo(dT)-Latex (OligotexTM-dT30-Super; Roche, Tokyo) as described by the manufacturer.

3.3. Trehalose-Enhanced Synthesis and Extraction of First-Strand cDNA

3.3.1. Synthesis of First-Strand cDNA (See Note 7)

1. To prepare the first-strand cDNA, put the following reagents in three 0.5-ml PCR tubes (A, B, and C).

   Tube A: In a final volume of 21.3 μl, add the following:
   - mRNA: 10 μg (precipitate with ethanol and resuspend directly into the primer and glycerol mixture)
   - 1st-BS primer (2 μg/μl): 7 μl (14 μg)
   - 80% glycerol: 14.3 μl

   2. Close the top of the tube and heat the mixture of mRNA and primers to 65°C for 10 min to melt secondary structures in the RNA. During the incubation, quickly prepare the reaction mixtures in Tubes B and C.

   Tube B: In a final volume of 142.5 μl, add the following:
   - 5x first-strand buffer: 28.6 μl
   - 100 mM dithiothreitol: 11.0 μl
   - Solution containing dATP, dGTP, dTTP, and 5-methyl-dCTP, each at a concentration of 10 μM: 9.3 μl
   - 4.9 M sorbitol: 55.4 μl
   - Saturated solution of trehalose: 23.2 μl
   - RNase H−SUPERSCRIPT II RT (200 units/μl): 15.0 μl

   Program a thermal cycler as follows:
   - 40°C for 4 min
   - 50°C for 2 min
   - 56°C for 60 min
   - 4°C for an indefinite time
Tube C: Transfer 1–1.5 μl of [α-32P]dGTP to the tube.

2. Use the following “hot start” protocol to start the reaction.
   a. Transfer Tubes A, B and C to the thermal cycler and start the cycling program.
   b. When the temperature reaches 42°C, quickly mix the contents of Tubes A and B.
   c. Transfer 40 μl of the A+B mixture to Tube C.
   d. Complete the cycling program.

3. At the end of the program cycle, stop the reaction by adding 0.5 M EDTA to a final concentration of 10 mM.

4. Measure the incorporation of radiolabeled precursor into first-strand cDNA (Tube C) by spotting aliquots of the radioactive reaction mixture onto DE-81 papers. Calculate the incorporation by measuring the radioactivity before and after three washings with 0.5 M phosphate buffer (pH 7.0). Calculate the yield of first-strand cDNA. Measure the size of the radiolabeled first-strand cDNA by alkaline agarose gel electrophoresis.

3.3.2. Organic Phase Extraction and cDNA Precipitation (See Note 8)

1. Transfer the large-scale first-strand cDNA reaction (Tube A+B) and the small-scale pilot reaction containing [α-32P]dGTP (Tube C) to a standard microfuge tube (PCR tubes cannot be used for the centrifugation at 15,000 rpm with organic solvents, because they may be crushed).

2. Add 3 μl of proteinase K (10 μg/μl) and 3 μl of SDS (10%). Incubate the reaction for 15 min at 45°C.

3. Perform phenol-chloroform and chloroform extraction and back extraction as follows:
   a. Add 0.5 volume of phenol and 0.5 volume of chloroform. Vortex moderately until the two phases mix.
   b. Centrifuge for 2 min at 15,000 rpm.
   c. Remove the aqueous phase (upper layer) and transfer it to a fresh 1.5 ml tube. Add an equal volume of chloroform and vortex gently.
   d. Centrifuge for 2 min at 15,000 rpm. Transfer the aqueous phase (upper layer) to a fresh, clean tube. Keep the tube with chloroform for the back extraction.

Perform back extraction to recover the residual cDNA from the residual aqueous phase that could not be recovered after the first extraction.
   a. Add 50 μl of water to the phenol-chloroform tube. Vortex gently.
   b. Centrifuge for 2 min at 15,000 rpm.
c. Transfer the upper phase to the chloroform tube. Vortex and centrifuge for 2 min at 15,000 rpm.
d. Transfer the upper phase to the chloroform tube.
e. Vortex and centrifuge for 2 min at 15,000 rpm.
f. Transfer the upper, aqueous phase to the previously extracted fraction of cDNA.

4. Add 0.5 volume of 10 M ammonium acetate.
5. Add 2.5 volume of ethanol. Incubate at –80°C for 30 min.
6. Centrifuge at 15,000 rpm for 15 min.
7. Wash the pellet with 70% ethanol. After ethanol precipitation, dissolve the cDNA in 46 μl of water.

3.4. Oxidation and Biotinylation of Diol Groups of mRNA (See Note 9)

3.4.1. Oxidation of the Diol Groups of mRNA

1. Add the following reagents directly to the 46 μl of cDNA samples obtained in Step 3.3.2.7:
   3.3 μl of 1 M sodium acetate (pH 4.5)
   100 mM sodium periodate to a final concentration of 10 mM.
2. Incubate the mixture on ice in the dark for 45 min.
3. Stop the reaction by adding 1 μl of 80% glycerol and vortexing (See Note 10).
4. Recover the nucleic acids by precipitation with 61 μl of isopropanol, 0.5 μl of 10% SDS, and 11 μl of 5 M sodium chloride. Incubate the mixture for 30 min in the dark at –80°C.
5. Centrifuge at 15,000 g for 15 min. Wash the pellet in 500 μl of 80% ethanol and centrifuge at 15,000 g for 2–3 min. Carefully discard the supernatant. Centrifuge again briefly and remove any remaining ethanol. Dissolve the nucleic acids in 50 μl of water.

3.4.2. Derivatization of the Oxidized Diol Groups

1. Add 160 μl of a freshly prepared solution of biotin hydrazine (long arm) to the first-strand cDNA-mRNA hybrid samples (50 μl). Mix the solution gently and then incubate it overnight (10–16 hours) at room temperature (20–24°C).
2. At the end of the incubation, precipitate the biotinylated cDNA-mRNA hybrids by adding:
   1 M sodium acetate (pH 6.1): 75 μl
   5 M sodium chloride: 5 μl
   ethanol: 750 μl
   After gentle mixing, incubate the solution for 30 min at –80°C or –20°C.
3. Recover the precipitate by centrifuging the sample at 15,000 g for 15 min. Discard the supernatant and remove any free biotin hydrazide from the pellet by washing three times with 80% ethanol, centrifuging after each wash.
4. Dissolve the final pellet in 70 μl of 0.1 X TE. Monitor for complete resuspension using a hand-held monitor.

3.5. Cap-Trapping and Release of Full-Length cDNA

3.5.1. Digestion with RNase I

1. To the cDNA sample (70 μl), add:

   - RNase I buffer (Promega): 20 μl
   - RNase I (5 units/μl): 1 unit/μg starting mRNA
   - (10 units of RNase I are used if the starting mRNA quantity is 10 μg.)
   - Water: to a final volume of 200 μl

   Incubate the reaction mixture for 15–30 min at 37°C.

2. Stop the digestion reaction by placing the sample on ice. Add:

   - 10% SDS: 4 μl
   - Proteinase K (10 μg/μl): 3 μl

   Incubate the reaction mixture for 15 min at 45°C.

3. Stop the proteinase K reaction by extraction with phenol. Transfer the supernatant to a clean tube and back-extract the organic phase with water. Precipitate the nucleic acids using isopropanol and 20 μg of tRNA as carrier. Dissolve the pellet in 20 μl of 0.1x TE.

3.5.2. Pretreatment of Magnetic Beads (See Note 11)

1. Pretreat MPG beads (500 μl per 25 μg of starting mRNA) with 100 μg of DNA-free tRNA. Incubate the beads for 30 min on ice with occasional mixing to prevent bead sedimentation.

2. Separate the beads by standing the tube in a magnetic rack for 3 min. Carefully remove the supernatant.

3. Wash the beads three times with 500 μl of binding buffer (4.5 M sodium chloride, 50 mM EDTA, pH 8.0).

3.5.3. Capture of Full-Length cDNA (See Note 12)

1. Resuspend the tRNA-treated MPG beads in 500 μl of wash/binding buffer (4.5 M sodium chloride, 50 mM EDTA, pH 8.0) and transfer 350 μl of the suspension to a fresh tube containing the biotinylated first-strand cDNA (20 μl of cDNA in 0.1x TE). After mixing gently, incubate the reaction for 10 min at 50°C, with slow rotation.

2. At the end of the incubation, transfer the remaining 150 μl of beads to the reaction tube and continue the incubation for a further 20 min at 50°C.

3. Separate the beads by standing the tube in a magnetic rack for 3 min.

4. Carefully remove the supernatant and wash the beads as follows:
2x with 500 μl of washing/binding solution (4.5 M sodium chloride, 50 mM EDTA, pH 8.0).
1x with 500 μl of washing solution 1 (0.3 M sodium chloride/1 mM EDTA).
2x with 500 μl of 0.4% SDS/0.5 M sodium acetate/20 mM Tris-HCl, pH 8.5/1 mM EDTA.
2x with 500 μl of 0.5 M sodium acetate/10 mM Tris-HCl, pH 8.5/1 mM EDTA.

3.5.4. Eluting the cDNA

1. After the final wash, elute the cDNA by adding 100 μl of alkaline treatment buffer: (50 mM sodium hydroxide, 5 mM EDTA). Stir the beads briefly and incubate them for 5 min at room temperature, mixing occasionally.
2. Separate the magnetic beads and transfer the supernatant (containing the eluted cDNA) to a fresh tube. Store the cDNA on ice.
3. Repeat the elution twice more, each time using 100 μl of the alkaline treatment buffer. Check that the elution is complete (80–90%) by comparing the amounts of the radioactivity in the eluate and on the beads.
4. Store the pooled eluates on ice to prevent any hybridization between cDNAs and any contaminating RNAs that might be present.

3.5.5. Treatment of Cap-Trapped First-Strand cDNA with RNase I

1. Add 50 μl of 1 M Tris-HCl (pH 7.0) to each tube (on ice) and mix quickly.
2. Add 1 μl of RNase I (5 units/μl) and mix quickly. Incubate for 10 min at 37°C.
3. To remove the RNase I, treat the cDNA with proteinase K, extract with phenol:chloroform, and back-extract the aqueous phase, as described in the above-mentioned RNase I treatment step.
4. Add 3 μg of glycogen to each tube. Concentrate the cDNA to 20 μl by ultrafiltration with a Microcon-100 filter at 2000 rpm and 25°C. Desalt the samples by adding 400 μl of 0.1x TE and then centrifuging as above for a total of three washes. Recover the DNA by inverting the filter in a new tube and centrifuging at 9000 rpm for 1 min.
5. Purify the cDNA samples obtained in Step 4 using the cDNA Spun Column Kit. This purification step is necessary to remove the primer-adapter, the blocking tRNA, and the residual digested RNAs before adding homopolymeric tails to the first-strand cDNAs or performing the single-strand linker ligation method (16).
6. Add 5 M sodium chloride solution to a final concentration of 0.2 M, and 1.5 µl of glycogen (2 mg/ml). Precipitate the cDNAs by adding 1 volume of isopropanol. Incubate at –80°C for 30 min.

7. Centrifuge at 15,000 rpm for 30 min. Wash the pellet twice with 80% cold ethanol. Dissolve the pellet in 10 µl of water.

3.6. Homopolymeric Tailing of First-Strand cDNA with dG Residues
(See Note 13)

3.6.1. Control Reaction: Using a Radiolabeled Oligonucleotide to Count the Number of dG Residues Added

1. Radiolabel the 5′ terminus of a single-stranded oligonucleotide, using [γ-32P]ATP and polynucleotide kinase, as described in Molecular Cloning 3rd Edition, Chapter 10 (14), Protocol 2. After purification, dilute the labeled oligonucleotide to a concentration of 1 ng/µl and store it at –20°C.

2. Transfer a small aliquot (~5–10%) of the tailing reaction containing cDNA (see below) to a fresh tube and add enough 32P-labeled oligonucleotide (usually ~0.5 µl) to detect easily by autoradiography after gel electrophoresis.

3. At the end of the pilot tailing reaction, analyze the labeled oligonucleotide by electrophoresis through the polyacrylamide gel using appropriate molecular-weight markers. Estimate the number of labeled dG residues added to the oligonucleotide.

3.6.2. Main Reaction: Adding dG Residues to First-Strand cDNA

1. Immediately before tailing, heat the cDNA samples to 65°C for 2 min to melt any secondary structure that might decrease the tailing efficiency. After heating, transfer the sample to ice.

2. To the cDNA sample, add:
   15 µl of saturated trehalose
   variable quantity of 50 µM dGTP, to a final concentration of
   20 µM for >750 ng cDNA
   15 µM for >400 ng cDNA
   10 µM for >150 ng cDNA
   5 µM for <150 ng cDNA
   10 µl of 5x TdT buffer
   5 µl of 10 mM magnesium chloride
   4 µl of terminal deoxynucleotidyl transferase (TaKaRa, 10U/µl)
   Adjust the final volume of the reaction to 50 µl with water.

3. Mix and transfer 2 µl of the main reaction to a tube containing ~0.5 µl of the 32P-labeled oligonucleotide.

4. Incubate both reactions for 15 min at 45°C.

5. Transfer the reactions to ice and measure the number of dG residues added to the 32P-labeled oligonucleotide.

6. After analyzing the tailing reaction, add 1 µl of 0.5 M EDTA to the tube containing the main reaction.
7. If the results of the pilot reaction are satisfactory, purify the cDNA with proteinase K digestion, extraction with phenol:chloroform and chloroform, and back-extraction. Precipitate the cDNA with isopropanol. After precipitation and washing with 70% ethanol, dissolve the cDNA in 15 μl of 0.1 X TE.

### 3.7. Second-Strand cDNA Synthesis

1. Set the following program on a thermal cycler:
   
   **Step 1:** 5 min at 65°C
   
   **Step 2:** negative ramp of -20°C (until 45°C), with slope of 0.1°C/second
   
   **Step 3:** 45°C for 10 min
   
   **Step 4:** 68°C for 20 min; repeat **Steps 3** and **4** once
   
   **Step 5:** 72°C for 10 min
   
   **Step 6:** 4°C for an indefinite time

2. Prepare Tubes A and B containing:
   
   **Tube A:**
   
   - 15 μl G-tailed cDNA
   - 6 μl of second-strand primer-adapter (100 ng/μl)
   - 30 μl of 2 x GCI buffer (for TaKaRa LA Taq)
   - 6 μl of solution of all four dNTPs, each at a concentration of 2.5 mM

   **Tube B:** 0.5 μl of \([\alpha-^{32}P]dGTP\)

3. Start the thermal cycler program and transfer Tubes A and B to the thermal cycler.

4. When the samples are at 65°C, add 3 μl of LA Taq (5 U/μl) to Tube A (final volume = 60 μl).

5. Mix the contents of the tube quickly but thoroughly, and immediately transfer a 3 μl aliquot to Tube B.

6. Allow the program to run its course and then transfer the tubes to ice.

7. Quantify the yield of the second-strand reaction using DE-81 paper.

8. Treat the second-strand cDNA with proteinase K, extract the solution with phenol:chloroform, back-extract with 0.1x TE, and precipitate the cDNA with isopropanol. Dissolve the cDNA in 20 μl of water.

### 3.8. Cleavage of the cDNA

1. Calculate the units of restriction enzymes to be used to cleave the cDNA based on the yield of second-strand cDNA and assuming that no significant amount of double-stranded cDNA has been lost during purification.
2. To the cDNA samples, add:

- 10x NEBuffer 2: 3 μl
- 10 mg/ml BSA: 0.3 μl
- *Bam*HI (20U/μl): 25 U/μg of cDNA
- *Xho*I (20U/μl): 25 U/μg of cDNA
- Water: to a final volume of 30 μl

3. Incubate the digestion for 1 hour at 37°C.

4. Add the appropriate amount (1.3 U/μg of cDNA) of shrimp alkaline phosphatase to dephosphorylate the cleaved cDNA. Incubate for 15 min at 37°C.

5. Add:

- 1 μl of 0.5 M EDTA
- 1 μl of 10% SDS
- 1 μl of 10 mg/ml proteinase K

Incubate for 15 min at 45°C.

6. Extract mixture once with phenol:chloroform and then once with chloroform. Precipitate the cDNA with isopropanol. After precipitation and washing with 70% ethanol, dissolve the cDNA in 50 μl of 0.1 X TE.

7. Purify the cDNA samples obtained in Step 6 using SizeSep 400 CL-4B spun column kit as described in the manufacturer’s instructions. This purification step is necessary to remove the primer-adapters.

8. Precipitate the cDNA from the pooled fractions containing the bulk of the cDNA with ethanol. Wash the pellet with 70% ethanol. Before redissolving the cDNA, measure the radioactivity in the pellet by scintillation counting. Calculate the amount of cDNA recovered from the column. Dissolve the cDNA in 3 μl of water.

### 3.9. Cloning of the cDNA into the Lambda Cloning Vector


2. Set up the ligation reactions in a final volume of 5 μl. Mix as much cDNA as possible with an equimolar quantity of λ arms.

3. Adjust the volume to 4 μl.

4. Add 0.5 μl of T4 DNA ligase and 0.5 μl of buffer.

5. Incubate the reaction overnight at 16°C.

6. Package the ligation products using 25 μl of a highly efficient packaging mix.
7. After packaging, measure the number of viable bacteriophages in an aliquot of the mixture by plaque formation on lawns of C600 bacteria. Usually, the entire cDNA cloning should yield $10^5$ to $10^7$ primary plaque-forming units of the recombinant bacteriophages.

8. Amplify the cDNA library according to the standard protocol (See Note 14).

4. Notes

1. When new libraries are constructed, the cDNAs are usually not normalized nor subtracted. The cDNA libraries are first constructed using the standard protocol shown in Fig. 4.1C, and then the normalization and subtraction procedures (17) are introduced in the construction of the full-length cDNA libraries to isolate new genes. The method is based on hybridization of the first-strand full-length cDNA with several RNA drivers, including starting mRNA as the normalizing driver and run-off transcripts from rearrayed clones as subtracting drivers. The detailed protocol for construction of the normalized and subtracted full-length cDNA libraries is described in http://www.dna-microarrays.org/content/restricted/protocols/pro01.pdf.

2. We have constructed 19 RAFL cDNA libraries (5) using various Arabidopsis plant materials. We have used the materials of the plants grown in the laboratories as the starting ones. If the materials from the field are used, the greatest care must be taken for the contamination of other organisms. Construction of the full-length cDNA libraries from all plant materials including cultured cells (6) is possible if sufficient RNAs can be prepared. It is recommended to use the mixture of RNAs isolated from various plant materials if construction of only 1 full-length cDNA library is planned for some plant species and isolation of more nonredundant cDNA clones from the library is expected.

3. Efficiency of RNA isolation depends on the plant materials used. About 10 µg of poly(A)$^+$ RNA is necessary for the construction of the standard plant full-length cDNA libraries.

4. Additional protocol using TRIZOL Reagent (Invitrogen, Carlsbad, CA) is also recommended for construction of the full-length cDNA libraries from the plant materials.
5. The 1st-BS primer contains the recognition sequence of $Bam$HI, which is used for cloning into the $\lambda$-FLC-1-B vector and also contains the recognition sequence of $Sst$I, which was used for construction of the RAFL 4–6 libraries (5).

6. This primer contains the recognition sequence of $Xho$I, which is used for cloning into the lambda vector.

7. The variant of the protocol using TaKaRa LA Taq buffer, which is commercially available from TaKaRa (Kyoto, Japan) and was originally developed for use in long PCR is also recommended. The protocol is described in STAGE IVb of the Web site http://www.dna-microarrays.org/content/restricted/protocols/pro01.pdf.

8. To purify the first-strand cDNA, an additional protocol by digestion with proteinase K and precipitation with CTAB (cetyltrimonium bromide) is also recommended. The protocol is described in STAGE VI of the Web site http://www.dna-microarrays.org/content/restricted/protocols/pro01.pdf.

9. The work should be done under RNase-free conditions until the biotinylation reaction has been completed.

10. Glycerol reacts with sodium periodate and quenches the reaction.

11. Streptavidin-coated porous glass beads are recommended because of their low level of nonspecific binding of nucleic acids and because of their high capacity for binding of biotinylated nucleic acids. To minimize nonspecific binding of nucleic acids to the beads, the beads are preincubated with a large excess of DNA-free tRNA. tRNA binds to and blocks any possible reactive “sticky” site on the beads.

12. Streptavidin, immobilized on the surface of the magnetic beads, binds biotin with very high affinity. Because the surface area of the beads is very large, streptavidin-coated beads have a very high capacity to bind biotin (>200 pmoles/mg); 1 mg of beads can bind 800 pmoles of a biotinylated 25-mer oligonucleotide. The combination of high affinity and high capacity allows the beads to capture virtually all biotinylated full-length cDNAs. The low level of nonspecific binding means that it is possible to use a large excess of beads in the binding reaction.

13. For creating a site for priming second-strand cDNA synthesis, an additional method, single-strand linker ligation method (16), which uses DNA ligase to add a double-stranded (ds) DNA linker to single-stranded (ss) full-length cDNA is also recommended.

14. It is recommended to amplify on the solid phase using up to 50,000 phage/large plate.
Acknowledgments

This work was supported by a grant for Genome Research from RIKEN, the Program for Promotion of Basic Research Activities for Innovative Biosciences, the Special Coordination Fund of the Science and Technology Agency, and a Grant-in-Aid from the Ministry of Education, Culture, Sports and Technology of Japan (MEXT) to KS. It was also supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (C) “Genome Science” from MEXT to MS and by a grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to YH.

References


Chapter 5

Preparation of Full-Length cDNA Libraries: Focus on Metazoans

Masako Harada and Yoshihide Hayashizaki

Abstract

Critical steps in a cDNA library preparation include efficient cDNA synthesis, selection of full-length cDNAs, normalizing their abundance, and the subtraction of redundant transcripts. The use of trehalose and sorbitol stabilizes the activity of the reverse transcriptase leading to efficient cDNA synthesis and the cap-trapping method is used for efficient full-length cDNA selection. Through the incorporation of additional normalization and subtraction steps that eliminate the size bias and expressed gene frequency, it is possible to attain cDNA libraries that include larger or rarely expressed genes. This chapter describes an efficient method to construct a full-length cDNA library, with a focus on metazoan samples.

Key words: mRNA, metazoan, full-length cDNA, CTAB-urea method, cap-trapper method, biotinylation of diol groups of mRNA, normalization and subtraction of cDNA.

1. Introduction

Until recently, ESTs (Expressed-sequence tags) were the most common method for gene discovery and investigation. However, ESTs still only represent fragments of a gene since the method is not specialized to give long transcripts; furthermore, rare ESTs are often hard to find among the large numbers of highly expressed ESTs. Even with subtracted EST libraries, it can be difficult to detect rare transcripts. With the advent of full-length cDNA library constructions, the full-length sequencing method has become an indispensable tool for the investigation and detection of new genes. The full-length sequencing method also captures not only the coding regions but also the noncoding transcribed region of DNA (which gives rise to noncoding RNA), in a higher degree
than ESTs. The combination of these features gives the full-length sequencing method the ability to supply more information about the genome and the genome organization and structure, and has therefore become vital to the new field of noncoding RNA research among others.

The already established cDNA sequencing techniques have a low proportion of full-length cDNA transcripts, due to the individual features of mRNAs and enzymes. These features were overcome in the full-length cDNA sequencing by the development of the following techniques. Firstly, the reverse transcriptase involved in first-strand cDNA synthesis from mRNA templates has an optimal temperature of around 42°C. This temperature leads to mRNA truncation due to an early termination of cDNA synthesis and detachment of the enzyme from mRNA itself. To overcome this problem, the disaccharide trehalose was introduced to maintain reverse transcriptase activity at high temperature and hence improve thermostabilization. Moreover, it was also found that trehalose can increase the activity of the enzyme at high temperature, the so-called thermo activation (1). Secondly, the difficulty of being able to select only full-length cDNAs from a pool of immature cDNAs of various lengths prompted the development of a selection method favoring mature transcripts. The biotinylated cap-trapper method is a selection method that introduces biotin groups to the diol residue of the cap structure of mature eukaryotic mRNA. Uncapped immature transcripts are then subsequently degenerated by RNase I (2). Thirdly, longer transcripts require a cloning vector capable of carrying an insert with various sizes. Furthermore, the plasmid also needs to be available at a high fraction within a cell. Consequently the λ-full-length cDNA cloning vectors were developed as new vectors suitable for full-length cDNA construction (3). Finally, for full-length cDNA library construction, further stages of normalization and subtraction are necessary, since mRNAs have a frequency bias against the abundance of individual mRNAs (rare transcripts are harder to find). This problem can be solved by library subtraction. In this process, hybridization of first-strand, full-length cDNA with several RNA drivers removes redundant, already discovered transcripts. This allows the detection of previously undetected new genes in a single step (4, 5).

This chapter deals with preparation of cDNA libraries with a focus on metazoans. However, it is to be noted that the method is equally applicable to the other organisms. The source of materials preferred for these methods are mainly tissues and cells, which are frequently used in laboratories. According to the original paper, the authors described that microgram quantities of RNA are sufficient for library generation (6). However, we suggest that milligram quantities of RNA are generally required for nonexperienced users.
In this chapter, the methods for full-length cDNA preparations are described. The methods start with RNA extraction and purification, followed by protocols describing how to normalize RNA populations for abundance and end with details of cloning and constructing the plasmid libraries. These protocols are more for large-scale full-length cDNA libraries, and should therefore be modified when used for small-scale cDNA library construction (see Note 1).

2. Materials

2.1. Preparation of RNA

2.1.1. From Organisms

2.1.1.1. Worm (Nematode)

2.1.1.2. Insect

2.1.1.3. Mammalian Brain

2.1.2. From Whole Tissues

M9 Medium [750 ml of sterile H₂O, 200 ml of 5 × M9 salts, 2 ml of 1 M MgSO₄, 20 ml of 20% solution of the appropriate carbon source (e.g., 20% glucose), 0.1 ml of 1 M CaCl₂, sterile deionized H₂O], Trizol reagent (Invitrogen), chloroform, isopropanol, 75% ethanol, RNase-free sterile H₂O, RNase-free microcentrifuge tube, RNeasy Mini kit, RNase-free DNase Set.

Phosphate-buffered saline (PBS), liquid nitrogen, Solution D (denaturing solution) [4 M guanidinium thiocyanate, 25 mM sodium citrate·2H₂O, 0.5% (w/v) sodium lauryl sarcosinate, 0.1 M β-mercaptoethanol], BME (2-mercaptoethanol) (Ameresco), Sterilized disposable tip, 2 M NaOAc (pH 4.0), water saturated phenol, chloroform/isoamyl alcohol (24:1), isopropanol, 75% ethanol, DEPC-treated H₂O.

Trizol reagent (Invitrogen), chloroform, isopropanol, 75% ethanol, RNase-free sterile H₂O, RNase-free microcentrifuge tube, RNeasy Mini kit (QIAGEN), RNase-free DNase Set (QIAGEN).

Chloroform/isoamyl alcohol (49:1, v/v), ethanol, isopropanol, liquid nitrogen, phenol, phosphate-buffered saline (PBS), 2 M sodium acetate (pH 4.0), solution D (denaturing solution) [4 M guanidinium thiocyanate, 25 mM sodium citrate·2H₂O, 0.5% (w/v) sodium lauryl sarcosinate, 0.1 M β-mercaptoethanol], cells or tissue samples for RNA isolation, Sorval SS-34 rotor or equivalent, Sorval H1000 rotor or equivalent, cuvettes for measuring absorbance at 260 nm, homogenizer, mortar and pestle washed in DEPC-treated H₂O, prechilled, polypropylene snap-cap tube, water bath or heating block preset to 65°C (7).

Sterilized dH₂O, CTAB (cetyltrimethylammonium bromide), 5 M NaCl, CTAB/urea solution [1% CTAB, 4 M urea, 50 mM Tris-HCL (pH 7.0)], 1 mM EDTA (pH 8.0), 1.2 M NaCl, 70% ethanol, H₂O.
2.1.3. From Cultured Cells

Trypsin, sterile ice-cold phosphate-buffered saline solution (PBS), sterile ice-cold PBS, ice-cold 0.5% NP-40 lysis buffer, vanadyl ribonucleoside, RNase inhibitor (RNase inhibitor is supplementary if the budget allows) (see Note 2), CTAB/urea solution [1% CTAB, 4 M urea, 50 mM Tris-HCl (pH 7.0), 1 mM EDTA (pH 8.0)], 7 M guanidinium chloride, phenol: chloroform (1:1) solution, chloroform, isopropanol, 70% ethanol, H2O.

2.1.4. Preparation of Poly(A)+ RNA from Total and Cytoplasmic RNA

Selection of poly(A)+ RNA

A) 2x Column-loading buffer [40 mM Tris-HCl (pH 7.6), 1 M NaCl, 2 mM EDTA (pH 8.0), 0.2% (w/v) sodium lauryl sarcosinate], elution buffer [10 mM Tris-HCl (pH 7.6), 1 mM EDTA (pH 8.0), 0.05% SDS], ethanol, RNase-free 5 M NaCl, 10 N NaOH, 3 M sodium acetate (pH 5.2), total RNA, Sorvall SS-34 rotor or equivalent, cuvettes for measuring absorbance at 260 nm, Dispo-column (Bio-rad) or a Pasteur pipette plugged with sterile glass wool, equipment for storage of RNA, oligo(dT)-cellulose, pH paper (pH test strips, Sigma), water bath or heating block preset to 65°C (see Note 2).

B) Absorption/washing buffer [TES containing 0.5 M NaCl], ammonium acetate, ethanol, ice-cold water, 5 M NaCl, TES, total RNA, cuvettes for measuring absorbance at 260 nm, microfuge fitted with speed control, oligo(dT)18-30 cellulose, rotating wheel, water bath or heating block preset to 55°C and 65°C (see Note 2), DEPC-treated, RNA-free H2O.

2.2. Cap-Trapper Method

2.2.1. Purification of Oligonucleotide Primers

Acetonitrile, 10 M ammonium acetate, n-butanol, formamide gel-loading buffer without tracking dyes, formamide-tracking dye mixture, methanol:H2O solution, oligonucleotide elution buffer, TE (pH 8.0), crude preparation of synthetic oligonucleotide, Millex HV filter (Millipore, 0.45-μm pore size), parafilm or fluorescent thin-layer chromatographic plate, short body Sep-Pack classic columns (Waters corporation), polypropylene syringes (5-cc and 10-cc), hand-held ultraviolet lamp (260 nm), water bath or heating block preset to 55°C, denaturing polyacrylamide gel, sharp-clean scalpel (7).

2.2.1.1. Sorbitol/Trehalose-Enriched Synthesis of First-Strand cDNA

Actinomycin D, 1 M KCl, 1 M MgCl2, Tris-HCL (pH 8.3 at room temperature), RNase inhibitor, first strand cDNA primer (2 μg/μl), sterilized dH2O, 5x first-strand buffer, 0.1 M dithiothreitol (DTT), dNTP mixture (see Notes 3 and 4), 4.9 M sorbitol, saturated trehalose, RNase H SuperSCRIPT II (Invitrogen), reverse transcriptase (200 U/μl), T4 gene 32 (1.5 μg/μl), [α-32P]dGTP, EDTA.

2.2.1.2. GCI/Sorbitol/Trehalose-Enriched Synthesis of First-Strand cDNA

2x GC-I buffer, dNTP mixture, 4.9 M sorbitol, saturated trehalose, RNase H SuperSCRIPT II, reverse transcriptase (200 U/μl), T4 gene 32 (1.5 μg/μl), EDTA.
2.2.2. Measuring the Yield of cDNA
DE-81 filter paper (see Note 2), 0.5 M NaPi (pH 7.0), dH₂O, 70% ethanol.

2.2.3. Purification of First-Strand cDNA by Digestion with Proteinase K and Precipitation with CTAB
[α-³²P]dGTP, proteinase K (10 μg/μl), 5 M sodium chloride, CTAB/urea solution (see Section 2.1.2), guanidinium chloride (7 M), ethanol, 80% ethanol, dH₂O.

2.2.4. Oxidation and Biotinylation of Diol Groups of mRNA
1 M sodium acetate (pH 4.5), 100 mM NaIO₄, 46 μl of cDNA solution (from Section 3.2.4), 10% SDS, 5 M NaCl, isopropanol, 70% ethanol, sterilized dH₂O, 1 M sodium acetate (pH 6.1), 10% SDS, 10 mM biotin hydrazide long-arm (Vector Biosystem), dH₂O, absolute ethanol, 80% ethanol, 0.1 × TE [1 mM Tris (pH 7.0), 0.1 mM EDTA].

2.2.5. Cap-Trapping and Release of Full-Length cDNA
MPG-streptavidin beads, DNA-free tRNA, washing/binding solution [2 M NaCl, 50 mM EDTA (pH 8.0), RNase I, RNaseI buffer (e.g., Promega), dH₂O, 5 M NaCl, 2x with washing/binding solution containing 0.4% SDS and 50 μg/ml tRNA, 1x with washing/binding solution containing 10 mM Tris-HCL (pH 7.5), 0.2 mM EDTA, 40 μg/ml tRNA, 10 mM NaCl, and 20% glycerol, 1x with 50 μg/ml tRNA in dH₂O, 50 mM NaOH, 5 mM EDTA, 1 M Tris-HCL (pH 7.5), 50 mM NaOH, 5 mM EDTA.]

2.2.6. Adding a Priming Site to the 5’ End of First-Strand cDNA
1 M Tris-HCL (pH7.0), RNase I (10 U/μl), proteinase K to remove RNase I, glycogen, ethanol, microcon 100 (Millipore) for 40–60 minutes at 2,000 rpm, 0.1 × TE.

2.2.6.1. Treatment of Cap-Trapped First-Strand cDNA with RNaseI
1 M Tris-HCL (pH7.0), RNase I (10 U/μl), proteinase K to remove RNase I, glycogen, ethanol, microcon 100 (Millipore) for 40–60 minutes at 2,000 rpm, 0.1 × TE.

2.2.6.2. Fractionation of First cDNA
MicroSpin S-400 HR column (Amersham-Biosciences), 1.5 ml screw-cap microcentrifuge tube.

2.2.6.3. Single-Strand Linker Ligation Method
Follow the protocol from Section 3.2.1.2 to purify oligonucleotides; therefore, follow the materials section from 2.2.1.

2.3. Normalization and Subtraction
100 mM of NaCl, Solution II in TaKaRa DNA Ligation Kit, Solution II (in TaKaRa DNA Ligation Kit), 0.5 M EDTA, 10% SDS, 10 μg/μl of proteinase K, dH₂O, phenol-chloroform, chloroform, column buffer, MicroSpin S-300 HR columns (Amersham-Biosciences).

2.3.1. Preparation of Minilibrary and Drivers
10 M ammonium acetate, bovine serum albumin (2 mg/ml, Fraction V, Sigma), 1 M dithiothreitol, ethanol, phenol-chloroform (1:1, v/v), placental RNase inhibitor (20 units/μl), 3 M sodium acetate (pH 5.2), 10x transcription buffer [400 mM Tris-HCL (pH 7.5 at 37°C)], appropriate restriction enzymes (see Note 2),
rNTP containing rATP, rCTP, and rUTP (each at 5 mM), 0.5 mM rGTP, template DNA, [α<sup>32</sup>-P]rGTP (10 mCi/ml, sp. Act. 400–3000 Ci/m mole), 0.5 ml microfuge tubes, Sephadex G-50 spun column [equilibrated with 10 mM Tris-HCl (pH 7.5)], water bath or heating block preset to 40°C (see Note 2), SOB-agarose/ampicillin plates, resuspension solution supplied by the Promega Wizard kit (Promega), restriction enzyme site (see Note 2), proteinase K, chloroform, T7 or SP6 RNA polymerase (Invitrogen), DNaseI (RQ1, RNase-free, Promega), RNAeasy kit (QIAGEN).

2.3.2. Biotin Labeling of Normalizing/Subtracting RNA Drivers

Label IT<sup>®</sup> Biotin Labeling Kit (Mirus Bio corporation), dH<sub>2</sub>O (DNase-RNase-free), 5 M NaCl, 99% ethanol, 80% of ethanol, 1x Mirus labeling buffer A.

2.3.3. Normalization/Subtraction

Proteinase K, phenol:chloroform (1:1), blocking oligonucleotide (biotin-dG<sub>16</sub>), 0.3 M NaCl, ethanol, 80% ethanol, buffer solution, dH<sub>2</sub>O or 1x TE.

2.3.4. Posthybridization Capture

10x RNase I buffer (Promega), RNase I, phenol-chloroform (1:1), chloroform, ethanol, dH<sub>2</sub>O, tRNA, 1 M NaCl, 10 mM EDTA, wash buffer, streptavidin beads, 1 M NaCl, 10 mM EDTA, binding buffer.

2.4. Generating the Library

2.4.1. Synthesis of Second-Strand cDNA

Microcon-100 ultrafiltration, isopropanol, 500 mM NaCl, 80% ethanol, 0.1x TE RNase I buffer (Promega), RNase I (10 unites/μl), 0.2% SDS, Microcon-100, 0.1x TE 5x Buffer A, 5x Buffer B, dNTP mix (final 2.5 mM each), sterilized distilled water, [α<sup>32</sup>-P]dGTP, Elongase<sup>®</sup> enzyme mix (5 units/μl), proteinase K, phenol-chloroform (1:1), 0.1x TE, ethanol or isopropanol.

2.4.2. Cleaving the cDNA in Preparation of Cloning

10x restriction buffer, 10x bovine serum albumin, cDNA [either of SstI and XhoI (λ–FLCII) or XhoI and BamHI (λ–FLCI)], H<sub>2</sub>O, BamHI, XhoI, 10x restriction buffer, 10x bovine serum albumin, 25 units/μg of cDNA [either of SstI and XhoI (λ–FLCII) or XhoI and BamHI (λ–FLCI)], glycogen, 0.5 M EDTA, 10% SDS, proteinase K (10 μg/μl), phenol-chloroform (1:1), chloroform, ethanol, 10x restriction buffer (MBI buffer), 0.5 M EDTA, 1 μl of 10% SDS, 10 mg/ml proteinase K, phenol-chloroform (1:1), chloroform, 5 mM EDTA, 100 mM NaCl (see Note 2).

2.4.2.1. Single-Step cDNA Cleavage and Two-Step cDNA Cleavage

Chloroform, 0.5 M EDTA (pH 8.0), ethanol, phenol:chloroform(1:1, v/v), SDS (10%, w/v), 3 M sodium acetate (pH 7.0), TE (pH 7.6), 1 M tri-HCl (pH 8.5), calf intestinal alkaline phosphatase (CIP), 10x CIP dephosphorylation buffer, proteinase K, restriction enzymes (see Note 2), cDNA sample, waterbath or heating block preset to 56°C, 65°C, or 75°C.
2.4.3. Preparation of the Cloning Vector

2.4.3.1. Preparation of l-Arms for Cloning

Chloroform, dialysis buffer [10 mM NaCl, 50 mM Tris-HCL (pH 8.0), 0.5 M EDTA (pH 8.0), ethanol, phenol, phenol:chloroform (1:1, v/v), SDS (10% w/v), 3 M sodium acetate (pH 7.0), TE (pH 7.6 and pH 8.0), proteinase K, 0.7% agarose gel cast in 0.5 × TBE containing 0.5 µg/ml ethidium bromide, Sorvall SS-34 rotor, borosilicate Pasteur pipette or Shepherd’s crook, dialysis tubing (boiled), water bath or heating block preset to 56°C, bacteriophage λ particles in CsCl suspension, 10 mM Tris-HCL (pH 7.5)/10 mM MgCl₂, 10X ligation buffer, ligase (New England Biolabs), restriction enzymes, 5 M NaCl, FseI, PacI, 5 M NaCl, 10X NEB3 buffer, H₂O, SwaI, XhoI, SalI, BamHI, proteinase K, 0.1% SDS, 20 mM EDTA, H₂O, 0.6% low-melting point agarose gel (Seaplaque, FMC), StyI, electrophoresis buffer (1x TBE), β-agarase (New England Biolabs).

2.4.3.2. Cloning the cDNA

λ-arms, T4 DNA ligase and 10x buffer, high-efficient packaging mix, XL-Blue MRA (P2) bacteria, E. coli (strain C600), 150-mm petri dish, chloroform, SM, LB agar plates, 0.7% LB top agarose, Sorvall SS-34, water bath or heating block preset to 47°C, screw or snap-cap polypropylene tubes (see Note 2).

2.5. Bulk Excision of cDNA

2.5.1. In Vivo Solid-Phase Excision

BNN132 cell stock, LB-agar plate containing 50 µg/ml kanamycin, LB containing 10 mM MgSO₄ and 0.2% maltose, 10 mM MgSO₄, LB-ampicillin plate.

2.5.2. In Vivo Liquid-Phase Excision

BNN132 cells, LB medium, 10 mM MgSO₄, λ cDNA library (MOI = ~1), LB containing 100 µg/ml of ampicillin, Wizard Plus Midiprep DNA Purification System (Promega), cell suspension solution, cell lysis solution, neutralization solution, gauze, DNA-binding resin, the filter, ethanol/salt buffer, nuclease-free water, 5 M NaCl, isopropanol, 80% ethanol, H₂O.

2.5.3. In Vitro Cre-lox-Mediated Excision

Wizard Lambda Preps DNA Purification System (Promega), Cre-recombinase (Novagen), phenol-chloroform (1:1), DH10B cells.

3. Methods

3.1. Preparation of RNA

Cytoplasmic RNA for reverse transcription and polyA RNA must be prepared at this stage. The source of materials can be any organism as well as cells and tissues.
3.1.1. From Organisms

The methods listed here are only examples of those that can be used for RNA extraction from different organisms. In general, RNA from any kind of organism can be used for the full-length cDNA library construction. Although brain samples always give greater diversity of gene expression, you may use any kind of tissues of your interest. It is, of course, possible to obtain an organism from nature despite the fact that there are xenogenic issues to be considered.

3.1.1.1. Worm (Nematode)

1) Harvest worms and wash twice with M9 medium (7) in a RNase-free microcentrifuge tube as a pellet (1,000–10,000 worms depending on stage).

2) Freeze in liquid nitrogen and keep at –80°C until use (see Note 5).

3) Thaw the sample at 65°C for 10 minutes and add 1 ml of TRIZOL reagent per pellet.

4) Vortex repeatedly for up to 15 minutes (see Note 6).

5) Add 200 µl of chloroform per pellet and vortex the tube for 15 minutes. Leave at room temperature for 10 minutes.

6) Centrifuge the tube at maximum speed for 15 minutes at 4°C.

7) Transfer the upper clear layer into a new RNase-free microcentrifuge tube. Try not to take any white clumps containing DNA and protein.

8) Add 0.5 ml isopropanol per 1 ml of Trizol and mix by inverting several times.

9) Centrifuge the pellet at maximum speed for 30 minutes at 4°C.

10) Discard supernatant by decantation and wash the pellet with 75% ethanol with 5 minutes centrifugation.

11) Take all the liquid out and dry the pellet for 10 minutes but do not over-dry.

12) Dissolve the pellet to 100 µl H2O.

13) Use RNeasy Mini kit (QIAGEN) together with RNase-Free DNase Set to further purify the RNA following the manufacturer’s instruction (8) (see Note 5).

3.1.1.2. Insect

1) Dissect the organ of your interest in PBS and freeze immediately in liquid nitrogen. Keep it at –80°C if not immediately used (see Note 5).

2) Add 500 µl of Solution D and 50 µl of BME into a 1.5 ml microcentrifuge tube on ice.

3) Add ~100 mg sample into the tube and pestle the sample with a sterilized disposable tip.
4) Add 50 μl of 2 M NaOAc (pH 4.0), 500 μl water-saturated phenol, and 100 μl chloroform/isoamyl alcohol (24:1) to the tube. Vortex vigorously and incubate on ice for 10 minutes.

5) Centrifuge the tube at 10,000 rpm for 15 minutes at 4°C and transfer the upper clear phase to a new microcentrifuge tube.

6) Extract RNA with an equal volume of chloroform/isoamyl alcohol (24:1).

7) Add 500 μl of isopropanol and incubate sample on ice for 10 minutes.

8) Centrifuge the tube at 10,000 rpm for 15 minutes at 4°C and discard the supernatant.

9) Wash the pellet with 75% ethanol, air-dry the pellet but do not over-dry.

10) Dissolve the pellet to an appropriate amount of DEPC-treated H₂O for 1–2 μg/μl. Store the RNA at –80°C (9) (see Note 5).

3.1.3. Mammalian Brain

1) Collect total brain or a part of brain to be approximately 130 mg and transfer into an RNase-free sterile microcentrifuge tube.

2) Use 1 ml of TRIZOL reagent per the above-mentioned tube and follow the procedure steps 4–13 in 3.1.1.1 for RNA extraction.

3.1.2. From Whole Tissues

1. Prepare cells or tissue samples as the following:
   1) Dissect the desired tissues and place them immediately in liquid nitrogen.
   2) Place ~100 mg of the frozen tissue into a mortar with liquid nitrogen and reduce it to a powder using a pestle.
   3) Homogenize the tissue in a polypropylene snap-cap tube containing 3 ml of solution D for 15–30 seconds at room temperature using a polytron homogenizer.

2. Place the homogenate in a fresh polypropylene tube and add 0.1 ml of 2 M sodium acetate (pH 4.0), 1 ml of phenol, and 0.2 ml of chloroform-isoamyl alcohol per 1 ml Solution D. Mix the contents thoroughly by inversion after addition of each reagent.

3. Vortex the tube for 10 seconds and leave it on ice for 15 minutes.

4. Centrifuge the tube at 9,000 rpm for 20 minutes at 4°C and transfer the upper aqueous phase (containing RNA) to a fresh tube.

5. Add an equal volume of isopropanol to the extracted RNA, mix well, and leave the tube at –20°C for 1 hour or more.

6. Centrifuge the tube at 9,000 rpm for 30 minutes at 4°C.
7. Decant the isopropanol and dissolve RNA pellet in 0.3 ml of Solution D per 1 ml of this solution used in Step 1.

8. Transfer the solution to a microfuge tube and vortex. Add 1 volume of isopropanol and leave the tube at –20°C for 1 hour or more.

9. Centrifuge the tube at maximum speed for 10 minutes at 4°C in a microfuge to precipitate RNA. Wash the pellet twice with 75% ethanol and centrifuge again to remove any remaining ethanol with a sterile pipette tip. Let the tube stand with the lid open for a few minutes. (see Note 2.)

10. Dissolve RNA in 4 ml of dH2O.

11. CTAB is used to remove the polysaccharides from the RNA solution. Add 1.3 ml of 5 M NaCl to adjust the ionic strength of the sample, then add 16 ml of CTAB/urea solution (2, 7).

12. Leave the sample at room temperature for 2 minutes, then centrifuge the sample at 10,000 rpm for 5 minutes at room temperature to collect the CTAB-RNA pellet.

13. Discard the supernatant (carefully not to disrupt the pellet) by using the pipette tip and resuspend the pellet in 300 μl of 1.2 M NaCl by vortex mixing.

14. Add 8 ml of ethanol to the RNA solution, mix the solution well, and store for 1–2 hours at –20°C.

15. Centrifuge the solution at 10,000 rpm for 15 minutes at 4°C to recover RNA.

16. Rinse the pellet with 70% ethanol followed by centrifugation at 4°C.

17. Let the pellet dry in the air and resuspend in H2O (100–500 μl) to required concentration. Store RNA at –80°C (see Note 5).

### 3.1.3. From Cultured Cells

Cultured cells are always preferred over tissues for mRNA extraction since tissues are always contaminated with unprocessed transcripts including heterogeneous nuclear RNA (hnRNA) (10), which affects the reverse transcription process (11).

1. After the trypsin treatment, collect approximately 1 × 10^7 cells on ice and resuspend them in 2 ml of sterile ice-cold phosphate-buffered saline solution (PBS).

2. Centrifuge the cells at 3,000 rpm for 5 minutes at 4°C to collect the cells.

3. Discard the supernatant and resuspend the pellet in 2 ml of sterile ice-cold PBS.

4. Mix the RNA solution with 5 ml of ice-cold 0.5% NP-40 lysis buffer together with vanadyl ribonucleoside complex to inhibit RNase activity, and 5000 units of RNase inhibitor (RNase inhibitor is optional and can be omitted).
5. Invert the tube several times immediately to mix the solution and place the tube on ice for 3 minutes.
6. Centrifuge the tube at 10,000 rpm for 3 minutes at 4°C to precipitate the nuclei.
7. Transfer the supernatant into a new tube containing 2 volumes of CTAB/urea solution (12).
8. Centrifuge the tube at 7,500 rpm for 10 minutes at 4°C to collect RNA as pellets.
9. Discard the supernatant carefully by using the pipette tip and add 4 ml of 7 M guanidinium chloride.
10. Purify the solution with 1 volume of phenol:chloroform (1:1) solution and with 1 volume of chloroform.
11. Add 1 volume of isopropanol and incubate at –20°C overnight (13).
12. Centrifuge the tube at 10,000 rpm for 15 minutes at 4°C to collect RNA and wash the pellet with 10 ml of 70% ethanol.
13. Centrifuge the tube at 10,000 rpm for 15 minutes again, discard supernatant, and try not to dry the pellet completely.

3.1.4. Preparation of Poly(A)+ RNA from Total and Cytoplasmic RNA

1. Selection of poly(A)+ RNA (see Note 2)
   A) Selection of large amounts of poly(A)+ RNA
   1) Suspend 0.5–1.0 g of oligo(dT)-cellulose in 0.1 N NaOH and pour this in a DEPC-treated Dispo-column.
   2) Wash column with 3 column volumes of sterile DEPC-treated H2O.
   3) Dilute 2x column-loading buffer to 1x with sterile DEPC-treated H2O and wash the column with sterile 1x column-loading buffer until the pH of the effluents is <8.0 checked with pH paper.
   4) Dissolve the RNA in double-distilled, autoclaved H2O and heat the solution to 65°C for 5 minutes. Cool it to room temperature quickly and add 1 volume of 2x column-loading buffer.
   5) Apply the solution of RNA to the column and collect the flow-through. After applying all the solution, wash the column with 1 column volume of 1x column-loading buffer and collect the flow-through continuously.
   6) Heat the flow-through to 65°C for 5 minutes and reapply it to the top of the column for the flow-through to be collected again.
7) Wash the column with 5–10 column volumes of 1x column-loading buffer, collecting 1-ml fraction into a sterile microfuge tube.

8) Measure the quartz or disposable methacrylate cuvettes to measure the absorbance at 260 nm of a 1:20 dilution of each fraction collected from the column using 1x column-loading buffer as a blank.

9) Precipitate the fractions containing the majority of the OD$_{260}$ material by the addition of 2.5 volumes of ethanol.

10) Elute the poly(A)$^+$ RNA from the oligo(dT)-cellulose with 2–3 column volume of sterile, RNase-free elution buffer. The fraction is equivalent in size to $1/3$ to $1/2$ of the column volume.

11) Pool the fractions containing the eluted RNA.

B) Selection of small amounts of poly(A)$^+$ RNA

1) Adjust the volume of each sample to 600 µl with TES in sterile microfuge tubes. Heat the sample to 65°C for 10 minutes and place it on ice for a few minutes. Add 0.1 volume of 5 M NaCl to each sample.

2) Add 50 mg (500 µl) of equilibrated oligo(dT)-cellulose to each tube and incubate the closed tubes on a rotating wheel for 15 minutes at room temperature.

3) Centrifuge the tubes at 1500–2500 rpm for 2 minutes at room temperature and transfer the supernatants to fresh microfuge tubes. Keep the tubes on ice.

4) Add 1 ml of ice-cold absorption/washing buffer to the pellets from Step 3. Gently vortex the tubes and incubate the closed tubes on a rotating wheel for 2 minutes at room temperature.

5) Centrifuge the tubes at 1500–2500 rpm for 2 minutes at room temperature, discard the supernatants, and repeat Steps 4 and 5 twice.

6) Resuspend the pellets of oligo(dT) in 0.4 ml of pre-chilled double-distilled autoclaved H$_2$O by gentle vortexing. Centrifuge the tube immediately for 2 minutes at 4°C and aspirate the supernatant.

7) Resuspend the pellets of oligo(dT)-cellulose in 400 µl of double-distilled autoclaved H$_2$O. Keep the suspensions for 5 minutes at 55°C and then centrifuge the tubes for 2 minutes at 4°C.

8) Transfer the supernatant to a fresh tube and repeat Step 7 twice, pooling the recovered supernatants.
9) Add 0.2 volume of 10 M ammonium acetate and 2.5 volume of ethanol to the supernatants. Keep the tubes at –20°C for 30 minutes.

10) Centrifuge the tubes at maximum speed for 15 minutes at 4°C and discard the supernatants. Try not to disturb the pellet and wash the pellets with 70% ethanol. Centrifuge briefly and aspirate the supernatants completely and leave the open tubes inverted for a few minutes to evaporate the residual ethanol.

2. Measure the prepared poly(A)^+ RNA concentration for quantification using OD ratio at OD260:OD280 (see Note 2).

3. Check the quality of RNA using following criteria (see Note 2):
   a. RNA is pure when the reading of OD230:OD260 is lower than 0.5 and OD260:OD280 is higher than 1.8 (around 2.0).
   b. Run RNA on agarose gel and check the band intensities. Ratio between 28S and 18S RNA should be 1:2 without genomic DNA contamination.

Dissolve the pellet to appropriate amount of DEPC-treated H₂O for 1–2 µg/µl. Store RNA at –80°C (see Note 5).

3.2. Cap-Trapper Method

The cap-trapping method was developed for efficient full-length cDNA selection. To overcome a problem associated with cloning of fragments with size bias, PCR steps are eliminated in this method for efficient and high-throughput full-length cDNA production. Recognition sites for restriction enzymes in the primer-adapter play an important role in the polyA tail removal.

3.2.1. Design and Purification of Oligonucleotide Primers

3.2.1.1. Design of Oligonucleotide Primers

Recognition sites are used for digesting nonhemimethylated cDNA, and in turn, cleaving poly(A) tails (14).

The conditions are set as follows:

A primer-adapter carries one or more of the recognition sites AccI, BamHI, BstXI, Hinfl, PstI, SacI, SacII, Sall, SmaI, SstI, and XhoI (15), as well as GsuI (type IIS restriction enzyme) (14). These recognition sites do not work when cDNA is hemimethylated and are therefore used for poly(A) tail removal. This primer-adapter can be used for both first and second cDNA synthesis.

As an example, XhoI (CTCGAG) is used as follows for the recognition site 5’-(GA)₈AACTAGTCGAG(T)_₁₀MN-3’). M represents G, A, or C and N is any nucleotide. Different vectors have different restriction sites; therefore, the restriction enzymes can be replaced with any of the enzymes listed above as required (16).

3.2.1.2. Purification of Oligonucleotide Primers

1. Prepare a 10-µM solution of the crude oligonucleotide in Milli-Q in a sterile microfuge tube. Vortex the solution thoroughly and centrifuge the tube at maximum speed for 5 minutes at room temperature.
2. Transfer the supernatant to a fresh sterile microfuge tube and extract the solution three times in succession with 400 µl of n-butanol. Discard the upper phase after each extraction.

3. Set the tube in a centrifugal evaporator and dry the solution until it becomes a yellowish pellet and creamy-white powder.

4. Dissolve them in 200 µl of Milli-Q.

5. Estimate the amount of oligonucleotide in the preparation using following formulae:
   a) The millimolar extinction coefficient of the oligonucleotide ($\varepsilon$):

   \[ \varepsilon (\text{mM}^{-1}\text{cM}^{-1}) = A(15.2) + G(12.01) + C(7.05) + T(8.4) \]

   [A, G, C, and T are the number of times each nucleotide is represented]

   b) The concentration of the undiluted solution of oligonucleotide ($c$):

   \[ c = \frac{\text{OD}_{260}(1000)}{\varepsilon} \]

6. Pour a denaturing polyacrylamide gel (see Note 2) of the appropriate concentration (according to the size of oligonucleotides, see Note 2:10.14) with the loading slots ~1 cm in length.

7. Run the gel at constant wattage (50–70 W) for ~45 minutes and then, turn off the power supply.

8. Mix an equal volume of formamide gel-loading buffer lacking dyes to the oligonucleotide solution and vortex. Heat the tube to 55°C for 5 minutes. While heating the sample, flush out the urea from the wells with 1xTBE.

9. Load ~2 OD$_{260}$ units of oligonucleotide mixed with the buffer onto the slots. Load 5 µl of formamide-tracking dye mixture into an unused slot.

10. Run the gel at 1500 V until the oligonucleotide has migrated approximately two thirds of the gel length.

11. Place the gel mold with the smaller plate upper most on plastic backed protective bench paper. Wait till the gel is <37°C before proceeding.

12. Remove the excess pieces and slowly and gently pry apart the place of the mold from the gel. The gel should remain attached on the longer glass plate.

13. Place a piece of cling-film on the gel, turn the glass plate over, and transfer the gel onto the film. Place a piece of Parafilm under the gel where the oligonucleotide is predicted to be.
14. Visualize the oligonucleotide on the gel using a hand-held UV lamp at 260 nm and excise each DNA band (the slowest-migrating) with a sharp, clean scalpel and put them into microfuge tubes. Ideally, the gels can be spread to three or four microfuge tubes.

15. Add 1 ml of oligonucleotide elution buffer to each tube and crush the gel piece with a sterile pipette tip using a circular motion and pressing the fragments of gel against the sides of the tubes. Seal the tubes tightly and incubate them for 12 hours at 37°C in a shaker incubator.

16. Centrifuge the tubes at maximum speed for 5 minutes at room temperature and pool the supernatants. Transfer them to a 5-cc disposable syringe and extract them through a Millex HV filter. Collect the flow-through in a 15-ml polypropylene tube.

17. Prepare a Sep-Pak C18 reversed-phase cartridge (Waters Corporation) as follows:
   a) Attach the barrel of a disposable 10-cc polypropylene syringe to the longer end of Sep-Pak C18 cartridge.
   b) Add 10 ml of acetonitrile to the barrel and slowly push it through the cartridge with the plunger of the syringe.
   c) Remove the syringe from the cartridge and then take plunger out of the syringe.
   d) Add 10 ml of Milli-Q to the barrel and slowly push it through the column with the plunger. Repeat Step c.
   e) Add 2 ml of 10 mM ammonium acetate to the barrel and push it slowly through the cartridge. Repeat Step c. Column is ready to use.

18. Apply the solution containing the gel-purified oligonucleotide from Step 16 to the barrel and slowly push it through the cartridge with the plunger. Collect the flow-through into a sterile 50-ml polypropylene tube. Repeat Step 17-c.

19. To wash the cartridge, apply 10 ml of H2O to the barrel and push it slowly through the column with the plunger. Repeat this washing step twice more.

20. Apply three aliquots of 1 ml of methanol:H2O solution to elute the bound oligonucleotide. Repeat Step 17-c after each elution. Collect each flow-through in a separate microfuge tube. Measure OD260 of the solution in each of the three microfuge tubes using the methanol:H2O solution as a blank. First fraction should have more than 90% of the oligonucleotide applied to the column.

21. Place the tube in a centrifugal evaporator to evaporate the solution to dryness.
22. Use 200 μl of H₂O or TE (pH 8.0) to dissolve the oligonucleotide.

23. Place 5 μl of the solution in a cuvette containing 995 μl of H₂O and mix well. Measure the OD₂₆₀ of the diluted sample and calculate the amount of oligonucleotide present in the total solution in Step 23 using the formulas in Step 5 of this protocol.

24. Adjust the concentration to 2 μg/μl.

3.2.2. First-Strand cDNA Synthesis

Reverse transcription is required for cDNA synthesis, and trehalose and sorbitol help to maintain the normal activity of the reverse transcriptase (I). Together with trehalose, sorbitol is employed to improve the reaction by thermostabilizing and activating trehalose mainly by reducing the formation of secondary structures in mRNA templates and thereby allowing the reverse transcriptase to move freely (17). Additionally, the use of the T4 bacteriophage gene 32 protein (T4gp32) in the reverse transcription reaction together with an RNase-deficient Moloney leukemia virus reverse transcriptase (MoLV-RT) can also enhance the yield of in vitro transcripts (12).

3.2.2.1. Sorbitol/Trehalose-Enriched Synthesis of First-Strand cDNA (see Note 7)

1. Set up 3 tubes (A, B and C) containing the following ingredients for each tube:

   Tube A:
   - mRNA (or total RNA) 2.5–25 μg
   - First-strand cDNA primer (2 μg/μl) 7 μl
   - Sterilized dH₂O to give a total volume of 21.3 μl

   (See Notes 7 and 8)

2. Heat the tube A at 65°C for 10 minutes. During this process, the tubes B and C need to be prepared quickly.

   Tube B:
   - 5x first-strand buffer 28.6 μl
   - 0.1 M dithiothreitol (DTT) 11 μl
   - dNTP mixture (see Note 3) 9.3 μl
   - 4.9 M sorbitol 55.4 μl
   - saturated trehalose 23.2 μl
   - RNase H SuperSCRIPT II Reverse transcriptase (200U/μl) 15 μl
T4 gene 32 (1.5 μg/μl) \( (see \textbf{Note 9}) \) 15.8 μl

The total volume of the mixture 158.3 μl

(Add RNase-free double-distilled water if required, to achieve the final volume.)

Tube C:

\( [\alpha^{-32}\text{P}]\text{dTgp} \) 1–1.5 μl

3. Set up the thermal cycler as follows:
   40°C for 2 minutes
   50°C for 2 minutes
   56°C for 60 minutes
   4°C for infinite.
   \( (see \textbf{Notes 10 and 11}) \)

4. Place all the 3 tubes onto the thermal cycler and start the cycle. Transfer the contents of tube B to tube A quickly when the temperature reaches 42°C, then 40 μl of the A and B mixture to tube C. Complete the cycling program \( (13, 17) \).

5. Add EDTA to give a final concentration of 10 mM to terminate the reaction \( (13) \).

---

3.2.2.2. GCI/Sorbitol/Trehalose-Enriched Synthesis of First-Strand cDNA

1. Set up 3 tubes (A, B, and C). Tubes A and C are prepared as in Section 3.2.2.1.

   Tube B:

   2x GC-I buffer \( (see \textbf{Note 12}) \) 75 μl
   dNTP mixture \( (see \textbf{Note 4}) \) 4 μl
   4.9 M sorbitol 20 μl
   Saturated trehalose 10 μl
   RNase H SuperSCRIPT II
   Reverse Transcriptase (200 U/μl) 15 μl
   T4 gene 32 (1.5 μg/μl) \( (see \textbf{Note 9}) \) 13.8 μl
   The total volume of the mixture 137.8 μl

2. Set up the thermal cycler as follows:
   42°C for 30 minutes
   50°C for 10 minutes
   55°C for 10 minutes
   4°C for infinite.
3. Follow the Step 4 of Section 3.2.2.1 for hot start and Note 10 for cold start (see Note 13).

4. Add EDTA to a final concentration of 10 mM to terminate the reaction (13).

3.2.3. Measuring the Yield of cDNA

1. Take 0.5 μl from tube C at the end of the reaction and spot it onto a small square of DE-81 paper. Keep another 0.5 μl of the reaction from tube C in a separate tube for alkaline gel analysis.

2. Measure the radioactivity before and after three 10-minute washings with 50 ml of 0.5 M NaPi (pH 7.0), followed by brief washing with dH₂O and 70% ethanol and quick air drying.

3. Calculate the total yield using following formulas:

\[
\text{Total dNTP} (\mu g) \times \left[ \% \text{ of incorporation} \right]/100 = \mu g \text{ of first-strand cDNA synthesized}
\]

* Example of this using the concentration used in Section 3.2.2.1

\[
\text{Total dNTP (moles)} = 9.3 \times 10^{-6} \text{ (liter)} \times 40 \times 10^{-3} \text{ (moles/liter)} = 3.72 \times 10^{-7} \text{ moles}
\]

\[
\text{Total dNTP (μg)} = 340 \text{ (average molecular weight of a residue)} \times 3.72 \times 10^{-7} = 126.4 \mu g
\]

Therefore, the yield is calculated as:

\[
126.4 \mu g \times \left[ \% \text{ of incorporation} \right]/100 = \mu g \text{ of first-strand cDNA synthesized}
\]

4. Analyze the size of the radiolabeled first-strand cDNA using alkaline gel analysis. The cDNA will be seen 2–3 Kb in alkaline gel analysis as a positive result. If not, it is possible that there was RNase contamination, decreased enzymatic activity, or failure of mixing the viscous components (see Note 2).

3.2.4. Purification of First-Strand cDNA by Digestion with Proteinase K and Precipitation with CTAB

This procedure removes all the unwanted RNA (if not purified) and the high molecular-weight precipitates from sorbitol treatment as well as the peptides and other products created by digestion of proteins with proteinase K and unincorporated \([\alpha^{-32P}]dNTP\).

1. Transfer the residue of the first-stranded cDNA reaction (the mixture of tubes A+B) and the cDNA mixture containing \([\alpha^{-32P}]dTTP\) (tube C) to a microcentrifuge tube.

2. Add 2 μl of proteinase K (10 μg/μl). Incubate the reaction for 15 minutes to 1 hour at 45°C to remove all the proteins contained in the cDNA mixture.
3. Add:

5 M sodium chloride 32 μl
CTAB/urea solution (see Section 2.1.2. for composition) 320 μl
Leave the sample for 10 minutes at room temperature (14).

4. Centrifuge the tube containing the mixture for 15 minutes at 10,000 rpm at room temperature and discard the supernatant (see Note 14).

5. Add 100 μl of guanidinium chloride (7 M) and dissolve the pellet.

6. Add 250 μl of ethanol and store the tube on ice at –20°C or –80°C for more than 30 minutes (see Note 15).

7. Centrifuge the tube at 13,000 rpm for 10 minutes and discard the supernatant. Add 800 μl of 80% ethanol and centrifuge the sample at 13,000 rpm for 3 minutes to wash the pellet. Discard the supernatant and repeat the 800 μl of 80% ethanol wash and discard the supernatant leaving the pellet.

8. Resuspend the pellet completely in appropriate volume of dH2O (46 μl is appropriate in this case, which is proportional to the first-strand cDNA reaction) (13).

3.2.5. Oxidation and Biotinylation of Diol Groups of mRNA

Oxidation procedure is required for the diol group of the cap and 3'-end of RNA before biotinylation.

3.2.5.1. Oxidation of the Diol Groups of mRNA

1. Add 3.3 μl of 1 M sodium acetate (pH 4.5) and 0.5 μl of 100 mM NaIO4 to 46 μl of cDNA solution (purified cDNA from Section 3.2.4).

2. Incubate the sample on ice in the dark for 45 minutes.

3. Add 1 μl of 80% glycerol and vortex the tube to stop the reaction.

4. Add 61 μl of isopropanol, 0.5 μl of 10% SDS, and 11 μl of 5 M NaCl. Incubate the sample in the dark either at –20 or –80°C for 45 minutes at 13,000 rpm.

5. Remove the supernatant carefully and check the amount of radioactivity before discarding.

6. Add 500 μl of 80% ethanol and centrifuge the sample at 13,000 rpm for 2–3 minutes at 4°C.

7. Discard the supernatant and rinse the cDNA again, twice with 50 μl of 70% ethanol and centrifugation at 15,000 rpm at 4°C for 3 minutes (see Note 16).

8. Resuspend in 50 μl sterilized dH2O.
3.2.5.2. Biotinylation of the Cap

1. Add 5 μl of 1 M sodium acetate (pH 6.1), 5 μl 10% SDS, and 150 μl of 10 mM biotin hydrazide long-arm to tube C prepared in the previous step containing cDNA in 50 μl of dH2O.

2. Incubate the sample overnight (10–16 hours) at room temperature (22–26°C).

3. Add 75 μl of 1 M sodium acetate (pH 6.1), 5 μl 5 M NaCl, and 750 μl absolute ethanol and mix gently but thoroughly.

4. Incubate the mixture on ice for 1 hour or at –20 to –80°C for 30 minutes.

5. Centrifuge the tube at 15,000 rpm for 10 minutes at 4°C.

6. Discard the supernatant leaving the pellet.

7. Wash the pellet once with 70% ethanol and once with 80% ethanol using centrifugation at 15,000 rpm (see Note 16).

8. Resuspend the cDNA pellet in 70 μl of 0.1x TE [1 mM Tris (pH 7.0), 0.1 mM EDTA].

3.2.6. Cap-Trapping and Release of Full-length cDNA

Using reverse transcription to synthesize cDNA, RNA-cDNA hybrids form. However, a portion of the hybrids do not carry full-length cDNA, instead they are still attached as partial fragments to intact mRNA. To select only full-length hybrids, RNaseI treatment digests biotin-attached poly(A) tail as well as RNAs not attached to cDNA, which carry incomplete cDNA. The magnetic beads are then used to capture and bind the biotinylated cap at the 5’-end of mRNA. RNA hydrolysis then breaks mRNA carrying biotinylated-caps, leaving full-length cDNA unharmed.

3.2.6.1. Preparation of Streptavidin Beads

1. Mix 500 μl of MPG-streptavidin beads and 100 μg DNA-free tRNA.

2. Incubate the mixture on ice for 30 minutes with occasional mixing.

3. Use a magnetic stand to separate beads for 3 minutes.

4. Remove the supernatant.

5. Wash the beads with 500 μl of washing(binding) solution stated in the materials section.

6. Resuspend the beads in 400 μl of washing(binding) solution.

3.2.6.2. RNaseI Digestion and Capture of Full-Length cDNA

1. Add appropriate units (1U per 1 mg of starting mRNA) of RNase I to 70 μl of cDNA solution, as well as the buffer supplied by the manufacturer (e.g., Promega) and dH2O, to a total volume of 200 μl.

2. Incubate the sample at 37°C for 15 minutes.

3. Place the sample on ice and add 100 μg tRNA and 100 μl 5 M NaCl to the tube.
4. Add the streptavidin beads suspended in 400 μl of washing/binding solution from the previous step to the tube and mix well.

5. Gently rotate the tube at room temperature for 30 minutes.

6. Let the tube stand in a magnetic rack for 3 minutes, and then discard the supernatant.

7. Gently wash the beads as follows with a wash volume of 0.5 ml.
   I. 2x with washing/binding solution containing 0.4% SDS and 50 μg/ml tRNA.
   II. 1x with washing/binding solution containing 10 mM Tris-HCL (pH 7.5), 0.2 mM EDTA, 40 μg/ml tRNA, 10 mM NaCl, and 20% glycerol.
   III. 1x with 50 μg/ml tRNA in dH₂O (see Notes 17 and 18).

3.2.6.3. cDNA Elution

1. Add 50 μl of 50 mM NaOH and 5 mM EDTA to the tube and incubate the mixture for 10 minutes at room temperature with occasional mixing.

2. Remove the beads magnetically and transfer the supernatant to a new tube containing 50 μl 1 M Tris-HCL (pH 7.5) on ice.

3. Repeat Steps 1 and 2 using 50 μl of 50 mM NaOH and 50 μl of 5 mM EDTA to recover 80–90% of cDNA, check the amount of cDNA using a hand-held monitor by monitoring the radioactivity.

4. Keep the tube on ice to prevent rehybridization of isolated cDNA and RNA traces (see Note 19).

3.2.7. Adding a Priming Site to the 5' End of First-Strand cDNA

After the extraction of first single-strand full-length cDNA, we apply single-strand linker ligation method (SSLLM) for second-strand cDNA synthesis. This method uses random 6-bp protruding ends attachment to single-stranded cDNA for second-strand cDNA synthesis, which simplifies the sequencing and protein translation step due to the elimination of the GC tail (14).

3.2.7.1. Treatment of Cap-Trapped First-Strand cDNA with RNaseI

This step completely removes residual RNA to obtain purified single-stranded cDNA.

1. Add 100 μl of 1 M Tris-HCL (pH7.0) and 1 μl of RNase I (10U/μl) to the recovered cDNA on ice; mix immediately after the addition.

2. Incubate the sample at 37°C for 10 minutes.

3. Use proteinase K to remove RNase I, then, precipitate cDNA with ethanol after phenol chloroform extraction and back-extract the aqueous phase.

4. Add 2–3 μg of glycogen and ethanol-precipitate the sample in a siliconized tube.
5. Concentrate the cDNA by using one round of ultrafiltration with a Microcon 100 (Millipore) for 40–60 minutes at 2,000 rpm at room temperature.

6. Redissolve the cDNA in 20 μl of 0.1 X TE.

3.2.7.2. Fractionation of First cDNA

Before going into the next step, the first-strand cDNA needs to be free of any trace of the primer-adapter used in first-strand synthesis as well as any residual RNAs. It is possible to eliminate these traces by using a MicroSpin S-400 HR column (Amersham-Biosciences). Follow manufacturer’s instruction as provided:

1. Resuspend the sample in the column by vortexing.
2. Loosen the cap one-fourth turn and snap off the bottom closure.
3. Place the column in a 1.5-ml screw-cap microcentrifuge tube for support. Alternatively, cut the cap from a flip-top tube and use this tube as a support.
4. Prespin the column at 3,000 rpm for 1 minute.
5. Place a column in a new 1.5-ml tube, remove and discard the cap, and slowly apply the sample to the top-center of the resin, being careful not to disturb the bead. Spin the column at 3,000 rpm for 2 minutes. The purified sample is collected in the bottom of the tube.

3.2.7.3. The Single-Strand Linker Ligation Method

3.2.7.3.1. Linker-Primers Preparation

1. Follow the protocol from Section 3.2.1.2 for purifying each oligonucleotide.
2. Mix GN₅, 5′, and N₆ linkers at a ratio of 4:5:1 to produce 2 μg/μl of total oligonucleotides.
3. Add 100 mM of NaCl to the linker-primers.
4. For annealing, incubate the sample at 65°C for 5 minutes, 45°C for 5 minutes, 37°C for 10 minutes, and 25°C for 10 minutes. At the end of this thermal cycle, the linker-primer mixture can be stored at –20°C.

3.2.7.3.2. Ligation of First-Strand cDNA

1. Use 2 μg of linker-primer for 1 μg of cDNA at most. Prepare 3 μl of cDNA solution containing 1 μg cDNA dissolved in dH₂O.
2. Incubate cDNA solution at 65°C for 10 minutes.
3. Place the tube containing cDNA on ice immediately after the incubation step.
4. Mix 2 μl of linker-primer mixture and 3 μl of cDNA solution on ice.
5. Add 5 μl of Solution II in TaKaRa DNA Ligation Kit.
6. Add 10 μl of Solution I in TaKaRa DNA Ligation Kit.
7. Incubate the reaction mixture at 10–16°C overnight (at least 10 hours).

8. Add 1 μl of 0.5 M EDTA, 1 μl of 10% SDS, 1 μl of 10 μg/μl proteinase K, and 10 μl of dH2O, then, incubate the solution at 45°C for 15 minutes to stop the reaction.

9. Use phenol-chloroform extraction, chloroform extraction, and back-extraction with 60 μl of column buffer to extract the ligation products.

10. After the ligation steps described above, the excess linker needs to be removed. To remove the excess linkers, use MicroSpin S-300 HR columns (Amersham-Biosciences), following the manufacturer’s instruction (same as MicroSpin S-400 HR column) (see Note 20).

3.3. Normalization and Subtraction

The major problems associated with generating full-length cDNA library are the size bias and expressed gene frequency. Shorter cDNA is always preferred in cloning and propagation while highly expressed genes will produce a high quantity of redundant mRNA transcripts. Normalization and subtraction are designed to overcome these obstacles, removing over abundant cDNAs as well as redundant sequences that already exist in a previously generated library. In our protocol, full-length cDNA is normalized and subtracted before cloning in order to amplify only novel cDNAs/cDNAs of interest. However, in the cases where you have less than 500 ng of cDNA, for purposes of size selection or small/very small-scale projects, skip this procedure (3.3) and go to the second-strand synthesis step.

3.3.1. Preparation of Minilibrary and Drivers

3.3.1.1. Subtraction Drivers

1. Amplify 20,000–50,000 colonies from minilibraries derived from 1,000–2,000 plasmid clones by growing them on SOB-ampicillin-agarose plates. These clones represent the highly expressed fraction of previously prepared libraries generated (e.g., from previous normalization experiments). This produces the minilibrary stock to create a bottleneck (see Note 21).

2. Incubate them overnight at 37°C.

3. Scrape bacterial cells from the plate into the resuspension solution supplied by the Promega Wizard kit. Extract the plasmid DNA following the manufacturer’s instructions.

4. Check the concentration of plasmid DNA by photo spectrometry and check the size by agarose gel electrophoresis.

5. Cleave the plasmid DNA from the 3’ end of the insert with the least frequently occurring restriction enzyme site available.

6. Use proteinase K to purify the linear insert, then purify the fragment with phenol chloroform extraction followed by chloroform extraction and ethanol precipitation.
3.3.1.2. RNA Driver Preparation

1. Prepare RNA drivers as follows:

1) Use an appropriate restriction enzyme to completely digest the superhelical plasmid DNA to prepare 5 pmoles of linear template DNA. Remove the 3’ protrusion of the digested DNA if any.

2) Purify the template DNA using phenol:chloroform and ethanol. Dissolve the DNA in H$_2$O to be 100 nM.

3) Warm the DNA solution to room temperature and add the components to a 0.5-ml microfuge tube in the following order at room temperature:

\[
\begin{align*}
\text{(pre-warm) Template DNA} & \quad 0.2 \text{ pmoles} \\
\text{(pre-warm) RNase-free H$_2$O} & \quad \text{to 6 µl} \\
\text{(pre-warm) 5 mM rNTP solution} & \quad 2 \text{ µl} \\
\text{(pre-warm) 100 mM dithiothreitol} & \quad 2 \text{ µl} \\
\text{(pre-warm) 10x transcription buffer} & \quad 2 \text{ µl} \\
\text{(pre-warm) 2 mg/ml bovine serum albumin} & \quad 1 \text{ µl} \\
\text{10 mCi/ml}[^{32}\text{P}]\text{rGTP} & \quad 5 \text{ µl}
\end{align*}
\]

Mix the components by gently tapping the tube. Then add:

- placental RNase inhibitor (10 units) 1 µl
- T7 or SP6 RNA polymerase (Invitrogen) 1 µl (~10 units)

Mix the components by gently tapping the tube again. Centrifuge the tube briefly to collect the contents at the bottom of the tube. Incubate the tube at a temperature appropriate for the polymerase for 1–2 hours.

4) Add 2 µl of 0.5 mM rGTP and incubate the tube for 1 hour at the temperature appropriate for the polymerase.

5) Add 1 µl of 1 mg/ml RNase-free pancreatic DNase I to the reaction tube to terminate the transcription. Mix the components by tapping the tube. Incubate the tube for 15 minutes at 37°C.

6) Add 100 µl of RNase-free H$_2$O, and extract the RNA with phenol:chloroform extraction.

7) Transfer the aqueous phase to a fresh microfuge tube and separate the radiolabeled RNA from undesired small RNAs and rNTPs by ethanol precipitation or spun-column chromatography (using Sephadex G-50 spun column) or Gel electrophoresis (see Notes 2 and 22). Either T7 or SP6 RNA polymerase (Invitrogen) can be
used depending on the site of cleavage. Follow the manufacturer’s instruction to prepare 5–10 μg of linearized plasmid or amplified PCR product.

2. Digest residual DNA with 1-2 μl of DNaseI (RQ1, RNase-free, Promega) at 37°C for 30 minutes followed by proteinase K digestion, phenol-chloroform extraction, and chloroform extraction. Precipitate the RNA with ethanol at the end.

3. Purify the RNA once again, using RNeasy kit (QIAGEN) following manufacturer’s instructions.

### 3.3.2. Biotin Labeling of Normalizing/Subtracting RNA Drivers

1. Evaluate the amount of RNA:
\[(\text{Target mRNA}) + (\text{Synthesized RNA from minilibraries}) = X \, \mu g\]

2. Use Label IT® Biotin Labeling Kit provided by Mirus Bio Corporation. Mix as follows:

<table>
<thead>
<tr>
<th>Component</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA mixture</td>
<td>X μg</td>
</tr>
<tr>
<td>Label IT reagent</td>
<td>X μl (1 μl of Label IT reagent for 1 μg of RNA)</td>
</tr>
<tr>
<td>10x Mirus labeling buffer A</td>
<td>1 μl</td>
</tr>
<tr>
<td>dH₂O (DNase-RNase-free)</td>
<td>to a total volume of 10 μl</td>
</tr>
</tbody>
</table>

Adjust the volume as appropriate.

3. Incubate the reaction at 37°C for 1 hour.

4. Precipitate the biotinylated RNA by adding 1/20 volume of 5 M NaCl and 2 volumes of 99% ethanol.

5. Wash the pellet with 80% of ethanol.

6. Centrifuge the tube at maximum speed for 3 minutes at 4°C.

7. Discard the supernatant by pipetting and dry the pellet in air.

8. Resuspend in 20 μl of 1x Mirus labeling buffer A. *(see Note 23)*

### 3.3.3. Normalization/Subtraction Procedure

1. Deproteinate the RNA drivers and cDNA using proteinase K treatment, followed by phenol/chloroform extraction.

2. Calculate the RoT value to be 5–500 *(see Note 24).*

3. In a 0.5-ml tube, mix SSSLM-treated single-strand cDNA, biotinylated RNA drivers, and blocking oligonucleotide (biotin-dG₁₆).

4. Precipitate the mixture with 0.3 M NaCl and 2 volumes of ethanol, following standard procedure; see Molecular Cloning *(18).*
5. Wash the pellet with 80% ethanol, centrifuge the sample at maximum speed for 3 minutes at 4°C, and remove the supernatant. Repeat washing again.

6. Dry the pellet in air but not completely, leaving it damp.

7. Use liquid scintillation spectroscopy to measure the amount of radioactivity in the tube.

8. Resuspend the pellet in 5 μl of buffer solution; dissolve the DNA completely in the buffer (see Note 24).

9. Place the tube at 42°C for the time required (see Note 24).

10. Add 2.5 volumes of cold ethanol.

11. Incubate the solution at –20°C for 30 minutes (can be placed on ice).

12. Centrifuge the sample at maximum speed for 10 minutes at 4°C.

13. Discard the supernatant, trying not to disrupt the pellet.

14. Centrifuge the sample at maximum speed for 3 minutes at 4°C.

15. Wash the pellet with 80% ethanol.

16. Centrifuge the sample at maximum speed for 3 minutes at 4°C.

17. Discard the ethanol; dry the pellet until damp.

18. Redissolve the pellet in 45 μl of dH2O or 1x TE.

19. Store the sample on ice until the next step.

### 3.3.4. Posthybridization Capture

Using RNase I digestion, all the undesired hybrids such as mismatched annealing products and small fragments of RNA attached to cDNA will be digested, increasing the specificity of hybridization. Prepare the magnetic beads simultaneously; see below.

#### 3.3.4.1. Digestion of Undesired Hybrids

1. Add 5 μl of 10x RNase I buffer (Promega) and 0.5 unit RNase I per 10 μg of driver RNA to the sample.

2. Place the sample at 37°C for 10 minutes.

3. Heat the mixture to 65°C for 10 minutes.

4. Purify the sample with phenol-chloroform extraction followed by chloroform extraction. Precipitate the sample with 2 volumes of ethanol with centrifugation.

5. Redissolve the pellet in 10 μl of dH2O.

#### 3.3.4.2. Preparation of the Magnetic Beads

Carry out the procedure to prepare the magnetic beads at the same time as the steps above. The required amount of CPG beads is 50 μl per microgram of biotinylated mRNA.

1. Add 10 μg of tRNA for every 50 μl of beads. This reduces the nonspecific binding of cDNA to the beads.
2. Place the beads at room temperature for 10 to 20 minutes, and then on ice for 30 to 60 minutes with occasional shaking.

3. Move the beads to the magnetic stand for 3 minutes and discard the aqueous phase.

4. Wash the beads three times with 1 M NaCl and 10 mM EDTA. Use equal volume of wash buffer as the starting volume of the beads.

5. Resuspend the beads in one volume of 1 M NaCl and 10 mM EDTA equivalent to the starting volume.

3.3.4.3. Removal of Abundant cDNA/Unwanted cDNA

1. Mix the washed beads and the cDNA/driver mixture.

2. Incubate the above mixture at room temperature for 15 minutes with occasional gentle mixing.

3. Place the mixture in a magnetic stand for 3 minutes.

4. Recover the supernatant containing the normalized/subtracted cDNA.

5. Wash the beads with 100 µl of binding buffer.

3.4. Generating the Library

3.4.1. Synthesis of Second-Strand cDNA

Before second-strand cDNA synthesis, remove residual RNA and purify the first-strand single-stranded cDNA. Use one of the following methods for this procedure.

3.4.1.1. Preparation for Second-Strand cDNA Synthesis

1. Microcon-100 ultrafiltration: Using Microcon-100 ultrafiltration, reduce the cDNA solution volume following the manufacturer’s instruction.

2. Isopropanol precipitation: Use 500 mM NaCl in the cDNA solution. Add 1 volume of isopropanol and leave the tube on ice for 30–60 minutes. Precipitate the sample by centrifugation at maximum speed for 15 minutes at 4°C. Remove the supernatant carefully. Wash the pellet with 100 µl of 80% ethanol. Centrifuge the tube for 2–3 minutes and discard the supernatant. Dissolve the pellet in 44 µl of 0.1x TE.

3.4.1.2. Purification of the Full-Length cDNA to Remove Traces of RNA

1. Add 5 µl of RNase I buffer (Promega) and 1 µl of RNase I (10 units/µl) to the cDNA solution.

2. Incubate the tube at 37°C for 15 minutes.

3. Add 350 µl of 0.2% SDS and leave the tube at room temperature for 2 minutes.

4. Transfer the mixture to a Microcon-100 and add 400 µl of 0.1x TE.

5. Centrifuge the tube at 2,000 rpm for 20 minutes and discard the filtrate.
6. Repeat Step 5 three times.

7. Invert the filter and centrifuge at 9,000 rpm for 1 minute (see Note 25).

A subsequent step requiring ligation of the first-strand cDNA with linker-primers is the second-strand cDNA synthesis.

### 3.4.1.3. Second-Strand cDNA Synthesis

1. Mix the following in tube A and tube B.

   **Tube A:**
   - template cDNA: X µl (up to 21 µl)
   - 5x Buffer A: 4.8 µl
   - 5x Buffer B: 6 µl
   - dNTP mix (final 2.5 mM each): 6 µl
   - Sterilized distilled water: to a total volume of 45 µl

   **Total volume:** 45 µl.

   **Tube B:**
   - [α-32P]dGTP: 0.5 µl.

2. Place the tube at 65°C for 1 minute.

3. Add 15 µl of Elongase Enzyme Mix (5 units/µl) and mix quickly by vortex.

4. Transfer 3 µl of tube A solution to tube B, containing 0.5 µl of [α-32P]dGTP at 65°C, quickly.

5. Set the thermal cycle as follows:
   - 65°C for 5 minutes
   - 68°C for 30 minutes
   - 72°C for 10 minutes
   - 4°C for ∞.

6. Measure the amount of second-strand reaction yield using DE-81 adsorption (follow Steps in 3.2.3). (see Note 26.)

7. Use proteinase K treatment followed by phenol-chloroform extraction and back-extraction with 0.1x TE. Precipitate the cDNA with either ethanol or isopropanol.

### 3.4.2. Cleaving the cDNA in Preparation of Cloning

#### 3.4.2.1. Single-Step cDNA Cleavage

1. Use the appropriate units of restriction enzymes for each digestion based on the yield of second-strand cDNA, assuming that no significant amount of double-strand cDNA has been lost during purification.

2. Add to the dissolved cDNA the following: 3 µl of 10x restriction buffer and 3 µl of 10x bovine serum albumin,
25 units/µg of cDNA [either of SstI and XhoI (λ–FLCII) or XhoI and BamHI (λ–FLCI)], and H₂O to a final volume of 30 µl.

3. Place the tube with the above mixture at 37°C for 1 hour.

4. Proceed with either dephosphorylation or purification before the cloning step (see Note 27).

### 3.4.2.2. Two-Steps cDNA Cleavage (Preferred for cDNAs Cloned by SSLLM)

1. Use 25 units of BamHI per microgram of cDNA and 50 units of XhoI per microgram of cDNA.

2. To the dissolved cDNA, add 3 µl of 10x restriction buffer and 3 µl of 10x bovine serum albumin, 25 units/µg of cDNA [either of SstI and XhoI (λ–FLCII) or XhoI and BamHI (λ–FLCI)], and H₂O to a final volume of 30 µl.

3. Place the mixture at 37°C for 1 hour.

4. Add 3.5 µg of glycogen, 1 µl of 0.5 M EDTA, 1 µl of 10% SDS, and 1 µl of proteinase K (10 µg/µl). Treat the solution with phenol-chloroform and chloroform for extraction and finally precipitate the reaction using ethanol.

5. Dissolve the pellet in 30 µl of H₂O.

6. Add 3 µl of 10x restriction buffer (MBI buffer) and 50 units/µg of cDNA of XhoI.

7. Incubate the sample at 37°C for 1 hour.

### 3.4.2.3. Dephosphorylation of the cDNA (see Note 28)

1. Add to the DNA

   10x CIP dephosphorylation buffer 5 µl
   H₂O to 48 µl.

2. Add the appropriate amount of CIP according the manufacturer’s instruction.

3. Incubate the tube at 37°C for 30 minutes, add the second aliquot of CIP, and incubate for a further 30 minutes.

4. To terminate the reaction, add SDS and EDTA (pH 8.0) to a final concentration of 0.5% and 5 mM, respectively. Mix the contents well and add proteinase K to a final concentration of 100 µg/ml. Incubate the tube at 56°C for 30 minutes.

5. Cool the reaction to room temperature and extract DNA using phenol:chloroform twice and chloroform alone once.

6. Transfer the aqueous phase to a fresh microfuge tube and recover the DNA with ethanol precipitation in the presence of 0.1 volume of 3 M sodium acetate (pH 5.2).

7. Dry the precipitate at room temperature and dissolve it in TE (pH 7.6) at a DNA concentration of >2 nmoles/ml.
3.4.2.4. Purification of the cDNA

1. After the dephosphorylation reaction, add 1 μl of 0.5 M EDTA, 1 μl of 10% SDS, and 1 μl of 10 mg/ml proteinase K.

2. Treat the sample with phenol-chloroform and chloroform, saving both the aqueous and the organic phase each time. Back-extract the two organic phases obtained from both procedures using 5 mM EDTA and 100 mM NaCl. Pool the aqueous phase obtained from both the primary extractions and back-extractions separately.

3.4.2.5. Fractionation of cDNA According to Size

The removal of primer- adapters and short full-length cDNAs is achieved in this stage. Use SizeSep Sepharose CL-4B spun columns (Amersham-Biosciences) following the manufacturer’s instructions.

3.4.3. Preparation of the Cloning Vector

A new class of cloning vector was developed specifically for full-length cDNA library construction. The major problems associated with the cloning vector for representative full-length cDNA library was that shorter cDNAs are more efficiently cloned into a vector than longer ones. To overcome this problem, we have developed λ-full-length cDNA (λ-FLC) cloning vectors. One advantage of using the λ vector is that, together with the use of Cre-lox system for bulk excision, λ-FLC cloning vectors make it possible to overcome the size bias and hence clone rarely expressed mRNA more effectively (3).

3.4.3.1. Preparation of λ-Arms for Cloning

1. Prepare the final λ-DNA constructs, extracting from bacteriophage λ, DNA from large-scale cultures using proteinase K and SDS.

   1) Place the prepared bacteriophage suspension in a section of dialysis tubing sealed at one end with a knot or a plastic closure. Close the other end of the dialysis tube and leave the sealed tube in the bacteriophage suspension for 1 hour at room temperature with slow stirring.

   2) Transfer the dialysis tube to a fresh flask of buffer and dialyze the bacteriophage suspension again for 1 hour.

   3) Transfer the bacteriophage suspension into a polypropylene centrifuge tube and add 0.5 M EDTA (pH 8.0) to a final concentration of 20 mM.

   4) Add proteinase K to a final concentration of 50 μg/ml and SDS to a final concentration of 0.5%. Mix the solution by gentle inversion of the tube several times.

   5) Leave the tube at 56°C for 1 hour and cool it to room temperature.

   6) Add an equal volume of equilibrated phenol to the digestion mixture and invert the tube several times until a complete emulsion has formed.

   7) Centrifuge the tube at 5,000 rpm for 5 minutes at room temperature. Transfer the aqueous phase to a fresh tube.
8) Extract the aqueous phase with a 1:1 mixture of equili-
brated phenol and chloroform and recover the aqueous
phase following Step 7. Repeat the extraction with an
equal volume of chloroform and transfer the aqueous
phase to a dialysis sac.

9) Dialyze the preparation of λ-DNA overnight at 4°C
through 3 changes of a 1000-fold volume of TE (pH 8.0).

10) Measure the concentration using OD_{260} and check the
integrity of λ-DNA using 0.7% agarose with an appropri-
ate marker. Keep the stock of λ-DNA at 4°C.

2. Place the 10 μg of λ-DNA in 180 μl of 10 mM Tris-HCL (pH
7.5)/10 mM MgCl₂, and incubate the sample at 42°C for
2 hours.

3. Add 20 μl 10x ligation buffer and 400 U of ligase (New
England Biolabs) and incubate the mixture at room tempera-
ture for 5 hours.

4. Incubate the mixture at 65°C for 15 minutes to inactivate
the ligase.

5. λ-DNA is then digested with appropriate restriction enzymes
as follows (see Note 29):
   i. Add 2 μl of 5 M NaCl, 6 U of FseI, and 8 U of PaeI to the
      vector and incubate the sample at 37°C for 4 hours or
      overnight.
   ii. Add 2 μl of 5 M NaCl, 30 μl of 10x NEB3 buffer, 270 μl of
       H₂O, and 20 μl of SwaI to the mixture and incubate at
       room temperature for 2 hours.
   iii. Heat the reaction tube to 65°C for 15 minutes.

6. Add 5 μl of 5 M NaCl, 40U of XhoI, 40 U SalI, and 40 U of
   BamHI and incubate the mixture at 37°C for 4 hours.

7. After restriction, purify the DNA with proteinase K treatment
with 0.1 % SDS and 20 mM EDTA.

8. Extract the DNA with phenol:chloroform (1:1) followed by
chloroform treatment and precipitate the sample with ethanol
(see Notes 2 and 30).

9. Resuspend the DNA in either 1x TE or H₂O for at least
30 minutes.

10 Separate the digested DNA in 0.6% low-melting point agarose
gel (Seaplaque, FMC) as follows:
   i. Set the well in the middle of the gel.
   ii. Digest the sample with StyI and carry out electrophoresis
      at 8 V/cm for 1.5 hour.
   iii. Cut out and discard the DNA fragments shorter than
19 kb (Step 1).
iv. Replace the electrophoresis buffer (1x TBE) with fresh buffer.

v. Apply 8 V/cm in the opposite direction for 2.5 hours.

vi. Cut out and discard the DNA fragment shorter than 19 kb (Step 2).

vii. Replace the electrophoresis buffer (1x TBE) with fresh buffer again.

viii. Apply 8 V/cm in the same direction as Step 2 for 30 minutes.

ix. Cut out the DNA fragment from the gel (Step 3) and equilibrate with TE buffer.

x. Using β-agarase (New England Biolabs), purify and check λ-arm (see Carninci, 1999 for the detail).

xi. Store the purified λ-arms in single-use aliquots at −80°C indefinitely or in 4°C for up to 1 week.

3.4.3.2. Cloning the cDNA

1. Mix cDNA and an equimolar quantity of λ-arms.

2. Adjust the volume to ~4 μl using a rotary evaporator if necessary (see Note 31).

3. Add 0.5 μl of 10x buffer and ~0.4 μl of T4 DNA ligase.

4. Incubate the reaction at 16°C overnight.

5. Package the products using ~25 μl of a high-efficient packaging mix, following the manufacturer’s instruction.

6. Measure the number of viable bacteriophages in an aliquot of the mixture by plaque formation on lawns of XL-Blue MRA(P2) bacteria (see Note 2).

7. Follow Steps 1 to 9 below to prepare and titer a plate stock of the cDNA library on lawns of XL-Blue MRA(P2) bacteria.

1) Prepare infected cultures for plating by mixing appropriate pfu of bacteriophage.

2) Add 3 ml of molten top agarose (40°C) for 10-cm plate or 7.0 ml of molten top agarose (40°C) for 15-cm plate to the first tube of infected cells. Mix the contents by gentle tapping or vortexing of the tube for a few seconds and immediately pour the entire contents of the tube onto the center of a labeled agar plate. Ensure an even distribution of bacteria and top agarose by gently swirling the plate. Repeat this step until the contents of each of the tubes have been transferred onto separate plates.

3) Incubate the plate at 37°C (without inversion) for 12–16 hours.

4) Take the plates from the incubator and add SM (see Note 2). Incubate the plates at 4°C for several hours on a shaking platform.
5) Transfer as much of the SM as possible into sterile screw-
or snap-cap polypropylene tubes using a separate Pasteur 
pipette for each plate.

6) Add 1 ml of fresh SM to each plate, gently swirl the plate, 
and store the plate in a tilted position for 15 minutes to 
allow the fluid to accumulate in one area. Remove the 
SM and combine it with the first harvest and discard the 
plate.

7) Add 0.1 ml of chloroform to each of the tubes containing 
SM, vortex the tubes briefly, and centrifuge at 5,800 rpm 
for 10 minutes at 4°C to remove the bacterial debris.

8) Remove the supernatants to fresh polypropylene tubes 
and add 1 drop of chloroform to each tube. Keep the 
resulting bacteriophage plate stock at 4°C.

9) Measure the concentration of infectious virus particles in 
each stock by plaque assay (see Note 2).

8. Plate 20–50,000 pfu on lawns of E. coli (strain C600) on 
150-mm petri dish to amplify the library (see Note 2).

### 3.5. Bulk Excision of cDNA

**3.5.1. In Vivo Solid-Phase Excision**

1. Streak a frozen BNN132 cell stock onto an LB-agar plate 
containing 50 μg/ml kanamycin and incubate the plate at 
37°C overnight, then store the plate at 4°C.

2. Pick an isolated single colony into 50-ml LB containing 
10 mM MgSO₄ and 0.2 % maltose. Incubate the culture at 
37°C with shaking until OD₆₀₀ = ~0.5

3. Centrifuge the tube containing the cells at 4,000 rpm and 
4 °C, and resuspend the cells in 10 ml of 10 mM MgSO₄.

4. Infect the suspension containing ~2.5 × 10¹⁰ cells with the 
amplified phage library, at the multiplicity of ~1.0, and plate 
the cells on LB-ampicillin plate at 37°C.

**3.5.2. In Vivo Liquid-Phase Excision**

1. Grow a fresh 50-ml culture of BNN132 cells in LB medium 
to an OD₆₀₀ of ~0.5 for each excision reaction.

2. Centrifuge the cells at 4,000 rpm, 4°C for 10 minutes and 
resuspend the cells in 10 ml of 10 mM MgSO₄.

3. Infect the bacteria with 1–5 × 10¹⁰ pfu of the λ cDNA library 
(MOI = ~1).

4. Incubate the bacterial suspension at 37°C for 20 minutes with 
shaking at 100 rpm.

5. Add the suspension to 90 ml of LB containing 100 μg/ml of 
ampicillin.

6. Place the culture in the shaking incubator at 30°C for 1–2 hours.
7. Extract the plasmid DNA using the Wizard Plus Midiprep DNA Purification System (Promega) as follows (see the manufacturer’s instruction):
   i Collect the cultured cell by centrifugation at 5,000 rpm for 5 minutes at 4°C.
   ii Discard the supernatant and resuspend the pellet in 4 ml of the cell suspension solution and mix well.
   iii Add 4 ml of cell lysis solution and incubate for 3 to 5 minutes.
   iv Centrifuge the tube at 6,000 rpm for 15 minutes at 4°C.
   v Recover the supernatant by decanting through gauze.
   vi Add 10 ml of DNA-binding resin and transfer the suspension to the filter.
   vii Wash the filter as described in the instruction with ethanol/salt buffer.
   viii Elute the plasmid DNA using nuclease-free water.

8. Add 0.05 volume of 5 M NaCl and 1 volume of isopropanol to precipitate DNA.

9. Place the solution at –20°C for 30 minutes.

10. Centrifuge the tube containing the solution at 15,000 rpm for 15 minutes at 4°C.

11. Discard the supernatant and wash the pellet with 80% ethanol.

12. Recentrifuge the tube at 15,000 rpm for 3 minutes.

13. Repeat Steps 11 and 12.

14. Discard the supernatant and redissolve the pellet in 50 μl of H2O.

15. Measure OD 260 of the DNA solution and check the quality by gel electrophoresis.

---

3.5.3. In Vitro Cre-lox-Mediated Excision

1. Use the Wizard Lambda Preps DNA Purification System (Promega) to isolate λ DNA from the phage library.

2. Transfer 1/4 of the purified DNA to plasmid as follows:
   i Add 1 unit of Cre-recombinase to the DNA solution.
   ii Incubate the DNA at 37°C for 1 hour as described in the manufacturer’s instruction (Novagen).
   iii Extract the DNA using the digested DNA by phenol-chloroform and chloroform extraction followed by ethanol precipitation.
   iv Transform the bulk-excised into DH10B cells by electroporation at 20 kV/cm.
3.6. Evaluation of Plasmid cDNA Libraries

Evaluate the cDNA libraries constructed following the described procedure. There are two methods to assess the size of the library. The first is the long PCR method using the lysate from individually transformed cells as template and two polymerases (6). The second is to digest the excised plasmids with a restriction enzyme (3). The average size of the cDNA inserts from excised plasmid cDNA libraries can be 2.9 kb for standard size and 6.9 kb for size-selected cDNA. The less common the full-length DNA are, the larger they tend to be in size. This means that the discovery of new genes depends on the size of the cDNA insert.

4. Notes

1. “Large-scale full-length cDNA libraries” indicates a full-length cDNA library covering the entire genome of the organism, where small-scale cDNA library refers to only a fraction of all the genes of the organism under consideration.

2. The detailed protocol can be found in 3rd edition of Molecular Cloning (5).

3. The solution contains 10 mM of each dATP, dGTP, dTTP, and 5-methyl-dCTP, comprising 625 μM dNTPs.

4. The solution contains 10 mM of each dATP, dGTP, dTTP, and 5-methyl-dCTP, comprising 312 μM dNTPs.

5. The organism from nature can be stored in –80°C for up to several months. The organism has to be frozen in liquid nitrogen immediately after harvesting. In the form of RNA, samples can be stored at –80°C for up to 1 year after freezing with liquid nitrogen, or at –4°C for 1 week without RNase contamination.

6. Worms do not dissolve completely. Do not leave samples more than 15 minutes.

7. The concentration of primer shown here is for the 40–50-mers. The concentration should be scaled up or down depending on the size if different primers are used (12).

8. The sample cannot contain ethanol from the previous purification step. Residual ethanol may affect the reverse transcription; thus, the sample needs to be free of ethanol (13).

9. Original concentration of T4 gene 32 is 1.5 μg/μl. For optimal use, 150 ng/μl is required in the reaction mixture.
10. For cold-start reactions, mix the contents of tubes A and B quickly on ice and transfer 40 μl of the mixture to tube C and run thermal cycler immediately (needs to be preheated at 40°C). The thermal cycle is the same as hot start. The major concern of the hot start is the internal priming (12).

11. When total RNA is used instead of mRNA, set the thermal cycler as follows: 40°C for 2 minutes, –0.1°C/sec to 35°C, 50°C for 2 minutes, 56°C for 60 minutes, and 4°C infinitely (17).

12. GC-I buffer is originally designed for the LA-Taq DNA polymerase (TaKaRa, Shiga, Japan) (13).

13. Hot-start is where the enzyme is added after the initial denaturation to avoid nonspecific amplification by primers misannealing.

14. At this stage, the supernatant contains most of the radioactive dNTPs, which should be treated carefully.

15. The incubation period at either –20°C or –80°C is preferably longer for smaller portions of the cDNA.

16. At the end of this step, ethanol needs to be removed completely by pipetting.

17. Take 50 μl of supernatant from Step 7-III and dilute it with 450 μl of dH2O. Store the dilution on ice for later analysis in Note 13.

18. The analysis with electrophoresis on alkaline gel will show the differences in the product size; usually the captured fraction will be slightly longer than the uncaptured fraction.

19. Take 10 μl of the elution solution in Step 4 and use this together with the diluted form.

20. There needs to be more than 200–300 ng of cap-trapped cDNA for normalization and subtraction, otherwise, skip the normalization and subtraction steps and move straight onto the second-strand cDNA synthesis (19).

21. Instead of using a minilibrary, nonredundant clones of the rearrayed library can be used if available. This could be made by plating a replica from a 384-well plate of a rearrayed library.

22. The use of commercially available kits is recommended for this step. In our group, Ambion kit is used.

23. Store the tube at –80°C if not progressing to the next step.

24. Calculate the amount of drivers required under the condition described in Section 3.3.3. Use RoT values (20) of 5–10 for normalization of a cDNA population, and up to 500 to subtract abundantly expressed cDNAs. RoT values of 5–10 will complete the hybridization of mammalian superprevalent cDNA, but will leave intermediately expressed cDNAs partial and leave rare cDNA as single-stranded. Use RoT values
of 1–2 for normalization of genome with lower complexity (e.g., protozoa, insects). The value depends on the complexity of the transcriptome. The formula used here is the following:

\[ RoT_{10} = 0.97 \, \text{µg/µl of RNA hybridized under standard condition for 1 hour} \]

Normalizing driver: The concentration of 0.97 µg/µl of normalized RNA driver gives \( RoT = 5 \) after 30 minutes. The effective amount of normalizing driver has an equal concentration of cDNA in the reaction mixture. The yield of first-strand cDNA will be equal to the normalization driver. For instance, the reaction mixture containing 500 ng of each normalization driver and cDNA will require 9.7 hours to achieve \( RoT = 10 \) and 4.5 hours to achieve \( RoT = 5 \).

Subtraction driver: The excess subtraction drivers require certain amounts of cDNA. For every 200 clones used as template, the value of \( RoT \) should be increased by 1. For example, the \( RoT \) value is 50 when the subtractive RNA driver is prepared from a minilibrary of 10,000 clones. For 50 as \( RoT \) value, subtractive mRNA needs to be up to 2.58 µg with the incubation period of 9.7 hours. Normalizing and subtractive \( RoT \) values should be calculated separately. The total \( RoT \) value is the sum of the normalizing and subtracting \( RoT \) values.

Sequence masking of RNA driver: The subtractive RNA driver often carries inappropriate sequences derived from PCR or plasmid DNA from previous steps. The synthetic oligonucleotides are essential to avoid the nonspecific capture on the streptavidin-coated magnetic beads. The oligonucleotides should contain sequences complementary to common sequences of the RNA driver. Use 150 ng for each of the 5' and 3' antisense oligonucleotides comprised of 40 to 50 bases per 1 microgram of subtractive RNA driver. These synthetic oligonucleotides will not interfere with second-strand cDNA synthesis. In the case that the sequences of the blocking oligonucleotides and the second-strand cDNA synthesis primer are the same, the blocking oligonucleotides need to be biotinylated for later removal, requiring the addition of extra beads. See CPG beads instruction sheet for more information.

25. Add H\(_2\)O to the filter and centrifuge again if required. The recovered cDNA may be up to 20 µl.

26. The yields can be as low as 20 ng of normalized cDNA for further library construction.

27. Do not use dephosphorylation if SstI or other 3’-protoruding enzymes are used to digest double-stranded cDNA.
28. Shrimp alkaline phosphatase (SAP) (Promega, Roch, etc.) can be used instead of CIP. Follow the manufacturer’s instruction accordingly.

29. For λ-FLC-II vectors, SalI may be omitted or it may be used to generate an alternative to the XhoI cloning site. For λ-FLC-I-B vectors, FseI, PacI, and SwaI are omitted, as they do not carry these sequences.

30. Keep the concentration of DNA less than 20 μg/ml to avoid problems associated with resuspension.

31. When using a rotary evaporator, do not allow the sample to dry completely in the evaporator, as desiccated DNA does not redissolve. At the same time, make sure the EDTA concentration remains <1 mM.

Acknowledgments

This work was supported by a Research Grant for the RIKEN Genome Exploration Research Project to Y.H and a grant of the Genome Network Project, both from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government.

We would like to thank Dr. Carninci P. for his advice, and also the colleagues in the Genome Exploration Research Group at the Genomic Sciences Center (GSC), RIKEN Yokohama Institute, for all the support.

References


Chapter 6

Expressed Sequence Tags: Normalization and Subtraction of cDNA Libraries

Marcelo Bento Soares, Maria de Fatima Bonaldo, Jeremiah D. Hackett, and Debashish Bhattacharya

Abstract

Expressed Sequence Tags (ESTs) provide a rapid and efficient approach for gene discovery and analysis of gene expression in eukaryotes. ESTs have also become particularly important with recent expanded efforts in complete genome sequencing of understudied, nonmodel eukaryotes such as protists and algae. For these projects, ESTs provide an invaluable source of data for gene identification and prediction of exon-intron boundaries. The generation of EST data, although straightforward in concept, requires nonetheless great care to ensure the highest efficiency and return for the investment in time and funds. To this end, key steps in the process include generation of a normalized cDNA library to facilitate a high gene discovery rate followed by serial subtraction of normalized libraries to maintain the discovery rate. Here we describe in detail, protocols for normalization and subtraction of cDNA libraries followed by an example using the toxic dinoflagellate *Alexandrium tamarense*.

Key words: cDNA library, dinoflagellate, expressed sequence tag (EST), normalization, reassociation kinetics, redundancy, subtracted cDNA library.

1. Introduction

1.1. Overview of the EST Approach for Gene Discovery

The generation of Expressed Sequence Tags (ESTs) from 3’ terminal exons of cDNA clones that are randomly picked from libraries is an efficient strategy for identifying transcripts (1–9). However, it is important to acknowledge that despite its advantages, there are several problems associated with the EST approach (10). Here we present a critical analysis of the EST strategy, and describe a procedure that has been developed in the Soares lab to specifically address some of the problems associated with the EST approach (10–12). This strategy, called Serial Subtraction of Normalized
Libraries, has been successfully applied in our EST program (http://genome.uiowa.edu/GeneDiscovery.htm; (13–15). We provide here an example of how we have implemented this strategy to develop a comprehensive EST set from the toxic, “red tide”-forming dinoflagellate alga *Alexandrium tamarense*.

The mRNAs of a typical somatic cell are distributed into three frequency classes: (I) prevalent, (II) intermediate, and (III) complex (16, 17). The classes at the two extremes (comprising 10% and 40–45% of the total, respectively) include members occurring at vastly different relative frequencies. On average, the most prevalent class consists of about 10 mRNA species, each represented by 5,000 copies per cell, whereas the class of high complexity comprises 15,000 different species, each represented by only 1–15 copies. Even though the most rare mRNA from any tissue is likely to be represented in a cDNA library of 5–10 million recombinants, its identification is very difficult (its frequency of occurrence may be as low as \(2 \times 10^{-6}\) on average or even \(10^{-7}\) for complex tissues such as the brain). The probability that a given mRNA will be represented in a cDNA library can be expressed by the equation \(P(x) = 1 - (1 - f)n\), where \(f\) = frequency and \(n\) = number of recombinant clones (12).

One of the problems commonly observed in large-scale EST programs is the redundant generation of ESTs corresponding to the most common RNAs (i.e., mRNAs of the prevalent and intermediate frequency classes, mitochondrial RNAs, and rRNAs). This is a problem that can significantly impair the overall efficiency of a gene discovery program that relies solely on the generation of ESTs from cDNA clones randomly picked from standard libraries. For example, it has been estimated that in a complex organ such as the vertebrate brain, there are about 36 prevalent mRNAs, 2,150 intermediate mRNAs, and as many as 45,000 different rare (complex) mRNAs, which comprise 16%, 46%, and 38% of the total mRNA mass, respectively. Thus, 62% of the clones picked at random from a whole brain library are expected to correspond to any one of the 2,000 or so class I + II transcripts present in the brain. As a result, one would anticipate that even during the early stages of a brain gene discovery program, novel ESTs would be identified at a frequency lower than 40% (12).

The use of normalized cDNA libraries has been shown to expedite gene discovery in large-scale EST programs. Because in a typical normalized cDNA library the frequency of all clones is within an order of magnitude range, redundant identification of the most common RNAs is greatly minimized. Normalized libraries can be generated by a number of reassociation-kinetics based procedures. It is noteworthy, however, that the process of normalization only contributes to minimize redundancies within libraries. This problem can be more effectively addressed by the use
of subtractive libraries that are progressively enriched for novel ESTs. This is the rationale behind our strategy to generate ESTs from serially subtracted, normalized libraries \((11, 18)\).

Serial subtraction of normalized libraries is an iterative strategy, whereby all arrayed cDNA clones from a library are pooled and used as a driver in a subtractive hybridization with the library from which they originated. Since the representation of the driver population is significantly reduced in the resulting subtracted library, redundant generation of ESTs is significantly minimized. Hence, every new library of a series is enriched for novel ESTs. It is noteworthy that this process can occur without compromising sequencing throughput because the processes of subtractive hybridization and EST generation are done simultaneously. In other words, while ESTs are being generated from a collection of arrayed clones (5,000 in average), the latter are being pooled and used as a driver in a hybridization reaction with their library of origin to generate the next subtracted library of the series. By the time generation of ESTs from all clones of the array is completed, a set of arrayed clones from the new subtracted library has been produced. Prediction of novelty becomes critical to allow us to determine the depth at which any given subtracted library of a series should be sequenced. This strategy has been applied successfully in many projects including gene discovery in \textit{A. tamarense}. In this species, we were able to generate 8,996 unique ESTs from 14,611 total sequences (61.57% novelty; 14).

### 2. Materials

#### 2.1. Normalized Library Construction

1. General materials: phenol, chloroform, 3M sodium acetate pH 7.0, TE buffer, ethanol, and glycogen, 0.5 M EDTA.

#### 2.1.1. Production of Single-Stranded DNA Tracer

1. A directionally cloned cDNA library constructed using a plasmid vector with an \textit{f1} origin for the production of single-stranded DNA (see \textbf{Note 1}).
2. \textit{E. coli} DH5\textalpha\textsuperscript{F'} phage-susceptible host cells.
3. 1 mm electroporation cuvettes.
4. 2\% YT bacterial growth media and agar plates.
5. Filter-sterilized 20\% (w/v) glucose solution.
6. Ampicillin (75 mg/ml).
7. M13 K07 helper phage.
8. 13 ml polypropylene centrifuge tubes.
2.1.2. Purification of Single-Stranded DNA

1. *Pvu*II restriction enzyme and reaction buffer.
2. Hydroxyapatite resin (HAP).
3. HAP Buffer (*see Note 2*).
4. Silanization solution (e.g., Sigmacote).
5. Water-saturated and dry 2-butanol.
6. Water-saturated ether.
7. Nensorb 20 nucleic-acid purification cartridge (PerkinElmer).

2.1.3. Production of DNA Driver by PCR

1. Qiagen Taq DNA polymerase kit.
2. T3 (5’ attaacccctcactaaaggga 3’) and T7 (5’ taatacgactcacta- taggg 3’) primers (20 μM).

2.1.4. Hybridization of Driver with ssDNA Tracer

1. Blocking oligonucleotides for pT3T7-Pac vector: 5’ blocking oligo (5’ taatacgactcactatagggaatttggccctcgaggccaagaattcgg- cacgagg 3’), 3’ blocking oligo (5’ a20nnnnngcggccgcaagtattccctttagtgagggttaat 3’), and tail blocking oligo (5’ a50 3’).
2. 1.2 M NaCl in 10 × TE buffer.
3. 20% SDS.
5. Deionized formamide.

2.1.5. Separation of ssDNA from the Hybridized DNA Duplex

1. Same materials as in *Section 2.1.2*.

2.1.6. Conversion of ssDNA to dsDNA

1. Sequenase v2.0 DNA polymerase and 5 × reaction buffer (USB).
2. M13 forward primer (5’ gttttccagtcacgac 3’).
3. dNTPs (10 mM).
4. 0.1 M DTT.

2.1.7. Amplification of the Normalized cDNA Library

1. 1 mm electroporation cuvettes.
2. *E. coli* DH10B phage-resistant electrocompetent host cells.
3. 2 × YT bacterial growth media and agar plates.
4. Qiagen-tip 100 high-speed Midi prep kit or equivalent.
### 2.2. Subtracted Library Construction

#### 2.2.1. Production of Single-Stranded DNA Tracer
1. Normalized cDNA library.
2. Same materials as in Section 2.1.1.

#### 2.2.2. Purification of ssDNA Tracer
1. Same materials as in Section 2.1.2.

#### 2.2.3. Production of Single-Stranded DNA Driver Template
1. Pool of cDNA clones sequenced from the normalized library.
2. Same materials as in Section 2.1.1.

#### 2.2.4. Purification of ssDNA Driver Template
1. Same materials as in Section 2.1.2.

#### 2.2.5. Production of DNA Driver by PCR
1. ssDNA driver template from Section 2.2.4.
2. Same materials as in Section 2.1.3.

#### 2.2.6. Hybridization of Driver with ssDNA Tracer
1. Same materials as in Section 2.1.4.

#### 2.2.7. Separation of ssDNA from the Hybridized DNA Duplex
1. Same materials as in Section 2.1.2.

#### 2.2.8. Conversion of ssDNA to dsDNA
1. Same materials as in Section 2.1.6.

#### 2.2.9. Amplification of the Subtracted cDNA Library
1. Same materials as in Section 2.1.7.

### 3. Methods

#### 3.1. Normalized Library Construction

##### 3.1.1. Production of Single-Stranded DNA Tracer
1. Electroporate start-library DNA into *E. coli* DH5α F’. The amount electroporated should be enough to result in a 5-fold increase over the number of recombinants in the start library, not exceeding 5 ng per electroporation.

2. Inoculate the bacteria into 100 ml of 2 × YT broth and incubate the culture at 30°C at 150 rpm for 1 h.

3. Plate 5 μl on a 2 × YT agar plate and incubate at 37°C overnight.
4. Add 100 μl of ampicillin (75 mg/ml) and 1 ml of 20% glucose. Grow culture at 30°C at 300 rpm overnight.

5. Count the colonies on the agar plate to determine the efficiency of transformation.

6. Inoculate 15 ml of 2 × YT broth with 200 μl of the overnight culture. Add 15 μl of ampicillin and 150 μl of 20% glucose.

7. Grow the culture at 30°C until the OD<sub>600</sub> = ~0.1, then grow the culture at 37°C until the OD<sub>600</sub> = ~0.2.

8. Superinfect the culture with M13 KO7 helper phage at a multiplicity of infection of 10. For OD<sub>600</sub> = 0.2 = 1.2 * 10<sup>8</sup> cells per ml. 10 ml of culture has 1.2 * 10<sup>9</sup> cells, so the culture should be infected with 1.2 * 10<sup>10</sup> pfu of phage.

9. Grow the culture at 37°C for 2 h in an orbital shaker.

10. Transfer the culture to a 13-ml centrifuge tube and add 0.4 g of polyethelene glycol MW 8000 and 0.3 g of NaCl per 10 ml of culture.

11. Precipitate on ice for 1 h or at 4°C overnight.

12. Centrifuge at ~15,000 RCF (e.g., in a Sorvall RC-5 with a SS-34 rotor at 11,500 RPM).

13. Remove the supernatant and resuspend the phage pellet in 0.5 ml of TE.

14. Transfer to a 1.7-ml centrifuge tube and add 0.5 ml of phenol.

15. Mix by vortexing and centrifuge in a standard bench-top microcentrifuge at full speed for 3 min.

16. Transfer the aqueous phase to a new 1.7-ml centrifuge tube and extract with 0.5 ml phenol:chloroform (1:1). Centrifuge as in step 15 and repeat 1 time.

17. Precipitate the ssDNA by adding 0.1 volume of 3 M sodium acetate (NaOAc), pH 7.0 and 2.5 volumes of 100% ethanol.

18. Precipitate at –20°C for 2 h.

19. Centrifuge at full speed for 30 min and remove the supernatant. Wash the DNA pellet with 75% ethanol, centrifuge for 15 min, remove the supernatant, and resuspend the DNA in 90 μl of TE.

20. Run 1 μl of the sample on an agarose gel to check the quality and amount of ssDNA.

### 3.1.2. Purification of Single-Stranded DNA

1. Digest 10 ng of ssDNA in a 100 μl PvuII reaction comprised of 10 μl of 10 × buffer and 2 μl of PvuII restriction enzyme (10 U/μl). Bring the reaction volume up to 100 μl with water. Incubate at 37°C for 2 h.
2. Prepare the HAP loading buffer (*see Note 2*). Preheat the buffer in a 60°C water bath.

3. Silanize the HAP column using Sigmacote (*see Note 3*).

4. Measure 0.4 g of HAP resin and resuspend the powder in 5 ml of HAP buffer.

5. Load the HAP suspension onto the column and apply positive pressure to drain excess buffer until the buffer is just above the settled HAP resin. Do not allow the resin to run dry. Throughout this procedure, a flow rate of 2–5 ml per minute should be achieved in all steps.

6. Add 1 ml of HAP buffer to the *Pvu* II reaction and load the reaction mixture onto the HAP column. Be careful not to disturb the settled HAP resin.

7. Collect the drops from the bottom of the column in a silanized 50-ml conical tube.

8. Reload the flow-through and collect the drops in the same tube.

9. Wash the tube containing the *Pvu* II reaction with 1 ml of HAP buffer and load it onto the column. Collect the drops in the same conical tube.

10. Wash the HAP column three times with 2 ml of HAP buffer, collecting the drops in the same conical tube. The final volume should be ~8 ml.

11. Extract the elution twice with 40 ml of water-saturated 2-butanol.

12. Extract once with 35 ml of dry butanol. This should reduce the volume of the aqueous phase. Extract a second time if the aqueous volume is greater than 5 ml.

13. Extract once with 10 ml of water-saturated ether.

14. Heat the sample at 65°C and remove ether with vacuum.

15. Purify the ssDNA using a Nensorb column according to the manufacturer’s instructions, except elute using 1 ml of 50% ethanol.

16. Precipitate the ssDNA by adding 50 µl of 3 M sodium acetate and 0.5 ml of 100% ethanol. Precipitate at –20°C for 2 h.

---

**3.1.3. Production of DNA Driver by PCR**

1. Centrifuge the precipitated ssDNA for 30 min, wash with 75% ethanol, and resuspend in 5 µl of TE. Run 0.5 µl of the ssDNA on an agarose gel to check concentration and quality.

2. Set up a PCR reaction using the Qiagen *Tag* polymerase comprised of 10 µl of 10 × buffer, 15 µl of 5 × Q-solution, 2 µl of 10 mM dNTPs, 5 µl each of T7 and T3 primers (20 µM), 0.5 µl *Tag* polymerase (5 U/µl), 1 µl of template ssDNA (1 ng/µl), and water to a final volume of 100 µl.
3. Run on a thermocycler with the following program: 94°C for 7 min (94°C 1 for min, 55°C for 2 min, 72°C for 3 min) × 20 cycles, 72°C for 7 min.

4. Purify the PCR products with two Centricep columns according to the manufacturer’s instructions.

5. Extract once with phenol:chloroform (1:1).

6. Precipitate the PCR product with 0.1 volume of 3 M sodium acetate pH 7.0 and 2.5 volumes of 100% ethanol.

7. Incubate at –20°C for 2 h.

3.1.4. Hybridization of Driver with ssDNA Tracer

1. Centrifuge the precipitated driver for 30 min, wash with 75% ethanol, and resuspend in 10 μl of deionized formamide.

2. Run 1 μl on an agarose gel to determine driver concentration.

3. Set up the hybridization reaction. The reaction should have a final concentration of 50% formamide and is comprised of 500 ng of PCR driver, 50 ng of ssDNA, 10 μg each of the 5’ blocking oligo, the 3’ blocking oligo, and the tail blocking oligo.

4. Cover the reaction with 50 μl of mineral oil.

5. Heat at 80°C for 3 min.

6. Add 1.2 M NaCl in 10 × TE to a final concentration of 1 × and 20% SDS to a final concentration of 1%. It is important that the NaCl and the SDS do not mix before they are added to the rest of the reaction. Add these to opposite sides of the tube and make sure that they completely fall through the mineral oil layer.

7. The ideal final volume is 20 μl.

8. Calculate the time needed to reach \( C_{o,t} = 5 \). \( (0.5 \text{ μg} / X \text{ μl}) \times 20 = \text{O.D.} \), \((\text{O.D.} / 2) \times 0.45 \times \text{time}(Y \text{ hours}) = C_{o,t} = 5 \). For example, 0.5 μg in 20 μl equals an O.D. of 0.5. Then O.D. of \( 0.5 / 2 \times 0.45 \times Y = 0.225Y \). Solve for \( Y \) to get \( 0.25 = 22.22 \text{ h} \).

9. Incubate the reaction at 30°C for \( Y \) hours.

3.1.5. Separation of ssDNA from the Hybridized DNA Duplex

1. Isolation of the unhybridized ssDNA from the reaction is an identical procedure to the original isolation of ssDNA tracer (Section 3.1.2). Skip the digestion in step 1 and follow steps 2–15 using the hybridization reaction instead of the PvuII reaction. Do not attempt to remove the mineral oil; it will have no effect on the HAP column.

2. Because of the low concentration of ssDNA following isolation, add 0.5 μl of glycogen to the ethanol precipitation.

3.1.6. Conversion of ssDNA to dsDNA

1. Spin down ssDNA and wash with 75% ethanol. Resuspend the pellet in 11 μl of H₂O.
2. Add 4 μl of 5 × Sequenase buffer and 1 μl of M13 forward primer (1 μg/μl).
3. Heat at 65°C for 5 min, then at 37°C for 3 min.
4. Add 2 μl of dNTPs (10 mM), 1 μl of 0.1 M DTT, and 1 μl of Sequenase 2.0 polymerase.
5. Incubate at 37°C for 30 min.
6. Add 1 μl of 0.5 M EDTA and vortex.
7. Add 80 μl of TE.
8. Extract with phenol:chloroform (1:1)
9. Precipitate the DNA with 0.1 volume of 3 M sodium acetate and 2.5 volumes of 100% ethanol and 0.5 μl of glycogen. Precipitate at –20°C for 2 h.

3.1.7. Amplification of the Normalized cDNA Library

1. Spin down precipitated DNA and wash with 75% ethanol.
2. Resuspend the DNA pellet in 3 μl of TE.
3. Electroporate all 3 μl (1 μl at a time) into three aliquots of E. coli DH10B cells (1 × 10⁹ cfu/aliquot).
4. Inoculate the bacteria into 100 ml of 2 × YT broth.
5. Incubate the culture at 30°C for 1 h at 15 rpm in an orbital shaker.
6. Plate 0.1, 1, and 10 μl of the culture on 2 × YT plates and incubate at 37°C overnight to determine the number of transformants.
7. Grow the liquid culture overnight at 30°C at 300 rpm.
8. Spin down the overnight culture and extract the plasmids using the Qiagen-tip 100 high-speed Midi prep kit or equivalent.

3.2. Subtracted Library Construction

Construction of the subtracted library is very similar to construction of the normalized library with a few important differences. Whereas in the normalization the driver is the library itself, in the subtraction the driver is the pool of cDNA clones that have already been sequenced. Because of this it is necessary to generate ssDNA from both the normalized library (tracer) and the pool of sequenced clones (driver). Then similar to normalization, a PCR driver is made from the pooled ssDNA that is hybridized to the normalized library ssDNA. The unhybridized DNA from this reaction will be enriched for unsequenced clones, because most of the previously sequenced cloned would have been removed by the HAP column. Also, unlike the normalization reaction in which the representation of clones is modified, the goal in subtraction is to completely remove the previously
sequenced clones from the library. As a consequence, the hybridization is much longer to allow the driver to hybridize to as many of the complementary sequences in the tracer as possible.

3.2.1. Production of Single-Stranded DNA Tracer

1. This procedure is identical to normalization step in Section 3.1.1 starting with the normalized library DNA rather than the starter library to make ssDNA.

3.2.2. Purification of Single-Stranded DNA Tracer

1. This procedure is identical to normalization step in Section 3.1.2.

3.2.3. Production of Single-Stranded DNA Driver

1. This procedure is identical to normalization step in Section 3.1.1 starting with the pool of sequenced clones rather than the starter library to make ssDNA.

3.2.4. Purification of Single-Stranded DNA Driver

1. This procedure is identical to the normalization step in Section 3.1.2

3.2.5. Production of DNA Driver by PCR

1. This procedure is identical to normalization step in Section 3.1.3 using the ssDNA tracer as the PCR template.

3.2.6. Hybridization of the Driver with ssDNA Tracer

1. This procedure is very similar to the normalization step in Section 3.1.4 but with some important differences that reduce hybridization time. This step is described in detail below.

2. Centrifuge the precipitated driver for 30 min, wash with 75% ethanol, and resuspend in 10 ml of deionized formamide.

3. Run 1 ml on an agarose gel to determine driver concentration.

4. Set up the hybridization reaction. The reaction should have a final concentration of 50% formamide and is comprised of 2.5 µg of PCR driver, 50 ng of ssDNA, 40 µg each of the 5' blocking oligo, the 3' blocking oligo, and 10 µg of the tail blocking oligo.

5. Cover the reaction with 50 µl of mineral oil.

6. Heat at 80°C for 3 min.

7. Add 1.2 M NaCl in 10 × TE to a final concentration of 1 × and 20% SDS to a final concentration of 1%. It is important that the NaCl and the SDS do not mix before they are added to the rest of the reaction. Add these to opposite sides of the tube and make sure that they completely fall through the mineral oil layer.

8. The ideal final volume is 20 µl.

9. Calculate the time needed to reach \( C_{ot} = 50 \). \( (2.5 \text{ µg}/X \text{ µl}) * 20 = \text{O.D.}, \ (\text{O.D.}/2) * 0.45 * \text{time} (\gamma \text{ hours}) = C_{ot} = 50 \). For example 2.5 µg in 20 µl equals an O.D. of 0.5. Then O.D. of 0.5/2 *0.45 * \( \gamma = 0.5625 \text{Y} \). Solve for \( \gamma \) to get 50/0.5625 = 88.88 h or 3.7 days.
10. Incubate the reaction at 30°C for γ hours.

3.2.7. Separation of ssDNA from the Hybridized DNA Duplex

1. This procedure is identical to the normalization step in Section 3.1.5.

3.2.8. Conversion of ssDNA to dsDNA

1. This procedure is identical to the normalization step in Section 3.1.6.

3.2.9. Amplification of the Subtracted cDNA Library

1. This procedure is identical to the normalization step in Section 3.1.7.

3.3. An Example: Gene Discovery in the Dinoflagellate A. tamarense

A. tamarense is a marine dinoflagellate alga that is responsible for toxic algal blooms (e.g., red tides) that cause paralytic shellfish poisoning through the production of saxitoxin. We are interested in this organism because of its affects on marine ecosystems and for several fascinating aspects of plastid and nuclear evolution that have occurred in the dinoflagellates (14, 15). Dinoflagellates, in general, have massive nuclear genomes (19). The haploid nuclear genome of A. tamarense is estimated to be about 100 gigabases, or 30 times larger than the human genome. Because of the large genome size, the EST approach was the most efficient method for large-scale gene discovery in this organism.

We constructed a starter, normalized, and a subtracted library from ~500 ng of polyA mRNA from A. tamarense. We produced a total of 14,611 high-quality sequences from the three libraries (483, 10,688, and 3,440 sequences from the starter, normalized, and subtracted libraries, respectively). All ESTs were sequenced from the 3’ end to maximize clustering accuracy using the 3’ untranslated region (UTR). The sequences were processed as previously described (20). Identification of a nonredundant “unigene” set of 8,996 unique clusters from the 14,611 sequences was accomplished using UIcluster v3.0.5. The overall novelty rate was 61.57%.

We sequenced very few clones from the starter library, in which a few highly represented transcripts dominated the library. After only 483 sequences, the novelty rate had already dropped to 78% (Fig. 6.1). Following normalization, the novelty rate increased to ~98%. We tracked the novelty of new sequences as they were generated and performed the subtraction when our novelty rate reached 40% from the normalized library. Following subtraction, the novelty of the new ESTs increased to ~80% and we continued to sequence from this library until the project funds were exhausted. The novelty rate from the final round of sequencing from the subtracted library was ~68%, indicating there remained many unsequenced transcripts in the subtracted library. Clustering of the sequences revealed that the vast majority of the clusters (6,158) were “singletons”, or unique ESTs (Table 6.1).
Our largest cluster contained 46 sequences of the histone-like protein of dinoflagellates (14). Using this approach, we were able to generate a large unigene set from *A. tamarense* while minimizing redundant sequencing, highlighting this approach as an efficient tool for gene discovery.

We are currently in the process of expanding the unigene set for *A. tamarense*. The original libraries were made from cells grown under replete, high-light culture conditions. We are presently generating 10,000 ESTs from cells that have been grown under different cell and life cycle stages, in the dark, or under limited nutrient conditions. We will use the subtraction method to maximize the discovery rate of genes expressed under these conditions. In this approach, we extract mRNA from *A. tamarense* grown under different conditions, pool the RNA, construct a starter library, and normalize this library. We then subtract the previously sequenced clones from the original *A. tamarense* libraries to remove them from the new library, made from cells grown under different growth conditions. In this manner, we reduce the amount of redundant sequencing and are able to discover more condition-specific genes using fewer clones. Because all of the libraries are constructed in the same vector, any combination of libraries can be used as the driver or tracer for the subtraction procedure.

**Fig. 6.1.** Schematic graph showing the change in novelty of the ESTs throughout the sequencing of the *A. tamarense* cDNA libraries. The arrows indicate when normalization and subtraction were done.
Table 6.1
Clustering results from the *A. tamarense* cDNA libraries (14)

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6158</td>
</tr>
<tr>
<td>2</td>
<td>1750</td>
</tr>
<tr>
<td>3</td>
<td>567</td>
</tr>
<tr>
<td>4</td>
<td>231</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td>1</td>
</tr>
</tbody>
</table>
4. Notes

1. This protocol describes normalization and subtraction of cDNA libraries constructed in the pT3T7-pac plasmid cloning vector. Other vectors may be used, but will require the synthesis of appropriate blocking oligonucleotides for the hybridization procedure. The blocking oligos prevent the hybridization of the driver to the ssDNA tracer through pairing with vector sequences flanking the cDNA inserts. These oligos must be customized for the specific vector used and cover the entire flanking vector sequence present on the PCR-generated driver.

2. At intermediate concentrations (0.11–0.13 mM sodium phosphate buffer at 60°C), double-stranded DNA binds to HAP, whereas single-stranded DNA flows through. A calibration must be performed every time a new batch of HAP or sodium phosphate buffer is used to be sure that binding conditions are optimized. For 0.12 M sodium phosphate buffer, mix fresh for each HAP column 6.0 ml of 1 M sodium phosphate pH 6.8, 1.0 ml 0.5 M EDTA pH 8.0, 2.5 ml 20% SDS, add H2O to a final volume of 50 ml.

3. The apparatus for HAP chromatography consists of a glass water-jacketed column with a glass filter and a stopcock. The column is attached to a circulating water bath set to 60°C.

References


Chapter 7

Generating EST Libraries: Trans-Spliced cDNAs

Cecilia Fernández and Rick M. Maizels

Abstract

Eukaryotes using trans-splicing for transcript processing incorporate a taxon-specific sequence tag (the spliced leader, SL) to a proportion (either all or a fraction) of their mRNAs. This feature may be exploited for the preparation of full-length-enriched cDNA libraries from these organisms (a diverse group including euglenozoa and dinoflagellates, as well as members from five metazoan phyla: Cnidaria, Rotifera, Nematoda, Platyhelminths and Chordata). The strategy has indeed been widely used to construct cDNA libraries for the generation of ESTs, mainly from parasitic euglenozoa and helminths.

We describe a set of optimised protocols to prepare directional SL-cDNA libraries; the method involves PCR-amplification of SL-cDNA and its subsequent cloning in a plasmid vector under a specific orientation. It uses small amounts of total RNA as starting material and may be applied to a variety of samples. The approach permits the selective cloning of mRNAs tagged with a particular SL from mixtures including large amounts of non-trans-spliced mRNAs. Thus, it allows exclusion of host contamination when isolating SL-cDNAs from parasitic organisms, and has other potential applications, such as the characterisation of the trans-spliced transcriptome from organisms in mixed pools of species.

Key words: Spliced leader, full-length cDNA, trans-spliced cDNA, 5′-untranslated region, trans-spliced transcriptome.

1. Introduction

1.1. Trans-Splicing Incorporates a Sequence Tag at the 5′-End of mRNAs

Spliced leader (SL) trans-splicing is a form of transcript processing whereby a ‘mini-exon’ (the SL-sequence) is donated from the 5′-end of a small, non-polyadenylated, nuclear RNA (the SL RNA) to unpaired splice acceptor sites on primary RNA transcripts to form a common 5′-terminal exon on mature mRNAs (for a recent review see (1); Fig. 7.1). This phenomenon was discovered in the kinetoplastid Trypanosoma brucei (2–4) and subsequently found in nematodes (5–7), plathyhelminths (8–10) and additional
species within the Euglenozoa (11, 12). More recently, it has also been found to occur in other unicellular eukaryotes [all major orders of dinoflagellates (13)], and in several metazoan phyla [Cnidaria (14), Chordata (15, 16, 17), and Rotifera (18)]. The current view is that trans-splicing has a mosaic phylogenetic distribution, but it is unknown whether it is an ancestral eukaryotic trait that has been lost in different lineages or whether it has independently arisen in diverse phyla (1, 19).

As a result of trans-splicing, a conserved taxon-specific sequence tag (Table 7.1) is incorporated to the 5′-end of a proportion of an organism mRNAs, ranging from 100% in kinetoplastids (20), to a majority in nematodes [50 to 90% acquire SL1 in different species, (7, 21)], and to a minority in flatworms [an estimated 25%, (10, 22)]. With the tag, trans-spliced mRNAs

---

**Fig. 7.1. The mechanism of SL trans-splicing: two examples in Metazoa.** (a) Trans-splicing of SL1 in nematodes illustrating the general situation (valid for all organisms except flatworms) where the AUG translation initiator is encoded within the first exon of the pre-mRNA. For further details, refer to the text in Section 1.1. (b) Trans-splicing in *E. granulosus* illustrating that the 3′-terminal AUG of flatworm SLs may act as the translation initiator of mature mRNAs. For further details, refer to the text in Section 1.2. The sequence of the trans-spliced SL is specified in each case. The diagrams include: (i) The substrates of the trans-splicing reaction, i.e. (1) the SL-RNA, which consists of the SL ‘mini-exon’ (in dark grey), followed by a splice donor site (arrowed) and an intron without a 3′-acceptor site (in white); (2) a recipient pre-mRNA, which consists of an outron (an intron lacking a 5′-splice site; in white), followed by two exons and an intron (in grey and white, respectively). (ii) A mature mRNA, which is generated after trans-splicing of the SL sequence at the 3′-splice site of the outron (the splice acceptor site, arrowed), cis-splicing of the intron and polyadenylation. Note that the SL forms the 5′-end of the mature mRNA.
receive the cap of the SL, a modified guanosine linked through a 5′-5′ triphosphate to the 5′-end SL nucleotide. In Euglenozoa, the cap is a 7-methyl-guanosine followed by four methylated nucleotides, a unique structure known as cap4 (23–25); in Metazoa, a 2,2,7-trimethyl-guanosine (8, 14, 22, 26–29); the cap of the dinoflagellate SL has not yet been analysed (13).

The biological role of SL trans-splicing is not entirely understood. Its best-known function is to resolve polycistronic transcripts into individual 5′-capped monocistronic mRNAs, an essential process in organisms whose transcripts are long polycistronic units like trypanosomatids (30) or whose genes are organised in operons like nematodes [(29, 31); see Table 7.1]. Apart from this role, SL trans-splicing is likely to have additional functions (actually, the majority of trans-spliced genes in metazoans are transcribed monocistronically). These are probably a combination of mRNA stabilisation through donation of the SL-cap, sanitisation of the 5′-untranslated region (UTR) through removal of deleterious sequences and regulation of mRNA translation via specific interactions of the SL sequence and the cap with the translational machinery (1, 29, 32, 33, 34).

1.2. Using the Spliced Leader to Prepare Full-Length-Enriched cDNA Libraries

Trans-spliced mRNA molecules carry conserved sequence tags at both ends (the SL at 5′ and a poly-A stretch at 3′). Such tags allow full-length cDNA to be cloned. Actually, the specific priming of reverse transcription with oligo-dT followed by second-strand cDNA synthesis with SL forward primers [the ‘SL-dT approach’; (32)] has been instrumental to isolate full-length copies of mRNAs (35, 36) and to prepare full-length-enriched cDNA libraries for the generation of expressed sequence tags (ESTs), mainly from parasitic organisms (Note 1). In the case of trypanosomatids, which can proliferate in vitro, cDNA cloning for library construction has usually been carried out from poly-A′ RNA and synthesis of the second strand with the Klenow fragment of DNA polymerase (37–41). In the case of metazoan parasites (helminths) that are not amenable to expansion in vitro, total RNA has frequently been used for reverse transcription and second-strand synthesis has been performed by PCR amplification (22, 42–46).

The presence of the SL offers several advantages for cDNA cloning in parasitic organisms (32, 47). Because the SL is parasite specific, SL libraries are free from host cDNAs. Host contamination is a major concern because parasite material is often recovered from either infected hosts or in vitro cultures including host cells as essential components. In addition, SL libraries are free from genomic DNA and mitochondrial transcripts, which is desirable whenever libraries are constructed for the generation of ESTs. In the case of organisms/life-cycle stages for which polyadenylated, mitochondrial RNAs constitute a predominant fraction of poly-A′ RNA, the latter provides a critical advantage (Note 2). Finally, the use of PCR amplification offers the possibility of preparing SL
<table>
<thead>
<tr>
<th>Phylum</th>
<th>Surveyed species</th>
<th>SL sequence</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euglenozoa</td>
<td><em>Trypanosomatids</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Trypanosoma brucei</em></td>
<td>AACGCUAUUUUAGAAGAGUUUCUGUACUAUUUG (35 nt)</td>
<td>(4)</td>
</tr>
<tr>
<td></td>
<td><em>Trypanosoma cruzi</em></td>
<td>AUCUACGCUAUUUAGAUCACAGUUCUGUACUAUUUG (41 nt)</td>
<td>(4)</td>
</tr>
<tr>
<td></td>
<td><em>Euglenoids</em></td>
<td>UUUCUGAGUGCUAUUUUUUCUG (24 nt)</td>
<td>(12)</td>
</tr>
<tr>
<td></td>
<td><em>Euglena gracilis</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dinophyta</td>
<td>Dinoflagellates (species from all major orders)</td>
<td>DCCGUAGCCAUUUUGGCUCAAG (22 nt, D = T, A or G)</td>
<td>(13)</td>
</tr>
<tr>
<td>Cnidaria</td>
<td><em>Hydra vulgaris</em></td>
<td>CAAACUUCUUUUUCUAAUAG (SL-A, 24 nt)</td>
<td>(14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACGGAAAAACACAUACUGAAACUUUUUAGUCCUGUAAUAG (SL-B, 46 nt)</td>
<td></td>
</tr>
<tr>
<td>Nematoda</td>
<td><em>Caenorhabditis elegans</em></td>
<td>GGUUUAUACCCAGUUGAG (SL1, 22 nt)</td>
<td>(5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGUUUAACCCAGUUGUACUAAG (SL2, 22 nt)</td>
<td>(81)</td>
</tr>
<tr>
<td>Rotifera</td>
<td><em>Adineta rivic</em></td>
<td>GGCUIAUAACCUUACCAAGAG (23 nt)</td>
<td>(18)</td>
</tr>
<tr>
<td>Plathyhelminths</td>
<td>Trematodes</td>
<td>AACCUGACCGGUUUAACGUUGUAGUUGUGCAUG (36 nt)</td>
<td>(8)</td>
</tr>
<tr>
<td></td>
<td><em>Schistosoma mansoni</em></td>
<td>AACCUGACCGGUUUCUGCAGGUAAUUGUGCAUG (37 nt)</td>
<td>(9)</td>
</tr>
<tr>
<td></td>
<td><em>Fasciola hepatica</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cestodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Echinococcus granulosus</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Echinococcus multilocularis</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Examples are provided from all phyla where trans-splicing has been described. Only a few species have been surveyed in Cnidaria, Rotifera and Chordata.

These examples and the information currently available on the use of the SL indicate that

i) SLs are not conserved across phyla (intriguingly, some similarity exists between rotifer SL and nematode SL1; (18)).

ii) The level of conservation within a given phylum is variable: the ‘taxon’-specific SL may be species-specific, as in trypanosomatids and plathyhelminths, or phylum-specific, as is the case of the nematode SL1 (see also Fig. 7.1).

iii) The proportion of transcripts processed by trans-splicing is also variable (see Section 1.1), and so is the number of SLs present in a given taxon:
- in trypanosomatids, the species-specific SL is added to the 5′-end of all mRNAs;
- in nematodes, a majority of transcripts is trans-spliced (65% in *C. elegans*); most of them receive the phylum-specific SL1. In the rhabditine lineage, the others receive one of the SL2 variants, which show some degree of species-specificity (83). SL2s are related to the resolution of operons and their use would be stage dependent (29, 63). As to the other lineages, it has recently been observed that operons and SL2-like SLs have evolved independently within the phylum. Thus, although all species appear to use trans-splicing for the resolution of operon-derived pre-mRNAs, the SL variant that is attached to the 5′-end of single gene mRNAs varies for the different nematode groups: the Rhabditina use SL2-like SLs, whereas some taxa utilise SL1 and others SL1-like variants (31).
libraries from small amounts of parasite material, thus making amenable to EST characterisation parasites and/or stages that are otherwise difficult to survey.

The taxon-specific SL tag also offers the possibility of isolating the genes expressed by a given organism or group of organisms from a mixed pool of species (e.g., an environmental sample). In such situations, the presence of the SL provides advantages similar to the ones pointed out for parasitic organisms: it allows the selective amplification of even small amounts of mRNA from the organism(s) to be analysed from mixtures that include large amounts of mRNAs from other species. It is therefore possible to uncover the genes expressed by eukaryotes carrying out trans-splicing coexisting with organisms not using this mechanism of transcript processing or, even, to analyse the trans-spliced transcriptomes of coexisting eukaryotes from different taxa (i.e., bearing different SLs; see Table 7.1). This type of approach was recently applied to the isolation of full-length dinoflagellate cDNAs from a natural assemblage without any prey-derived contaminants (13). The selection of SL-tagged mRNAs from samples of natural populations represents a potentially powerful analytical tool that remains to be fully explored.

1.3. Trans-Spliced cDNAs Provide Full-Length Transcripts

Full-length cDNAs contain invaluable information about a gene and its products (Note 3). They include not only the complete coding region but also the 5′- and 3′-UTRs, which provide information about the transcriptional start and regulatory motifs related to translation efficiency and stability of the mRNAs. The preparation of full-length-enriched libraries for the characterisation of ESTs is thus a major aim of sequencing projects. Aside from gene discovery, EST data obtained from such libraries also facilitate gene finding and the correct annotation of UTRs in genomic sequences.

Although SL-dT cDNAs are by definition full-length, they differ from full copies of non-trans-spliced mRNAs. More specifically, because the SL is added at acceptor sites at the 5′-UTR of pre-mRNAs, SL-cDNAs retain only a portion of the 5′-UTR; consequently, they do not enable the characterisation of transcriptional start sites (Note 4). An extreme situation is observed for some platyhelminth transcripts. Flatworm SLs include a 3′-terminal AUG, which has been identified as a putative translation initiator codon in mRNAs from various species (10, 22, 45, 48, 49); such SL-cDNAs would not contain any portion of the pre-mRNA 5′-UTRs (see Fig. 7.1b).

Because they contain the full-coding region, SL-dT cDNAs are valuable for the identification of secreted and membrane-bound proteins. These proteins generally carry a ‘signal sequence’ at their N-terminus [‘signal peptide’ or ‘signal anchor’, respectively (50, 51)] that directs their transport and localisation in the
cell. This is especially relevant for parasitic organisms because such molecules are likely to include signals and receptors, which are critical for maintaining infection. Hence, due to their importance, membrane receptors and secreted proteins typically form the focus of projects aimed at understanding parasite biology and/or identifying drugs and vaccine targets. EST projects were not usually viewed as a means of uncovering signal sequence-bearing proteins; various ‘selection’ strategies have instead been designed for such purposes (52–57). Although elegant, these approaches are laborious and often prone to isolation of false positives (58, 59).

In contrast, even small-scale EST projects allow the efficient identification of signal sequence-encoding genes when 5′-end enriched cDNA libraries are used for the generation of ESTs (45, 60) (Note 5).

1.4. Trans-Spliced cDNAs Versus ‘Oligo-Capped’ cDNAs

Taking into account that trans-splicing introduces a leader sequence at the mRNA 5′-end, the mechanism resembles the oligo-capping procedure (61), whereby an RNA oligo is ligated to the 5′-end of full-length mRNAs, a strategy that has been successfully applied to full-length cDNA cloning (62) (Note 6). In spite of this similarity, the products of trans-splicing and oligo-capping differ: trans-spliced mRNAs are capped (see Section 1.1), whereas oligo-capped mRNAs are not (the cap is actually replaced by the RNA oligo, which is ligated onto decapped mRNAs); in addition, oligo-capped mRNAs retain the full 5′-UTR of the parent mRNAs whereas, as already mentioned, trans-spliced mRNAs do not (see Section 1.3).

More importantly, for metazoan organisms that trans-splice a fraction of their transcripts, SL-dT libraries provide information on a subset of mRNAs; in contrast, oligo-capped libraries would include data from the whole mRNA set. This has actually been observed for Caenorhabditis elegans: 65% of cDNAs isolated from oligo-capped libraries contained SL-sequences (63), a figure that correlates well with the estimated proportion of mRNAs whose maturation involves trans-splicing in this organism (21, 29). However, we have found that trans-spliced mRNAs from the cestode Echinococcus granulosus are oligo-capped most inefficiently (45).

Although unexpected, this result indicates that, even in the case of organisms that trans-splice a minority of their mRNAs, the SL-dT approach constitutes a useful strategy to obtain full-length-enriched libraries. Furthermore, the EST datasets derived from SL-dT libraries may complement rather than overlap those obtained with cDNAs synthesised using strategies such as oligo-capping that lead to the isolation of high-quality cDNAs (Parkinson, Maizels and Fernández, unpublished data).

Finally, the protocol designed to prepare 5′-end-enriched libraries from oligo-capped mRNA (62) could be similarly applied to SL mRNAs. If random hexamers were used instead of oligo-dT for reverse transcription, it should be possible to
prepare 5'-end-enriched SL libraries to allow the characterisation of the 5'-ends from long (> 5 kb) trans-spliced mRNAs (‘SL-random hexamers approach’).

2. Materials

Reagents (including water) of the highest quality/purity should be used throughout and RNase-free during the steps that involve handling of RNA (Note 7). Materials should conform to similar standards. Whenever possible, use sterile disposable plasticware and barrier tips certified nuclease-free. The utmost care should be taken for every manipulation to avoid RNase contamination in the initial steps as well as damage/contamination of the material during the rest of the procedure. Be aware that how the library is constructed will ultimately determine the quality of the sequence information that may be extracted from the isolated clones. Below is a list of specific reagents we have successfully used to carry out the methods in the next section, organised according to the subheadings of the protocols.

1. Preservation of RNA in starting material: RNA later RNA Stabilisation Reagent (Qiagen) as the reagent of choice; Trizol Reagent (Invitrogen) or Buffer RLT from the RNeasy kit (Qiagen) as suitable alternatives.

2. Isolation of total RNA: Trizol Reagent (Invitrogen) and/or RNeasy kit (Qiagen).

3. DNase treatment: DNase I (RNase-free) (10 U/μl, Roche).

4. Carrier for ethanol precipitation of RNA and cDNA: glyco- gen (20 μg/μl, Molecular Biology Grade–Sigma; or Ultrapure- Invitrogen).

5. First-strand cDNA synthesis and directional cloning of cDNA: SuperScript Plasmid System for cDNA Synthesis and Plasmid Cloning kit (Invitrogen), as detailed below.
   i) First-strand cDNA synthesis: DEPC-treated water, Not I primer-adaptor (0.5 μg/μl), 5x first-strand buffer, 0.1 M DTT, 10 mM dNTP mix (10 mM each of dATP, dCTP, dGTP, dTTP), SuperScript II reverse transcriptase (200 U/μl, Note 8), E. coli RNase H (2 U/μl). RNase- OUT (40 U/μl, Invitrogen), not provided with the kit, should be included during first-strand synthesis.
   ii) Ligation of Sac I adaptors: 5x T4 DNA ligase buffer, Sac I adaptors (1 μg/μl), T4 DNA ligase (1 U/μl).
   iii) Not I digestion: REACT 3 buffer, Not I (15 U/μl).
iv) Column chromatography of cDNA: cDNA size fractionation columns (2 ml plastic columns containing 1 ml of Sephacryl S-500 HR).

v) Directional cloning of cDNA: 5x T4 DNA ligase buffer, NotI-SalI-cut pSPORT 1 (50 ng/μl), T4 DNA ligase (1 U/μl)

6. Transformation of the library: E. coli ELECTROMAX DH10B (Invitrogen, not included in the cDNA cloning kit).

7. PCR amplification of cDNA: Taq DNA polymerase (5 U/μl) (Qiagen) or Advantage 2 PCR kit (Clontech) (Note 9). Reverse NotI primer: 5’-TAGATCGGAGCGGCCGCCCTTT-3’ (10 pmoles/μl; HPLC-purified, custom order, Sigma Genosys); forward SL primer (10 pmoles/μl; HPLC-purified).

8. Polishing of PCR-amplified cDNA: PCR Polishing kit (Stratagene), including 10x Pfu reaction buffer and cloned Pfu DNA polymerase (2.5 U/μl).

9. Quality assessment of the library:
   i) Culture of recombinant clones: LB medium (10 g/l tryptone, 5 g/l yeast extract and 5 g/l NaCl) containing ampicillin (100 μg/ml).
   ii) Amplification of cloned inserts: PCR primers (M13 reverse: 5’-AGCGGATAACAATTTCACACAGG-3’ and M13 forward: 5’-CCCAGTCACGACGTTGTAAAACG-3’; 10 pmoles/μl; custom order, Sigma Genosys).
   iii) Cleaning of PCR products: Shrimp Alkaline Phosphatase (SAP, 1 U/μl; Amersham) and Exonuclease I (ExoI, 10 U/μl; Amersham).
   iv) Sequencing reactions: BigDye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems); sequencing primers (3.2 pmoles/μl; custom order, Sigma Genosys), either 5’-end sequencing primer (T7: 5’-TAATACGACTCACTATAGG-3’) or 3’-end sequencing primer (SP6: 5’-ATTTAGGTGACACTATAG-3’).

3. Methods

We describe below a general protocol for the construction of directional SL libraries using a PCR strategy for cDNA amplification. The outlined methods are based on our experience with the platyhelminth parasite E. granulosus (45). We presume that you intend to prepare a library using total RNA isolated from small amounts of material. As a rule of thumb, in such cases, it is
important to reduce the number of steps; for this reason, some procedures are marked as optional under the assumption that it is desirable but not essential to carry them out. Although obvious, we also presume that, unless otherwise specified, you stick to the protocols in the manuals supplied by the manufacturers of the recommended reagents.

3.1. Preservation of RNA in Starting Material

The quality of the final product depends on the quality of the starting material. Every care should thus be taken to maintain the integrity of the starting RNA. You may work with fresh material – straight after removal from the host, in case you are dealing with a parasite. If this is not possible, (i) freeze the material immediately and keep it frozen in liquid nitrogen or at –70°C (after snap-freezing in liquid nitrogen); alternatively, (ii) store it in RNAlater stabilisation reagent (this is very convenient because the RNA is stable for up to 7 days at room temperature, which allows safe transport of the sample); or (iii) homogenise (disrupt and lyse) the sample in a guanidinium isothiocyanate-containing reagent such as Trizol or buffer RLT (RNeasy kit) and store it at –70°C until use (see also Section 3.2; Notes 10 and 11).

3.2. Isolation of Total RNA

The starting RNA must be of the highest quality. We have been able to prepare RNA of suitable quality using two commercially available reagents: either Trizol or the RNeasy kit. These may also be applied sequentially to prepare high-quality total RNA (Note 12). We provide a protocol using Trizol, which is applicable to a wide variety of tissues.

1. Homogenise 50 mg of parasite tissue in 1 ml of Trizol reagent (it is convenient to use 1.5 ml Eppendorf tubes; if working with large volumes, use Falcon tubes instead) (Notes 13 and 14).
2. Remove insoluble material by centrifugation at 12,000x\(g\) for 10 min at 2 to 8°C. Transfer the cleared homogenate to a fresh tube and discard the pellet.
3. Add 0.2 ml of chloroform; cap the tube securely. Shake by hand for 15 sec and incubate at room temperature for 2 to 3 min.
4. Centrifuge at 12,000x\(g\) for 15 min at 2 to 8°C. Transfer carefully the upper aqueous phase to a fresh tube and discard the pellet.
5. Precipitate the RNA with 0.5 ml of isopropanol. Let stand for 10 min at room temperature (Note 16). Centrifuge at 12,000x\(g\) for 10 min at 2 to 8°C.
6. Discard the supernatant and wash the pellet with 1 ml of 75% ethanol. Centrifuge at 7,500x\(g\) for 5 min at 2 to 8°C.
7. Briefly dry the RNA pellet and dissolve it in up to 50 \(\mu\)l of RNase-free water (Notes 17 and 18).
3.3. Quality Assessment of the Isolated RNA

1. Estimate the yield and purity of the product from Step 3.2.7 by checking the absorbance ratios $A_{260}/A_{280}$ and $A_{260}/A_{230}$ of a suitable dilution (such as 1:50 or 1:100). Dilute the sample in water to determine the RNA concentration (an aqueous RNA solution of 40 µg/ml has an $A_{260}$ of 1.0). A solution of pure RNA has $A_{260}/A_{280}$ of 2.0 and $A_{260}/A_{230}$ higher than 2; however, expected values may occur outside that range. Trizol-isolated RNA diluted in water has $A_{260}/A_{280}$ of 1.6–1.8; lower ratios are typically due to protein contamination (Note 19). Nucleotides have minimum absorbance at about 230 nm; thus, the ratio $A_{260}/A_{230}$ provides an indication of contaminants such as carbohydrates or guanidinium thiocyanate carried over during the precipitation steps.

2. Analyse the integrity of the preparation by electrophoresing an aliquot (about 0.5 to 1 µg, if possible) in a 1.5–2% agarose gel; heat denature the sample at 65°C for 2–3 min before loading onto the gel (Note 20).

3.4. DNase Treatment (Optional; Note 21)

1. Remove the residual DNA with DNase I by combining 39 µl of the sample, 5 µl of 0.05 M MgSO$_4$, 5 µl of 1 M sodium acetate pH 5.2, 1 µl of DNase I (10 U/µl) and incubating for 1 h at 37°C.

2. Add 50 µl of RNase-free water and 100 µl of phenol-chloroform and vortex vigorously for 30 sec. Centrifuge at maximum speed in a microcentrifuge for 5 min at room temperature. Transfer the aqueous phase to a fresh tube (Note 22).

3. Ethanol-precipitate the RNA by adding 1 µl of 20 µg/µl glycogen, 10 µl of 3 M sodium acetate pH 5.2 and 220 µl of absolute ethanol (Note 23). To pellet RNA, centrifuge at 12,000xg in a microcentrifuge for 20 min at 4°C.

4. Remove the supernatant and rinse the pellet with 500 µl of 75% ethanol. Centrifuge at 12,000xg in a microcentrifuge for 2 min at 4°C. Air-dry the pellet for 1 to 2 min at room temperature.

5. Dissolve in 10 µl of RNase-free water (Note 18).

3.5. First-Strand cDNA Synthesis

1. Synthesise first-strand cDNA with SuperScript II reverse transcriptase in a 20 µl reaction volume by combining 10 µl of the sample (1 to 5 µg RNA; Note 24), 1 µl of 0.5 µg/µl of NotI primer-adaptor (Note 25), 1 µl of dNTP mix (10 mM each), 4 µl of 5x first-strand buffer, 2 µl of 0.1 M DTT, 1 µl of RNaseOUT (40 U/µl), 1 µl of SuperScript II RT (200 U/µl) (Note 8) and incubating at 42°C for at least 50 min (Note 26).

2. Inactivate the enzyme by heating at 70°C for 15 min.

3. Degrade the template RNA with 1 µl of RNase H (2 U/µl), by incubating at 37°C for 20 min (Note 27).
3.6. Quality Assessment of First-Strand cDNA (Optional)

In order to confirm the success of reverse transcription, it is convenient to assess the integrity of first-strand cDNA. For this purpose, if the necessary information about the organism is available, we recommend that you carry out PCR reactions using gene-specific primers to find out whether the cDNA preparation (i) includes 5'-ends from copies of long transcripts, and from transcripts of genes expressed at low level and (ii) is free from genomic DNA (using primers which span an intron and thereby give different size products from cDNA and genomic DNA). Set up 25 µl-PCR reactions with 0.5 to 1 µl of the first-strand cDNA preparation and amplify for 30–35 cycles.

3.7. PCR Amplification of SL-cDNA

1. Determine the conditions for the PCR amplification of SL-cDNA. Set up a 50 µl-PCR reaction with 1 to 2 µl of the first-strand cDNA preparation (the amount depends on the quality and quantity of the starting RNA). Combine the template with 31 to 30 µl of dH₂O, 5 µl of 10x Taq buffer, 10 µl of 1 mM dNTPs, 1 µl of 10 pmoles/µl –solution of each primer: forward SL primer and reverse NotI primer (Note 28) and 1 µl of Taq DNA polymerase (5 U/µl) (Note 9). Thermocycle for 15 to 20 cycles at 94°C, 1 min; 57°C, 1 min; 72°C, 5 min.

2. Analyse the profile of the SL-cDNA by checking an aliquot (10 to 20 µl) in a 1% agarose gel; include a MW marker with a quantified band, so as to estimate the yield of amplification. Determine the optimum number of cycles to PCR-amplify the SL-cDNA (Fig. 7.2a; Note 29).

3. Scale up the double-stranded cDNA synthesis to produce about 500 ng of cDNA: carry out several 50- to 100-µl reactions using the previously optimised conditions (Note 30).

4. After the PCR, combine the products and extract them with an equal volume of phenol-chloroform (as described in Step 2, Section 3.4). Carefully remove the aqueous phase and transfer it to a fresh tube (Note 31).

5. Ethanol-precipitate the cDNA with 7.5 M ammonium acetate (use 100 µl of salt solution followed by 600 µl of absolute ethanol for 180 µl of cDNA). Centrifuge at 12,000xg for 20 min at room temperature (Note 32).

6. Remove the supernatant and rinse the pellet with 500 µl of 75% ethanol. Centrifugate at maximum speed in a microcentrifuge for 2 min. Dry for 10 min at 37°C and proceed to Section 3.8 (Note 33).

3.8. Polishing PCR-Amplified cDNA

Taq polymerase possesses template-independent terminal transferase activity that results in the addition of a single unpaired nucleotide at the 3′-end of amplified fragments. For this reason, prior to ligation with the blunt-end of SalI adaptors, the cDNA must be
Fig. 7.2. Preparation and quality assessment of an SL-cDNA library: results of some steps. The picture includes agarose/ethidium bromide gels obtained when constructing an SL-cDNA library from *E. granulosus* larval worms (protostrongylus, PS) using the methods described in Section 3. (a) PCR amplification of SL-cDNA. Two PS samples were processed in parallel from equal amounts of starting RNA (4 μg); the gel shows the SL-cDNAs amplified from 1 μl of the product of reverse transcription after 18 or 20 cycles (20 μl from 50 μl PCR-reactions were loaded in each lane). *E. granulosus* SL-cDNA is visualised as a smear (~ 400 to 3000 bp) with some discrete bands. The gel highlights how the optimum number of cycles has to be determined for every preparation: PS1 was amplified for 18 cycles; 19 cycles were carried out for PS2 (see Note 29). (b) Size fractionation of SL-cDNA. SL-cDNA from PS1 was ligated to *Sal I* adaptors, digested with *Not I* and fractionated by column chromatography as described in Section 3.11; 3 μl aliquots of fractions 8 to 14 (eluted between 425 and 635 μl) were loaded onto the gel. Note that (i) In spite of the minute amounts (the 500 bp band of the marker contains 25 ng), faint smears of SL-cDNA are visualised from fraction 8. (ii) A slight decrease is observed in the size range of the material eluted between consecutive fractions. (iii) The adaptors (arrowed) are clearly visible in fractions 13 and 14: As indicated below the gel, two sub-libraries were prepared with the cDNA from tubes 8+9 and 10 (sub-libraries 1 and 2, respectively), which, according to the profile in the gel, would predominantly include cDNAs in the range 700 to 3000 bp. (c) Distribution of SL-cDNA insert sizes. Following the protocols in Section 3.13, 96 randomly isolated colonies from sub-library 1 were PCR-amplified using M13 forward and reverse primers (that yield PCR products about 200 bp longer than the corresponding inserts). Note that even if the fractions were pooled so as to maximise cloning of long cDNAs and avoid cloning of adaptors: (i) The majority of cDNAs have sizes between 500 and 1500 bp with only a few in the range 2000–3000 bp. (ii) Two recombinants (about 2%) contain cloned adaptors (arrowed). Because they carry *Not I* and *Sal I* ends, these may clone as ‘monomers’, ‘trimers’ (250 and 300 bp products in the bottom and top gels, respectively); in other sets of clones, we have also found ‘pentamers’ (400 bp PCR products).
treated with a polymerase with 3' to 5' exonuclease activity (such as T4 or Pfu DNA polymerases) to ensure that its termini are blunt. We provide a standard protocol using Pfu polymerase.

1. Dissolve the SL-cDNA in 7 μl of dH2O and combine with 1 μl of 10x Pfu buffer, 1 μl of 10 mM dNTPs and 1 μl of Pfu (2.5 U/μl).

2. Incubate at 72°C for 30 min. The end-polished cDNA may be added to the ligation reaction without further purification.

3.9. **Ligation of SalI Adaptors to SL-cDNA**

1. Set up a 50-μl reaction to ligate the polished cDNA to SalI adaptors: combine the 10 μl of SL-cDNA from Step 3.8.2 with 19 μl of dH2O, 10 μl of 5x T4 DNA ligase buffer, 6 μl of SalI adaptors (1 μg/μl; Note 34) and 5 μl of T4 DNA ligase (1 U/μl). Incubate the reaction at 16°C for a minimum of 16 h.

2. Extract with an equal volume of phenol-chloroform (as described in Step 2, Section 3.4). Carefully remove the aqueous phase and transfer it to a fresh tube.

3. Ethanol-precipitate the cDNA with 7.5 M ammonium acetate (as described in Steps 5 and 6, Section 3.7; use 25 μl of salt solution and 150 μl of absolute ethanol). Dry the pellet and proceed to Section 3.10.

3.10. **NotI Digestion of SalI-Adapted SL-cDNA**

1. Set up a 50-μl reaction to digest the ligated cDNA with NotI: dissolve the SL-cDNA in 41 μl of dH2O and combine with 5 μl of REACT 3 buffer and 4 μl of NotI (15 U/μl). Incubate for 2 h at 37°C.

2. Extract with phenol-chloroform and ethanol precipitate, as described in Steps 2 and 3, Section 3.9. Dry the pellet and proceed to Section 3.11.

3.11. **Size Fractionation of SalI-adapted NotI-digested SL-cDNA**

1. Use a cDNA fractionation column to remove residual SalI adaptors and the NotI fragments of the primer-adaptor released by restriction digestion as well as to select adapted digested cDNAs >500 bp. **Strictly** follow the protocol and recommendations of the manufacturers (Note 35).

2. Run a 1% agarose gel with an aliquot (about 3 μl) of each fraction eluted between 350 and 550 μl; include a MW marker with a quantified band (about 10 ng) to estimate the amount in every fraction (cDNAs > 500 bp usually elute within that range; you may check fractions eluting before 350 μl and after 550 μl; the latter helps visualising the adaptors to ensure you do not include them). Analyse the profile of the eluted fractions to decide which ones to include in the library and how/whether to pool them; take into account that about 10 ng of cDNA should be used for ligation to the vector (Fig. 7.2b; Note 36).

3. Pool the selected fractions and ethanol-precipitate the cDNA with glycogen as carrier (as described in Steps 3 and 4, Section 3.4; Note 37). Dry the pellet and dissolve in a maximum of 14 μl of dH2O.

1. Ligate the fractionated cDNA from Step 3, Section 3.11 to NotI-Sall-cut vector using 20 μl reactions: combine an adequate volume of cDNA (about 10 ng) with 4 μl of 5x T4 DNA ligase buffer, 1 μl of NotI-Sall-cut pSPORT 1 (50 ng/μl) and 1 μl of T4 DNA ligase (1 U/μl). Incubate for 3 h at room temperature or overnight at 4°C.

2. Ethanol-precipitate the ligation reaction with glycogen as carrier (as described in Steps 3 and 4, Section 3.4; Note 37). Dry the pellet and dissolve the ligated cDNA in 3 to 5 μl of dH2O.

3. Use 1 μl of the library to transform competent E. coli ELECTROMAX DH10B by electroporation following the manufacturer’s protocol.

3.13. Quality Assessment of the Library

The following protocols are based on those described by Whitton et al. (64). They provide a simple and efficient way of checking the quality of the library through analysis of 100 to 200 isolated colonies (handled in 1–2 microtiter plates). Depending on the outcome of the initial assessment, the same protocols can be used to carry out a small-scale EST project.

Analysis of the distribution of insert sizes

1. Randomly pick independent colonies into 200 μl of LB with ampicillin in 96-well microtiter plates. Incubate overnight at 37°C.

2. Transfer 100 μl of the cultures into a new 96-well PCR plate. Add an equal volume of LB/glycerol 60/40 (v/v) to the remainder of the overnight cultures and freeze the original plate at −70°C (glycerol stocks of isolated clones).

3. Spin down the cells in the new plate and discard the supernatant. Add 50–100 μl of dH2O to the pellets, and lyse the cells by heating at 95°C for 5 min in a PCR block. Spin down the cell debris. Use 5 μl of the supernatants to set up 20 μl-PCR reactions using M13 forward and reverse primers.

4. Analyse 3–5 μl of the PCR products in a 1% agarose gel and evaluate the distribution of insert sizes in the library (Fig. 7.2c).

3.14. Sequence Analysis of Sampled Inserts

1. Clean the remaining 15 μl of the PCR products by treating them with 1 μl of SAP (1 U/ml) and 1.5 μl of ExoI diluted 1/10 in 50 mM Tris-HCl pH 8.0. Incubate the plate at 37°C for 40 min and then at 94°C for 15 min in a PCR block.

2. Set up the sequencing reactions by combining 4 μl of ExoSAP-treated PCR products with 3 μl of 1x sequencing buffer, 1 μl of T7 sequencing primer (3.2 pmoles/μl; Note 38) and 2 μl of BigDye; thermocycle the plate following the instructions of the manufacturer.

3. Determine the sequences (usually good-quality reads of >500 nt); process and analyse the data in individual sequences; check for the presence of the SL, open reading frames (ORF) and putative starting methionine and potential signal sequence (Table 7.2);
### Table 7.2
Comparative analysis of *Echinococcus granulosus* SL-cDNA libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Titer (cfu/µg RNA)</th>
<th>ESTs</th>
<th>Host DNA (gDNA+cDNA)</th>
<th>Median insert size (range) bp&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Putative start Met + ORF&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Possible start Met from SL&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Putative SignalSequence&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSU-SL</td>
<td>~ 1 x 10&lt;sup&gt;5&lt;/sup&gt;</td>
<td>95</td>
<td>0</td>
<td>700 (350-1800)</td>
<td>87 (92%)</td>
<td>47 (54%)</td>
<td>26 (27%)</td>
</tr>
<tr>
<td>PSP-SL</td>
<td>~ 7 x 10&lt;sup&gt;5&lt;/sup&gt;</td>
<td>86</td>
<td>0</td>
<td>750 (350-1800)</td>
<td>66 (77%)</td>
<td>30 (45%)</td>
<td>10 (12%)</td>
</tr>
<tr>
<td>CW-SL</td>
<td>~ 8 x 10&lt;sup&gt;5&lt;/sup&gt;</td>
<td>104</td>
<td>0</td>
<td>1050 (400-2800)</td>
<td>92 (88%)</td>
<td>42 (46%)</td>
<td>17 (16%)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Insert sizes were estimated by electrophoresis analysis of the corresponding PCR products, as described in **Section 3.13** (see also Fig. 7.2c).

<sup>b</sup>An ATG towards the 5′-end was considered a putative starting Met if it was followed by an open reading frame of at least 99 nt. When available, similarity with an orthologous protein was also considered.

<sup>c</sup>Putative starting codons contributed by the *E. granulosus* SL (see **Table 7.1**). Percentage is of sequences with identified start sites.

<sup>d</sup>A signal sequence was considered to be present when at least two parameters scored positive using the SignalP algorithm (http://www.cbs.dtu.dk/services/SignalP/).

Note that host contamination was successfully excluded from the libraries (the same was true for copies of mitochondrial transcripts; see **Section 1.2**). Modified from Fernández et al. (45).

Abbreviations: cfu, colony forming units; gDNA, genomic DNA.

Library codes: PSU, Untreated protoscoleces; PSP, Pepsin/HCl-activated protoscoleces; CW, Cyst Wall.
perform BLAST searches, etc. From the global data, you will get an idea of the transcripts that are highly expressed by the organism(s)/stage(s) and of the redundancy of the library, which is also an indication of its suitability for EST generation. In the case of the libraries in Table 7.2, because the primary assessment showed they were of reasonably good quality, further sequencing was carried out – about 1,500 ESTs from each library.

4. Notes

1. Recently, a global approach has been launched for the characterization of SL-mRNAs from a non-parasitic organism, the chordate Ciona intestinalis (17).

2. We have encountered such problems when working with material from parasites of the genus Echinococcus [(45); Fernandez and Brehm, unpublished observations]. A predominance of polyadenylated large subunit mitochondrial rRNA has been reported for other organisms, notably Drosophila melanogaster (65) and the platyhelminth Fasciola hepatica (66).

3. We refer to protein-coding genes; in any case, it is worth noting that the generation of ESTs from full-length-enriched cDNA libraries has recently emerged as an excellent source of information about mRNA-like non-coding RNAs (67–70).

4. The reason for making this point is that 5′-ESTs generated from full-length cDNA libraries are considered good resources for the large-scale identification of transcriptional start sites (71, 72). Such information cannot be obtained from full-length SL-cDNA libraries.

5. The advantages provided by the characterisation of ESTs from 5′-end-enriched cDNA libraries for the identification of signal-sequence encoding genes are by no means restricted to parasitic organisms. Large-scale initiatives have recently been undertaken with this purpose using human cDNAs (73, 74).

6. Two chapters in previous volumes of this series contain an outline of the principles and detailed protocols for the preparation of cDNA libraries using oligo-capping. Please, refer to them for further information on that strategy (72, 75). In addition, two chapters in the current volume (Chapters 4 and 5 by Motoaki Seki and Yoshihide Hayashizaki, respectively) refer to the cap-trapper procedure, another method for the
construction of full-length-enriched cDNA libraries; this approach is based on the selective recovery of full-length cDNAs by means of streptavidin purification of cDNA/mRNA hybrids after biotin-labelling of the mRNA cap and RNase I treatment (76).

7. Authors appear to be divided regarding the use of DEPC, the ‘classic’ inhibitor of RNases: some recommend using it always to treat water and rinse glassware, whereas others say it should never be used. Either with or without DEPC treatment, make sure you use RNase-free reagents and materials; if employing DEPC, also make sure you eliminate it properly.

8. We have also obtained excellent results with PowerScript reverse transcriptase (Clontech); this enzyme, which is also devoid of RNase H activity, may be used instead of SuperScript II.

9. In principle, we recommend using Tag DNA polymerase, which so far appears to have an unparalleled performance for the amplification of complex templates. We have also obtained very good results with a mixture of Tag and a proof-reading polymerase (the enzyme mix from the Expand Long Template PCR system, Roche). This system yielded slightly longer cDNAs than Tag (see also Note 36).

10. Tissues can be stored for several months at –70°C; nevertheless, it is usually safer to keep samples in a stabilising medium, either RNA later or a guanidinium isothiocyanate-containing solution.

11. Although the suppliers do not specifically recommend it for such purpose, we have also been able to isolate good-quality RNA from samples that had been transported at room temperature, after homogenisation in Trizol (see also Note 13).

12. Trizol uses the popular acid-guanidinium isothiocyanate-phenol-chloroform method that, according to Suzuki and Sugano (72), yields RNA containing a lot of fragmented RNA and also genomic DNA. These authors recommend purifying Trizol-isolated RNA with the RNeasy kit, which includes a silica gel-based membrane that selectively binds RNA in the presence of high salt.

13. Homogenisation of samples is critical to ensure a good recovery of RNA. You should consider which method is better suited in your case. Soft tissues tend to be easily disrupted and homogenised by pipetting up and down in Trizol. An efficient procedure that is also useful for small amounts of material is grinding the sample to a fine powder under liquid nitrogen using a mortar and pestle. The RNeasy handbook includes highly instructive descriptions of various alternatives.
14. It is important to determine the yield of your material in order to estimate the correct (and the minimum) amount of starting sample. In principle, we recommend that you use at least 1 μg of total RNA for the preparation of the library. Unless this is not possible, we strongly advise you to store the material until you have gathered enough (see also Note 17).

15. Re-extraction with Trizol prior to precipitation of the RNA reduces DNA contamination, and may provide an alternative to DNase treatment (Section 3.4) if working with small amounts of material. Add 1 ml of Trizol to the aqueous phase and repeat the extraction. Adjust the volume of isopropanol in Step 5, Section 3.2, accordingly.

16. Some parasite materials are peculiar animal tissues and require customised modifications of standard protocols. One such material is the wall of hydatid cysts, the larvae of *E. granulosus* – precipitation of Trizol-isolated RNA was carried out in the presence of high salt (0.25 ml of isopropanol + 0.25 ml of a solution containing 0.8 M sodium citrate and 1.2 M NaCl were added to the aqueous phase), as recommended to prevent co-precipitation of contaminating carbohydrate-rich molecules (45). Similarly, when using the RNeasy kit, binding of RNA to the membrane was carried out with the conditions recommended for plant tissue (450 μl RLT + 225 μl of 100% ethanol) instead of those for animal tissue (600 μl of RLT + 600 μl of 70% ethanol).

17. The yield of RNA very much depends on the quality of the sample and the type of material being processed. Working with *E. granulosus*, 50 mg wet tissue usually yields about 5 μg of total good-quality RNA in the case of the cyst wall, immature adults or pepsin/H⁺-activated larval worms; but only 2 μg for untreated worms. These amounts are far below those recovered from mammalian tissues (>1 μg/mg).

18. The ‘RNA storage solution’ (Ambion) provides an excellent alternative to water, especially if you will not proceed immediately with cDNA synthesis. It is a 1 mM sodium citrate solution of pH 6.4; base hydrolysis is minimized by the low pH and the chelating effect of citrate on free cations. We prefer not to use water and to avoid the standard 0.1 mM EDTA or TE buffer for long-term storage of RNA.

19. The $A_{260}/A_{280}$ is considerably influenced by pH and ionic strength – low ionic strength and low pH increase $A_{280}$. For this reason, some authors recommend measuring absorbance in 10 mM Tris–HCl pH 7.5, so as not to lose sensitivity in the detection of protein contamination. Purified RNA has an $A_{260}/A_{280}$ ratio of 1.9–2.1 in this buffer (77).
20. The sharpness of the rRNA bands provides an indication of the quality of the preparation. Be aware that manual descriptions of ‘expected profiles’ derive from data of model species, and that other organisms may have ‘atypical’ profiles. For example, in most cases, the 28S rRNA band is present at twice the intensity of the 18S rRNA band. In *Echinococcus* spp. and other platyhelminths, 28S rRNA is naturally nicked; when denatured, it co-migrates with 18S rRNA (78); thus, heat-denatured Echinococcus rRNA is visualised as a single sharp band.

21. Although the SL-dT approach should yield cDNA free from genomic contamination, it is convenient to remove residual DNA. This step may be omitted when working with very small amounts of material (see also Note 15).

22. Instead of the phenol-chloroform extraction, in case an RNeasy kit is available, you may follow the RNA clean-up protocol to remove DNase. Depending on the amount of material, it may be necessary to ethanol-precipitate the RNA and re-suspend it in up to 10 μl before proceeding to cDNA synthesis.

23. The addition of glycogen significantly increases the recovery of small amounts of nucleic acids by ethanol precipitation. Glycogen does not interfere with most downstream applications of RNA and DNA; it is insoluble in ethanolic solutions and forms a precipitate that traps the target nucleic acids. Where dry ice is available, it may be convenient to freeze the tube for 10 min to facilitate precipitation. Alternatively, the tube may be kept at −70°C or −20°C overnight.

24. Up to 500 ng of mRNA may be used instead of 5 μg of total RNA. If enough material is available, purify polyA⁺-RNA using one of the many commercially available kits. Be aware that oligo-dT-purified RNA may still contain large amounts of ‘irrelevant’ polyadenylated transcripts (see Section 1.2 and Note 2).

25. To prepare a 5'-end-enriched SL library, use 1 μl of random hexamer adaptor-primers instead of the tagged oligo-dT (NorI-adaptor-primer); allow the reverse transcription reaction to proceed at 12 to 25°C for 1 h, prior to incubation at 42°C.

26. Suzuki and Sugano (72, 75) recommend incubating at 42°C for more than 3 h to ensure that reverse transcription is completed.

27. The template RNA may also be degraded by treatment with NaOH at 65°C. We have not tried such an alternative, but it is included in the protocols by Suzuki and Sugano (72, 75).
28. Although obvious, these conditions presume that the 5’-SL primer forms a good set with the NotI 3’-primer, and is compatible with a cycling programme that uses 57°C as annealing temperature. For example, we used the oligonucleotide 5’-CACCGTTAATCGGTCCCTACCTT-3’ as SL primer, when working with Echinococcus spp. (22).

29. The number of cycles critically determines the quality of the final library; you should aim to use the minimum giving a ‘good’ yield of cDNAs, while minimizing the accumulation of copies from abundant and/or low molecular weight transcripts that tend to be preferentially amplified. As you are probably aware, different thermocyclers may yield different results; depending on your PCR machine, you may have to modify the PCR parameters. To determine the optimum number, we recommend that you check the yield after 15, 17 and 19 cycles (or 14, 16, etc; see Fig. 7.2a). To avoid wasting first-strand cDNA, you may take samples of 10 µl after ‘n’ cycles, and check them on a gel by comparison with an equal volume after ‘n + 2 to 3’ cycles. Apart from determining the optimum number of cycles, this allows you to estimate the amount of amplified cDNA/reaction.

30. Setting up several tubes in parallel under the optimal conditions and pooling them minimizes the chance of a single PCR error being propagated in the library. In the case of E. granulosus, depending on the quality of the starting material, 2 to 4 independent 100-µl PCR reactions have yielded about 500 ng of SL-cDNA, using the product reverse transcribed from 1 µg of total RNA.

31. Instead of the phenol-chloroform extraction, if a PCR clean-up kit (Qiagen) is available, you may use it to purify the amplified SL-cDNA.

32. If working with small amounts of material, use the protocol described in Steps 3 and 4, Section 3.4 that includes a carrier for ethanol precipitation.

33. If you have succeeded in preparing good-quality SL-cDNA (in terms of size distribution and presence of long and lowly expressed transcripts), but the total amount is below the one we recommend (about 400–500 ng), do not move into the following steps whose purpose is to prepare the cDNA for directional cloning. Directional libraries are valuable cDNA collections in the context of EST projects (see Note 38); however, because their construction involves several steps, a certain amount of material is inevitably lost while carrying them out. Thus, if the total amount of cDNA is low, we advise you to ligate it to a general-purpose PCR cloning vector [TA vector, such as TOPO-TA (Invitrogen) or pGEM-T
A small, good-quality library was prepared by TOPO-TA cloning of *E. multilocularis* SL-cDNA (22). TA vectors are designed for one-step cloning of PCR products prepared with *Taq* DNA polymerase that, due to its terminal transferase activity, adds an extra deoxyadenosine (dA) at the 3'-ends of amplified molecules. The plasmids are supplied linearised and carry 3'-deoxythymidine (dT) overhangs at both ends, thus efficiently ligating to PCR fragments with 3'-dA overhangs.

34. We recommend that you reduce the amount of *SalI* adaptors in the reaction to 6–7 μl (instead of 10 μl, as recommended by the manufacturers), so as to minimise the amount of cloned adaptors in the library. These tend to appear, even if you follow exactly the protocol for the column fractionation (Section 3.11; see Fig. 7.2c), especially if you are working with a small amount of cDNA.

35. Size fractionation of cDNA is another factor critically determining the quality of the final library (see also Notes 29 and 36). We have found that, when handling small amounts of material, column chromatography is safe, and provides a reasonably good selection of cDNAs >500 bp. The procedure may be carried out twice to ensure exclusion of small fragments (41). Alternatively, cDNA fractionation may be achieved by agarose gel electrophoresis (72, 75, 79). This is probably the method of choice when dealing with higher amounts, because it allows better size resolution; 2D-agarose gels have also been proposed as an improved alternative (80).

36. The other factor critically determining the quality of the final library is the cloning step. We recommend that you prepare at least two sub-libraries using cDNA from different fractions or pooled fractions (such as 8+9 and 10+11; or 8+9 and 10, in the example shown in Fig. 7.2b) by setting up independent ligation reactions. Ideally, ligation of cDNA to the vector should be carried out under conditions allowing large fragments to be cloned as efficiently as smaller ones. This is a major technical difficulty, because larger cDNAs do not compete well with smaller ones during the ligation step (especially if they correspond to genes expressed at low level). The insert size in the library is thus determined by the performance of the cloning step in such a way that the longer cDNAs generated during PCR amplification may not be represented in the final library. Setting up independent ligations should decrease the competition between the different sizes and maximise the chances of cloning long cDNAs (79, 80) (see Fig. 7.2c).

37. You may alternatively use 5 μl of 1 μg/μl yeast tRNA, and 0.5 volumes of 7.5 M ammonium acetate followed by 2 volumes of absolute ethanol.
38. At this stage, we recommend that you sequence the clones from the 5'-end only. Because sequencing from the 3'-end may be problematic due to the presence of the poly-A tract, it is usually carried out at a later stage, once the library has been found to be of suitable quality for larger scale EST generation.

Acknowledgments

The protocols we describe were optimised during a visit of CF to RMM’s laboratory with the support of the Wellcome Trust (International Travelling Research Fellowship, Ref 061168). The authors are grateful to Dr. Gustavo Salinas (Universidad de la Republica, Uruguay) for critical reading of the manuscript and fruitful discussions throughout its preparation.

References


Expressed sequence tags (ESTs) have proven to be one of the most rapid and cost-effective routes to gene discovery for eukaryotic genomes. Furthermore, their multipurpose uses, such as in probe design for microarrays, determining alternative splicing, verifying open reading frames, and confirming exon/intron and gene boundaries, to name a few, further justify their inclusion in many genomic characterization projects. Hence, there has been a constant increase in the number of ESTs deposited into the dbEST division of GenBank. This trend also correlates to ever-improving molecular techniques for obtaining biological material, performing RNA extraction, and constructing cDNA libraries, and predominantly to ever-evolving sequencing chemistry and instrumentation, as well as to decreased sequencing costs. This chapter describes large-scale sequencing of ESTs on two distinct platforms: the ABI 3730xl and the 454 Life Sciences GS20 sequencers, and the corresponding processes of sequence extraction, processing, and submissions to public databases. While the conventional 3730xl sequencing process is described, starting with the plating of an already-existing cDNA library, the section on 454 GS20 pyrosequencing also includes a method for generating full-length cDNA sequences. With appropriate bioinformatics tools, each of these platforms either used independently or coupled together provide a powerful combination for comprehensive exploration of an organism’s transcriptome.

**Key words:** Expressed sequence tags, EST, sequencing cDNA, capillary sequencing, ABI 3730xl, pyrosequencing, GS20.

1. Introduction

The number of ESTs in dbEST that were submitted over the past 10 years greatly correlates to the evolution of DNA sequencing techniques and to the cost of sequencing. DNA sequencing procedures date back to the mid-1970s and of the two early
techniques: Maxam-Gilbert chemical analysis method \((I, 2)\) and Sanger’s enzymatic procedure employing chain-terminating nucleotides \((3)\), only the latter is in routine application for high-throughput sequencing today. Since their introduction, numerous technological refinements in DNA sequencing have been reported, aimed at producing more information per sequencing experiment, and increasing instrument throughput/capacity. The sequencing of cDNAs was subsequent to the introduction, in 1986, of fluorescent DNA labels for the Sanger sequencing, and the use of thermostable polymerases that enabled a “cycled sequencing” protocol to be performed. Once the cycle sequencing reaction is completed, the sample is analyzed using a capillary-based DNA sequencer such as the Applied Biosystems instruments.

Due to the proliferation of large-scale DNA-sequencing projects in recent years, and the thrust to re-sequence entire human genomes in the context of better understanding disease-specific mutations, alternative sequencing methods to reduce time and cost have become available. One such “massively parallel” sequencing platform is the GS20 sequencer from 454 Life Sciences. This is a sequencing system that offers a 100-fold increase in throughput over the current state-of-the-art Sanger sequencing technology on capillary electrophoresis instruments \((4)\). The apparatus uses a novel fiber-optic slide (PicoTiterPlates) of individual 40 \(\mu\)m wells in which beads containing individual DNA fragments, amplified by an emulsion PCR step, are subjected to sequencing by synthesis using a pyrosequencing protocol optimized for solid support and picoliter-scale volumes.

Constantly improving methods for tissue extraction, RNA/mRNA isolation, and cDNA library construction, in addition to the DNA sequencing improvements, have contributed to a constant increase in the number and quality of ESTs in dbEST. By combining methods for generating full-length cDNA libraries with the increased read lengths from capillary sequencers, it has been possible to capture a larger portion of the transcript through single-pass sequencing. Discussion of different protocols used to generate cDNA libraries or approaches to the analysis of the generated sequences are beyond the scope of this chapter, but are covered in Chapter 2. In addition, detailed methods used to generate different cDNA libraries from a variety of sources are described in Chapters 3–7.

In this chapter, we will provide protocols for sequencing cDNAs on two distinct sequencing platforms: the Applied Biosystems ABI 3730xl and the 454 Life Sciences GS20. We will also briefly describe the basecall extraction, sequence processing, and data submission procedures for sequences generated on both platforms.
2. Materials

With suitable laboratory infrastructure, ESTs can be quickly and inexpensively generated in large numbers. This chapter describes the high-throughput sequencing of ESTs as performed at the Washington University Genome Sequencing Center. The robots and ancillary instruments are not recommended, but rather reflect the ones used in this center at the time of writing. Similarly, the programs used for the instruments are specific to our LIMS, and should therefore be replaced with what is commercially available or modified to accommodate your own LIMS setup.

2.1. Plating

1. Large agar plates with appropriate antibiotic (the example in this chapter uses Kanamycin 0.07 ug/ml; IPTG 0.25 ug/ml). The antibiotic depends on the antibiotic-resistance gene encoded in the plasmid vector used for cloning
2. SOB media (used as a cell growth medium to ensure maximum transformation efficiency)
3. GC10 Thunderbolt Electro competent Cells (Gene Choice) (GC10 Thunderbolt Electro Competent Cells are comparable to the DH10B strain and carry recA1 and endA1 mutations that aid in plasmid stability and improved quality of prepared plasmid DNA. GC10 Competent Cells are offered at a high-efficiency grade for subcloning and generation of cDNA libraries.)
4. Plasmid ligation dilution
5. Broad stainless steel plate spreader
6. 1 mm cuvettes
7. 1.7 ml microfuge tube
8. Electroporator (BIO-RAD GenePulser Xcell)
9. 70% ethanol and reservoir
10. Electric hot-air dryer
11. 10% bleach (General Distribution).

2.2. Picking

1. Costar 384-well microtiter trays
2. 1% bleach (General Distribution)
3. 70% Ethanol
4. Sterile dH₂O
5. Media with antibiotic (Terrific Broth with Kanamycin, TB+Kan)
6. WellMate microplate dispenser (Matrix Technologies Corporation) with sterile tubing
7. Clear adhesive seals 5.5” × 3 3/16” with tab on 3 3/16 side (Grainger Inc, cat. # 809S).

2.3. Prepping
1. 384-well overnight growth archives (plates that have been picked and incubated as described in Section 3.2)
2. Eppendorf 384 well trays
3. Archive Prep Homogenous Solution (per well: 10 μl 19.5% PEG/1.65 M sodium chloride, 8μl Eppendorf solution 2, 2μl prepared beads)
4. Dilute lysol solution: 1:200
5. dH₂O
6. Gravity feed reservoir (in-house made reservoir)
7. Archive Prep Magnets (Magnaworks)
8. Diapers and Qtips (Fisher Scientific).

2.4. Sequencing
Cocktail for EST sequencing:
1. Bulk Big Dye Cocktail FWD primer (the primers depend upon the cloning vector): ABI Big Dye V3.1 mix; ABI 5X buffer (400 mM Tris-HCl pH 9.0/10 mM MgCl₂), FWD primer @ 10 μM Stock
2. Bulk Big Dye Cocktail REV primer (the primers depend upon the cloning vector): ABI Big Dye V3.1 mix; ABI 5X buffer (400 mM Tris-HCl pH 9.0/10 mM MgCl₂), REV primer @ 10 μM Stock
3. Sigma water (cat # 3501)
4. Beckman Coulter BioMek FX 384 pipette tips (P30 sterile)
5. 384 well V-groove reservoirs for Big Dye cocktails and precipitation
6. Q-Fill (Genetix) or Multidrop (Thermo Electron)
7. Floor Vortexer (Lab-Line Instruments, Inc., Titer Plate Shaker, Model # 4625)
8. Velocity 11 Plateloc Sealer with sealing tape (Velocity 11 Plateloc pierceable heat seal)
9. Precipitation Premix (100:1 solution of 100% EtOH/1ul NaOAc pH 5.2), 10 μl per reaction
10. 70% ethanol
11. ddH₂O.
2.5. Loading the ABI 3730xl Analyzer

1. 1X concentration 3730 Buffer with EDTA (Applied Biosystems)
2. dH₂O
3. POP7 polymer bottle (Applied Biosystems)
4. 96- or 384-well septa
5. 96- or 384-well plate bases and retainers.

2.6. Full-Length cDNA Preparation for 454 Life Sciences GS20

1. PowerScript Reverse Transcriptase (100 mM DTT, 5x First Strand Buffer, PowerScript Reverse Transcriptase) (Clontech, cat# 639500)
2. Advantage 2 Polymerase Mix (Advantage 2 50X Polymerase Mix, Advantage 2 10X PCR Buffer) (Clontech, cat# 639202)
3. PCR Nucleotide Mix, 10 mM each dNTP – 2,000 µl (Roche, cat# 11814362001)
4. RNaseOUT Recombinant Ribonuclease Inhibitor, 40 units/µl (Invitrogen, cat# 10777-019)
5. Dynabeads M-280 Streptavidin, 10 mg/ml (Invitrogen, cat# 112-06D)
6. Bovine Serum Albumin (BSA), 25 mg–10 mg/ml (New England BioLabs, cat# B9001S)
7. MmeI, 500 units–2,000 units/ml (New England BioLabs, cat# R0637L)
8. MB Dilution Buffer (10 mM Tris Acetate pH 7.8, 1 mM EDTA pH 8.0)
9. 2xBinding and Washing (B&W) Buffer (For M-280 bead prep) [10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl]

<table>
<thead>
<tr>
<th>5’ Smart_Oligo</th>
<th>5’- AAGCAGTGGTAA CAACGCATCCGA CGCrGrGrG-3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>3’Oligo_dT_SmartIIA</td>
<td>5’- AAGCAGTGGTAAACAACG CATCCGACTTTTTTTTT TTTTTTTTTTVN-3’</td>
</tr>
<tr>
<td>NEW_SmartIIA</td>
<td>5’- Biotin-TEG- AGCAGTGGTAAACAACGCATCCGAC -3’</td>
</tr>
</tbody>
</table>

2.7. 454 Life Sciences GS20 Sequencing Run

1. 50% ethanol
2. Bead Buffer 2 (454 Life Sciences).
2.8. Sequence Extraction, Processing, and Submission

The PC for running the ABI 3730xl or the 454 G20 is provided with the instrument.

1. For EST sequence processing, we recommend a Dell Precision 470n with the following specifications (Note 1):
   - 2.80 GHz Xeon processor
   - 2 Gb RAM
   - 80 Gb hard drive
   - Debian Linux running KDE
2. Internet connection

3. Methods

3.1. Plating

This protocol explains GC10 cell transformation by electroporation and plating the transformed cells onto large (20 × 20 cm) agar plates. Our experience is that transformation by high-voltage electroporation is an effective method for transforming E. coli with plasmid DNA.

3.1.1. Preparing Materials

1. Retrieve the number of agar plates needed for plating (Kanamycin + IPTG), and place them in a 37°C incubator with the lids ajar for about 1 hour to dry. Once dry, remove from incubator and label them accordingly.
2. Retrieve the required number of tubes of GC10 electro-competent cells and immediately place on ice to thaw for a minimum of 15 min before using. Each tube contains 100 μl of cells.
3. Place the appropriate number of 1-mm cuvettes on ice to pre-chill (1 cuvette/tube is used under normal plating conditions).
4. Add 900 μl of SOB media to each GC10 electro-competent cell tube and replace lid. Keep the cells on ice during this time.

3.1.2. Electroporation Procedure

1. Add 20 ul of GC10 cells to each pre-chilled cuvette (cell amounts are determined by the type of plating and plating conditions).
2. Add the required amount of cDNA ligation dilution to the GC10 cells in each cuvette. Change tips between each cuvette to avoid contaminating the ligation mixture. Keep the cuvettes on ice during these additions. (If bubbles exist in the cells/dilution mixture inside the cuvette, tap the cuvette gently on the bench top to eliminate the bubbles).
3. Quickly place each cuvette in the electroporator pod (ensure that the electroporator is set to a program that will pulse at 2000 Volts, capacitance 25 μF, resistance 200 ohms).

4. Following the pulse delivery, immediately add 100 μL of SOB media to the cuvette. The cuvette should rest on ice for 5 min. Transfer all contents of the cuvette into a 1.7-ml microfuge tube containing 900 μl of SOB media.

5. Follow Steps 2.1–2.5 for all ligation dilutions.

6. When all dilutions have been electroporated, place all 1.7-ml microfuge tubes in a 37°C water bath for no longer than 30 min. After 30 min remove tubes from the water bath and prepare for the spreading procedure.

3.1.3. Spreading Procedure

1. Fill a reservoir (a clean tip box lid will work) with 70% ethanol and place on the bench in a safe, but accessible location.

2. Dip the stainless steel spreader in 70% ethanol and shake excess ethanol off. The stainless steel spreader is held to a hot-air dryer for no less than 15 seconds. Disposable plastic spreaders may also be used, without the need to ethanol sterilize if a clean plastic spreader is used for each plate.

3. Extract entire contents of 1.7-ml microfuge tube, transfer to the surface of a dry agar plate and, using the sterile spreader, gently spread the tube contents across plate surface evenly, just until the liquid has absorbed.

4. Follow Steps 3.1–3.3 for each microfuge tube and agar plate until completed.

5. Place all agar plates, inverted, into the 37°C incubators for a 18 to 22-hour growth time.

3.2. Picking

These instructions assume the use of a Genetix Qpix II instrument for picking and the WellMate (Matrix Technologies Corporation) instrument for filling the plates with media. This protocol explains how to pick E. coli colonies containing plasmids from agar plates into 384-well polystyrene microtiter trays containing growth media. After an overnight incubation period, the trays (“archive plates”) are read with a spectrophotometer to measure the amount of growth in each well, as a determinant of whether the archive plate can pass on to prepping. While these instructions assume the use of robots due to the size of the task, the picking of the colonies can be performed manually if only several hundreds or thousands of ESTs are needed (Note 2). Ideally a cDNA library should contain representatives of every sequence of cDNA molecules from every different mRNA. However, each non-normalized cDNA library contains many clones of high abundance and while
beyond the scope of this chapter, one should be aware that the probability of picking up redundant clones can be reduced by different methods of selection.

3.2.1. Fill Archive Plates with Media

1. Using the WellMate, dispense 70 l of media (TB+Kan and glycerol) into each well of the 384-well trays. Be sure to replace the lid as soon as the tray is removed from the WellMate, to maintain sterility and minimize evaporation.

2. Store the filled trays in a 4°C refrigerator for up to three days until needed.

3.2.2. Prepare the Genetix-Qpix II for Picking

1. To launch a window, double-click the “Picking” icon to open the software.

3.2.3. Attaching the Picking Head and Filling the Wash Baths

1. Under “Picking”, click Diagnostics to display the plate deck.

2. On the display, left-click on a position next to the door, then right-click for an option menu. Choose “move actuator to…” to move the actuator arm to that position on the deck.

3. Being careful not to bump the camera, attach a sterile picking head to the actuator arm and fasten with a screw.

4. Close the door, then right-click on the display, and select “Home Drives” from the options menu to return the actuator arm to the Home position on the deck.

5. Check that the Air indicator light on the front panel is lit to assure adequate air pressure in the instrument.

6. Under “Picking”, select Pin Firing Test. Examine the picking head to make sure that each of the 96 pins fire.

7. Fill wash bath #1 to the top of the bristles with 1% bleach.

8. Fill wash bath #2 to the top of the bristles with sterile dH2O.

9. Fill wash bath #3 to the top of the bristles with 70% ethanol.

3.2.4. Aligning the Camera

1. Under “Picking”, check to make sure that the correct settings are selected as below:

2. Plate: Costar Plate 384 Well

3. Color: White

4. Max Plates To Use: Max plates should be set according to the target shown at the “count colony” step on the touch screen.

5. Retrieve an empty agar plate and place onto the light table, adjusting the plate corner guards against the pegs. Under “Picking”, select Align Camera. The picking head will punch a single hole into the surface of the agar. When the camera image is displayed on the computer screen, align the camera as follows: align the crosshairs on the center of the hole using the x-axis and y-axis arrow buttons. (To
improve accuracy, reduce the increment settings for the x-axis and y-axis arrow buttons and use the **Zoom** feature to enlarge the image on the screen.)

6. With the crosshairs centered, click **OK** to set the position. At the next prompt, click **YES** to return the actuator arm to the Home position on the deck.

7. Remove the agar plate from the instrument deck.

---

### 3.2.5. Adjusting the Light Threshold and Calibrating the Camera

1. Retrieve an agar plate with colonies for picking (make sure the plate lid is wiped dry) and place the agar plate onto the light table and adjust the plate corner guard so the plate is held tightly in position.

2. Under “Picking”, select **Test Image**. In the window, click on the center of the plate to highlight it, then click **Picture** to display the image.

3. On the tool bar, select **Tools** and then **Criteria** tab to check for these settings:

   4. Minimum Diameter: 7 (only colonies with 7-mm diameter or larger will be picked).

   5. Maximum Diameter: 50 (only colonies with 50-mm diameter or smaller will be picked).

   6. Roundness: 0.75.

   7. Axis Ratio: 0.75.

8. Click on the **Threshold** tab, select **Show Threshold** to display a black and white threshold image. Check for these threshold settings:

   9. Center Frame Minimum: 0

   10. Center Frame Maximum: 155

11. The agar should be all black except that the colonies should appear as white on the threshold image. Click **OK** to return to **Test Image** menu.

12. On the tool bar, select **Calibrate**. After scanning for colonies, the software computes the “Aspects Ratio”. This ratio should return at a value of 0.94 or higher.

13. After successfully calibrating the instrument, click **Done** to return to the **Picking** screen.

---

### 3.2.6. Preparing a Full Picking Script

1. Under **Picking**, select **Full**. Answer the series of prompts as below:

2. “Calibrate Lens?” click **NO**.

3. “Align Camera?” click **NO**.

4. “Prepare picking script?” click **YES**.
5. After answering these questions, the QPix will begin to scan the entire plate for colonies.
6. Ignore options such as “View Generated script code”, “Change head before continuing” etc.
7. Under the “Load Plate Holder” message, click **OK** to display the plate assignment. A minimum of one 384-well archive tray is required per agar plate.
8. Use the archive trays that have been kept on 4C in the refrigerator (*see Section 3.2*). Remove the lids and place onto the plate deck according to the diagram. Align the tray so that the A1 well is in the lower-right corner of the holder.
9. Close the QPix door and click **Done** on the interface.
10. The message “Picking is now setup and ready to start. Are you ready to continue?” will display. Click **YES** to begin picking.
11. When the instrument is finished picking, the interface will display a “Script has successfully completed” message. Click **OK**.
12. In response to the “Carry over pin and well settings?” prompt, choose **NO**.
13. Remove each archive tray, cover the wells with clear adhesive seals, and replace the lid.
14. To pick the next plate, place a new agar plate on the light bed and repeat the above steps.
15. When all plates are finished, break down the instrument by removing the picking head, sterilizing the plate deck with dH2O and UV Light, and power down the QPix.

**3.2.7. Incubating the Picked Archive Trays**

1. Place batches of picked archive trays into an incubator set at 37°C and incubate the trays for 18–24 hours.
2. After the incubation period, remove the trays from the incubator, carefully remove the clear tape, and then replace the lids.
3. Read each tray using a spectrophotometer to determine growth in 384-well plates. Prior to the sample read, perform a verification test on the Bio-Tek Microplate Scanning Spectrophotometer to ensure that it is reading samples uniformly and accurately (*Note 3*).

**3.3. Prepping**

Plasmid preps are used to extract plasmid DNA from bacterial cell suspensions and ours is based on the use of magnetic beads that help to ensnare the plasmid DNA molecules following cellular lysis by detergent (5), and to selectively precipitate them out of solution by application to a magnetic field. A washing of the magnetic beads with an ethanol solution keeps the beads and DNA in place while removing contaminants that might interfere with the
sequencing reactions to follow. Air drying of the bead rings to remove excess ethanol is followed by elution of the DNA from the beads by the addition of water.

Our instructions assume the use of Biomek FX (Beckman Coulter) robot and Cytomat carousel stackers for archive prep-ping. The Beckman Coulter Biomek FX is an automated pipetting instrument with dual multi-channel pipette heads and is therefore a suitable tool for automating routine simple prep procedures. By this type of instrument, large-scale sequencing centers can implement automated 384-well plasmid preps and increase the amount of generated sequencing data.

For a smaller number of samples one can use a mini- or mid-iprep kit (Note 4).

3.3.1. Setting Up the Biomek FX Instrument

1. Calibrate the pipetting head before using the robot: open the Biomek software, click on the Instrument heading at the top of the interface, select Manual Control, and select Home all Axes. Wait until the robot has completed the calibration before proceeding. If the pipette head jumps and an error message displays on the computer screen, click OK and click on Home all Axes once again.

2. On the Windows desktop, double-click on the Biomek FX icon, which launches the FX software. Once the software is open, click on Project and select Open Project.

3. Under File, select Open, then List, and then double-click on the Eppendorf Arch Prep icon. After opening the script, click on Instrument Setup to display the plate deck diagram so one can arrange 384 Costar archive trays and Eppendorf plates onto the deck according to the diagram.

4. Place a box of P30 tips (without lid) onto the deck according to the Instrument Setup display. Retrieve 3 tip box lids and place them to hold the required liquids (Homogeneous Solution, 85% EtOH, and Lysol) according to the Instrument Setup diagram. Fill trays with the appropriate liquids. The gravity feed tray will be used for waste. Ensure the gravity waste pump is set on “automatic” before proceeding.

5. Place black archive prep magnets into appropriate positions according to the Instrument Setup diagram.

6. Prime the wash station and fill it with water. Click on the Green Arrow to begin the Eppendorf Arch Prep. Verify proper deck layout and click OK.

7. Remove trays from the deck. Allow Eppendorf trays to dry on the benchtop for at least 15 min. Trays may be sequenced immediately or kept at 4°C overnight to be sequenced the next day. Return archive trays to a freezer for storage.
3.4. Sequencing

These instructions assume the use of BECKMAN COULTER BIOMEK FX, the 1/48X BIG DYE TERMINATOR cocktail, Eppendorf 384-WELL polycarbonate thermal cycling plates (for forward reactions, we use Archive plates and for reverse reactions, we use Red 384-well plates), and the use of an ABI 3730xl. The archive prep is sequenced directly without prior pellet re-suspen-
sion. The purpose of this procedure is to describe the assembly of 384-well Big Dye Terminator sequencing reactions with a total sequencing volume of 6uL, using a 1/48X sequencing cocktail made with Big Dye Version 3.1. This protocol uses 3ul of cocktail and 3ul of cDNA subclone for each well.

3.4.1. Setting Up the Biomek FX Instrument: 384-Well Head

1. Fill the water carboy with dH2O. When replacing the water line, make sure that the hose is adequately submerged in the water carboy. On the Windows desktop, double-click on the Biomek FX icon, which launches the FX software. Under File, go to Open. Under Methods, double-click on Seq_fulldeck_384s_brewaddition.bmt.

2. After opening the script, click on Instrument Setup to display the plate deck diagram.

3. Place a box of 384-well P30 tips (without lid) into the tip nest according to the Instrument Setup display. Retrieve archive prepped Eppendorf trays. Place one Big Dye reservoir onto the deck according to the diagram and pour the appropriate Big Dye cocktail into the reservoir.

4. Using the touchscreen, under “Sequence” purpose, “sequence” process, “unique” output devise, scan the brew (sequencing cocktail) barcode to designate Forward or Reverse (Note 5), scan each cycleplate. Press “Confirm” (Note 6).

3.4.2. Starting the Script

1. Check the level of all the solutions. On the tool bar, click the green arrow button. Throughout the run, examine the instrument and check that the tips are aspirating/dispensing evenly; that the wash station is functioning properly and fluid levels maintained, and that there are no signs of general mechanical failure. Open the light curtain protecting the plate deck if you experience any malfunction.

2. When finished with the program, seal plates with silicone plate mats. Centrifuge the cycleplates on 1500 rpm for 30 seconds. Using the MJ Tetrad or other thermal cycler, thermal-cycle the plates using:

- 95°C for 30 seconds
- 50°C for 15 seconds
- 60°C for 2 min
- 34 cycles (Note 7).
3. Make sure that the cycleplates are sitting properly in the Tetrad’s 384-well alpha blocks. This cycling time is 2 hours and 20 min.

3.4.3. Precipitation and Resuspension [Using the Multidrop (Thermo Electron) or Qfill2 (Genetix) to Dispense Ethanol Premix and 70% Ethanol]

1. Connect sterile dispense head to the Multidrop or Qfill2 and set to dispense 10 μl per well. Pour precipitation “Premix” into a sterile reservoir and prime the tubing. “Quick spin” cycleplates before removing silicone plate mats. Using the Multidrop or Qfill2, dispense 10 μl of ethanol premix into each well. Seal cycleplates with clear plastic tape and centrifuge on 3500 rpm for 30 min. Retrieve cycleplates from centrifuge, remove tape, and place upside down on a clean and sterile paper towel. Place plates back into the centrifuge (still upside down) and spin at 500 rpm for 30 seconds, to remove the premix.

2. With the Multidrop/Qfill2 setting remaining at 10 μL, dispense 70% ethanol into each well of the cycleplates. Seal with a clear plastic tape and centrifuge at 3500 rpm for 15 min. Retrieve cycleplates, remove tape, and place upside down on a clean and sterile paper towel. Place plates back into the centrifuge (still upside down) and quick spin at 500 rpm for 30 seconds, to remove the ethanol.

3. Dry cycleplates in a speed vacuum for 15–30 min or until dry. Connect a sterile dispense head to the Multidrop/Qfill2 and set to dispense 15 μl. Fill the water jar with sterile autoclaved dH2O and prime the tubing. Then, dispense water into each well of the cycleplates.

4. Seal the cycleplates using the Velocity plate sealer and store at 4°C before loading. (Before loading on ABI 3730xl, vortex the plates for a minimum of 3 min followed by a “quick spin”.)

3.5. Loading the ABI 3730xl Analyzer

Loading trays on the Applied Biosystems 3730xl Analyzer – These instructions assume the use of an Applied Biosystems 3730xl analyzer, explaining how to load trays on this instrument. Different programs can be used when loading complete or partial trays, or subset of quadrants of a 384-well plate. We will only describe the program used for full 384-well plates that are being run with the machine’s default run module (Note 8).

General information for Batman program – Batman allows for Just In TimeTM loading of trays by maintaining continuous communication between blade servers and the 3730 Data Collection software. As a plate is taken from the in-stacker of the 3730 DNA Analyzer and scanned internally, a sample sheet is automatically generated. With the use of this program, the user does not have to manually scan or process trays/sample sheets. As each run is completed, Batman will schedule and carry out the transfer and
trace analysis of the data generated. This distribution of transfer and analysis throughout the day, as opposed to once daily, greatly decreases peak database and processor loads in addition to decreasing the number of “race condition” related trace analysis fails.

3.5.1. Loading Full Trays with Default Run Module

1. On the PC desktop, restart ABI Data Collection.
2. If the computer was restarted, ABI data collection must also be restarted.
3. Ensure that all instances of non-Batman software (e.g., Batmobile) are shut down.
4. Open the Batman software by double-clicking on the icon.
5. Ensure that the run module for the samples on the ABI is set for Batman loading.
6. Open and close the loading door to see that Batman records the open and close events.
7. Place trays to be loaded in the stacker and close the door.
8. Go to the Run Scheduler to ensure the instrument mode is set to auto and when the machine is ready, click the green arrow to start the load.

3.6. Sequence Extraction, Processing, and Submission of ABI 3730xl ESTs

More details on sequence extraction and processing, prior to submission to the dbEST division of GenBank, are provided in Chapter 9.

3.6.1. Basecalling and Detection of High-Quality Region

Processed data from the ABI 3730xl exists in the form of AB1 format chromatographs, or trace files. The ABI 3730xl software itself assigns basecalls to the sequence data obtained from each capillary for each run, and assigns a quality value to each base. Transfer of data from the ABI computer to a primary filesystem can be accomplished using a common FTP program such as WinFTP, Fetch, or Fugu.

One might choose to re-basecall sequence data rather than use the ABI machine’s default basecalling. This allows reformatting the trace files to other chromatograph format, if desired. Two commonly used basecallers are ABI’s own KB basecaller and the Phred basecalling program (6, 7). These algorithms provide both the basecalls themselves and an associated quality value (quality score) for each base. For example, Phred quality scores range from about 4 to 60, with the higher number corresponding to a more confident basecall.

Run Phred and generate basecalls and quality values (Note 9). The Phred program, in addition to basecalling and the assignment of a Phred quality value for each base, will define a region of overall high quality in each trace (8, 9) (Note 10). Phred can be used to create SCF format trace files from the original AB1 format files (Note 11).
The STADEN package’s “Experiment file format” is a convenient flat file format for storing sequencing data and all the ancillary information that accumulates as one performs downstream analysis. It allows storage of the entire length of sequence basecalls and quality values, while also storing positional information that defines the beginning and end of the “high quality region”, and where vector-insert boundaries exist, and so forth. For details on the experiment file format visit the STADEN package homepage at: http://staden.sourceforge.net/manual/formats_unix_18.html.

3.6.2. Trim Vector and Linker, Screen for Contaminants and “Low Complexity” Regions, and Derive Putative IDs

A summary for these steps is provided here; however, for more details on sequence trimming, screening, and annotation see Chapter 9. The high-quality region of the trace (as determined by Phred) is used when searching for the vector-insert boundaries. Use an already available vector/adaptor trimming program, such as “vector_clip” from the STADEN package (see http://staden.sourceforge.net/manual/vector_clip_unix_toc.html) to identify these boundaries. We take the vector-free, high-quality insert and screen for possible contaminants by performing a simple BLAST vs. databases of expected contaminants, such as: structural RNA (noncoding RNA subunits, e.g., 18S); bacterial contamination; vector sequence (sometimes sequencing vector can be cloned as an insert); and other contaminants (such as databases of other species being worked with in the lab). We discard all reads that align to sequences from these contaminant databases (Note 12).

Furthermore, any region with an unusual composition can create problems in sequence similarity searching (10) (causing false positive hits) and should be removed. Use a software tool such as “dust” for this purpose (dust <fasta file>). The output will be a fasta sequence masked for what “dust” considers to be low-complexity sequence (all regions of low complexity will have all their bases converted to “N”).

Finally, an automated annotation step in which ESTs are BLASTed against a protein database (such as GenBank’s nr db: all nonredundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF) provides a putative id for each EST (blastx vs. GenBank’s “nr” database).

3.6.3. Format EST File for Submission to NCBI’s GenBank dbEST Division

1. For each batch submitted, an NCBI “EST file” must be created. For a detailed description of an NCBI EST file, see the URL (NCBI website): http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html#EST%20Files.

2. Build and submit ancillary information files to the NCBI before the initial submission. In addition to the EST file that can be prepared and submitted as data come off each sequencing machine, a number of ancillary files must be prepared and
submitted to the NCBI in advance. These ancillary files need only to be submitted once, they do not need to be resubmitted for additional ESTs sequenced from the same source. Each EST file needs to reference a specific Publication, Library, and Contact file.

3.6.3.1. Publication File

This file details the project that is sponsoring the ESTs being submitted. Typically these projects lead to a publication (thus the official name “Publication File”). A single publication file often will be used with numerous different Library files (described below), as a single project/publication can involve many libraries. For a detailed view of a Publication file see: http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html#Publication%20Files.

3.6.3.2. Library File

Each different library must be detailed individually in the Library file (although you are allowed to include multiple library records in a single Library file you submit to the NCBI). Library files describe the numerous traits associated with the various libraries from which you have generated ESTs. Often a single project will include many library file entries. For a detailed view of a Library file see: http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html#Library%20Files.

3.6.3.3. Contact File

Each Publication file must be accompanied by a Contact file. The Contact file includes information about the submitter and the submitter’s lab. For a detailed view of a Contact file see: http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html#Contact%20Files.

3.6.4. Submit the ESTs to dbEST Repository and the Traces to the NCBI Trace Archive

1. Concatenate all files using a simple UNIX command “cat” into one file: cat *dbest > my_dbest_submission.dbest.

   Compress the file before submitting it using any appropriate compression tool. One example is the UNIX gzip compression utility: gzip my_dbest_submission.dbest.

   Send the file (my_dbest_submission.dbest.gz) to the dbEST repository (batch_sub@ncbi.nlm.nih.gov) as an attachment if you have a small number of ESTs (e.g., <2,000). Note if you are sending dbEST sequence files as attachments they should not be compressed, but rather attached as plain text. If you have a larger number of sequences (e.g., >2,000), then we would recommend compressing all files including the dbEST sequence files and using the NCBI’s FTP site to upload the file for larger submissions: ftp ftp.ncbi.nih.gov.

   In this case, communicate with the NCBI and obtain a username/password for submission before uploading data. A good email for first contact communication would be: info@ncbi.nlm.nih.gov.
Once a username and password for access to their FTP site has been obtained, each time data are uploaded, also send a notice to batch-sub@ncbi.nlm.nih.gov describing the submission. For more details on proper submission protocol for the NCBI, please visit their website at: http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html.

2. Deposit trace files in the trace archive. Trace data and supporting files should be placed on the NCBI secure FTP site. The first thing to do is to contact trace@ncbi.nlm.nih.gov to obtain a secure FTP account, including identifying information for your group such as a group name and trace archive acronym.

Note that submissions to the NCBI trace archive are automatically picked up by Ensembl as well (and vice versa). Each submission consists of a single, compressed UNIX tarfile. When extracted that tarfile should expand into a directory containing the following items (do NOT put traces in the top level of this directory tree):

a. TRACEINFO.xml or TRACEINFO.txt – this is the main file describing the submission.

b. MD5 – this is an MD5 footprint of this directory, for verification purposes.

c. README – free text file describing the submission and contents of this directory.

d. subdirectory structure for the traces (usually a directory named “trace”).

For full details on preparing these files, please visit the website: http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=submit&m=doc&s=submit.

The README file is free form text and is often as simple as this: This submission contains 15,225 SCF format trace files. This volume was created on 8/25/2006.

And finally, the traces themselves need to be placed under a subdirectory of the outermost directory (do not put traces into the outermost directory). Note that when referring to traces in the TRACEINFO.xml (or TRACEINFO.txt) files, you need to use the relative path to the trace, such as:

traces/HBBA/HBBAA1U0001.scftraces/HBBA/HBBAA1U0002.scfetc.

For full details, please visit the NCBI trace archive website at: http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=submit&m=doc&s=submit.

3.7. Full-Length cDNA Preparation for 454 Life Sciences GS20 (Note 13)

This protocol describes a novel, noncloning-based shotgun approach to full-length cDNA (fl-cDNA) sequencing that is performed using a novel sequencing technology platform. The approach was developed to improve throughput and cost
considerations associated with fl-cDNA sequencing projects and
to deliver full-length cDNA sequences on a genomic scale, such
that the needs addressed by cDNA projects might become more
economically and temporally tractable. In particular, pyrosequen-
cing by synthesis using the 454 instrument platform provides
high-throughput sequence data acquisition, resulting in hundreds
of thousands of ~100mer sequences in only a few hours. Furthermore, our unique strategy addresses two potential limitations of
cDNA sequencing on the 454 platform that impact read quality
and efficiency, by minimizing both long homopolymer stretches
[i.e., 3’ poly (A) tails] and 5’ adapter sequences.

**3.7.1. Modified SMART First-Strand cDNA Synthesis**

This step is aimed at producing full-length cDNA copies by oligo
dT priming and optimized RT conditions.

1. Prepare each of the following mixtures in a separate 0.2-ml
PCR tube. If there is sufficient RNA, one could choose to
prepare a double reaction (40 µl total volume).
Mix #1 (for a total volume of 5.0 µl):
   1.0 µl  12 µM 5’ Smart_Oligo
   1.0 µl  12 µM 3’ Oligo_dT_SmartIIA
   1.0 µl  Clontech 10 mM dNTP Mix.
Add appropriate µl of total or poly A+ RNA (0.025–1 µg poly A+
or 0.05–1 µg total RNA) and water to a total volume of 5.0 µl.
Mix #2 (for a total volume of 15.0 µl):
   4.0 µl  Clontech 5X First-Strand Buffer [250 mM Tris-HCl
           (pH 8.3), 375 mM KCl, 30 mM MgCl2]
   0.4 µl  Clontech 100 mM DTT
   0.5 µl  NEB 10 mg/ml BSA
   0.5 µl  Invitrogen 40U/µl RNaseOUT
   7.6 µl  80% Trehalose or Water
   2.0 µl  Clontech PowerScript RT.
2. Incubate mix #1 for 5 min at 65°C and hold at 45°C.
3. Incubate mix #1 and mix #2 for 15 seconds at 45°C. Transfer
   mix #2 to mix #1 without removing the tubes from the
   thermocycler. Pipet up and down to mix.
4. Incubate the tubes using the following conditions:
   1.4.1:  45°C for 2 min (hot start)
   1.4.2:  Negative ramp: go to 35°C in 1 min (gradient
           annealing)
   1.4.3:  35°C for 2 min (complete annealing)
   1.4.4:  45°C for 5 min
   1.4.5:  Positive ramp: +15°C (until 60°C) at +0.1°C/sec
1.4.6: 55°C for 2 min
1.4.7: 60°C for 2 min
1.4.8: Go to Step 6 for 10 additional times
1.4.9: +4°C hold.

3.7.2. PCR Cycling Optimization (Titration)

These steps are used to determine the threshold cycle number for minimally amplifying the full-length cDNA products. We have found this titration to be critical, due to differences in cDNA libraries among different organisms, differing RNA qualities, etc.

The optimization is conducted within a single 100-μl PCR reaction:

- 10.0 μl Clontech 10X Advantage 2 PCR Buffer
- 2.0 μl 10 mM dNTP Mix
- 2.0 μl New_SMARTIIA (for primary PCR reactions)
- 2.0 μl Clontech 50X Advantage 2 Polymerase Mix
- 2.0 μl ss cDNA
- 82. μl Water.

1. Perform 18 cycles and remove 10 μl for gel electrophoresis
2. Perform 2 additional cycles and remove 10 μl (20 total cycles)
3. Perform 2 additional cycles and remove 10 μl (22 total cycles)
4. Perform 2 additional cycles and remove 10 μl (24 total cycles)
5. Perform 2 additional cycles and remove 10 μl (26 total cycles)
6. Run all of the reactions on a gel to determine the optimum number of thermal cycles, defined as the point at which exponential amplification begins. (Note 14; Fig. 8.1).

Fig. 8.1. Determine the optimum number of thermal cycles. The optimal number of cycles will have a bright smear >300 bp and <4–6 kb. The smear should not reach the bottom of the well on the agarose gel. In this example, the optimal number of cycles is 21.
7. Preheat the thermal cycler to 95°C, use the following cycling program: 95°C for 1 min × cycles (as determined in 6); 95°C, 15 sec; 65°C, 30 sec; 68°C, 6 min. Hold at 4°C.

3.7.3. PCR Template Production for Sequencing

Prepare 8–10 100-µl PCR reactions (yield: ~1–3 µg per reaction) using the same amount of first-strand cDNA from the cycling optimization step and cycle for the determined optimum number of cycles to ensure that you will have 15 µg for downstream steps. After the PCR reaction is complete, purify the PCR product with Qiagen QIAquick columns. Each 100-µl reaction should produce 1–2 µg of DNA so take care to not overload the QIAquick columns (binding maximum for these columns is 10 µg). Follow the QIAquick protocol with a final 30-µl elution in water. Assay the CDNA concentration (e.g., via Nanodrop). Save ~100 ng cDNA for future gels.

3.7.4. Restriction Exonuclease Digestion for 5’ and 3’ Adaptor Excision

The presence of the Mme1 recognition sites in the modified adapters allows the removal of the adapters and much of the polyA tail, prior to 454 library preparation.

1. Prepare the following mixture:
   10 µl NEB 10x Buffer 4
   1 µl NEB 10 mg/ml BSA
   5 µl NEB 2 U/µl MmeI
   2 µl 3.2 mM S-adenosylmethionine (SAM) (1:10 dilution of NEB 32 µM stock)
   5 µg filtered RT-PCR product

   Add water (77 µl ) to a final volume of 100 µl.

2. Incubate the mixture 30 min at 37°C.

3. Incubate 20 min at 80°C.

4. Store at –20°C or continue with Streptavidin-bead purification.

5. Perform 2 digest reactions that contain 5 µg PCR product per reaction (Note 15).

6. Purify the digested DNA using 1 QIAquick column for every 10 µg of input DNA. Follow the QIAquick protocol with a final 30 µl elution in water.

7. Assay the DNA concentration (e.g., via Nanodrop). Save ~100 ng DNA for future gels.

3.7.5. 5’ and 3’ Adaptor Removal Using Streptavidin Paramagnetic Beads

1. Bead Washing
   - Transfer 100 µl (1 mg) Dynabeads M-280 Streptavidin beads to a clean tube.
   - Place the tube on the MPC for 2 min.
   - Remove the supernatant with a pipette.
– Remove the tube from the MPC and resuspend the beads with 100 µl 2x B&W Buffer.
– Repeat Steps 2, 3, and 4 for a total of two washes.

2. Nucleic Acid Immobilization
– Add equal volumes of 2x B&W buffer and purified MmeI-digested PCR product to the washed streptavidin beads.
– Incubate for 15 min at room temperature with inversion (Labquake).
– Place the tube on the MPC for 2 min.
– Transfer the supernatant to a clean tube.
– Remove the tube from the MPC and resuspend the beads with 100 µl 1x B&W buffer.
– Repeat Steps 3 through 5 for a total of 2 washes.

3. Purify the digested DNA using 1 column for every 5 µg of input cDNA. Follow the MinElute protocol (Qiagen) with 2–13 µl elutions in water.

4. Assay the DNA concentration (e.g., via Nanodrop). Save ~100 ng DNA for future gels.

5. Use the nebulized cDNA for sequencing (Fig. 8.2).

3.8. 454 Life Sciences GS20 Sequencing Run

These instructions assume the use of 454 Life Sciences Sequencer GS20. The procedure starts with a sequencing run that should be carried out after the prewash run, followed by PicoTiterPlate preparation (bead deposition), and the prep run. This section will explain the sequencing run; for carrying out the previous steps consult the GS20 manufacturer (http://www.454.com/products-and-reagents/reagent-kits.asp).

Fig. 8.2. Modified SMART Flashgel (~50 ng DNA per lane). Ln 1. Modified SMART RT-PCR product (23 cycles); Ln 2. MmeI-digested Modified SMART RT-PCR product (the <100 bp band is assumed to be the biotinylated 5’ and 3’ ends of the digested RT-PCR product); Ln 3. Dynabeads M-280-purified MmeI-digested DNA (the <100 bp band was eliminated by the paramagnetic beads); Ln 4. Nebulized DNA.
3.8.1. Clean (Replace) the PicoTiterPlate™ Cartridge and the Camera Faceplate

1. After the prep run has completed, open the camera cover. Fluid will be pumped to waste. If the PicoTiter™ Plate cartridge is not completely emptied of liquid after opening the camera cover, close the camera cover, wait for 5 seconds, and reopen it to remove any remaining liquid. (If the cartridge will be replaced, install the Camera Faceplate Guard on the camera face by hanging from the metal faceplate seating pins.)

2. Remove the PicoTiterPlate™ from the PicoTiterPlate™ cartridge by pressing its frame spring latch to lift the frame from the cartridge, and then sliding out the used plate.

3. Remove the PicoTiterPlate™ cartridge seal from the PicoTiterPlate™ cartridge.

4. Squirt 50% ethanol on the cartridge and wipe it with a Kim-wipe. Allow the cartridge to air dry fully.

5. If required, change the cartridge to the appropriate size (Note 16).

6. Gently wipe the camera faceplate with a new Zeiss moistened cleaning tissue and allow the camera faceplate to air dry completely.

3.8.2. Load and Set the Run Script and Other Run Parameters

The 454 instrument has some flexibility in terms of how the picotiter plate is sectioned, loaded, run, and analyzed. As such, it is important to follow the instructions listed below for a successful run of the instrument.

1. A small Run complete window will open when the prep run is completed. Click OK to continue. At the top of the Instrument Run window is a single menu: File. Select Run Script, and the Settings window will open.

2. To set up a new sequencing Run, in the Scripts field at the top of the Settings window, select the sequencing Run script for this Run from:

   70x75\TACG\42x_TACG_70x75
   10.2.4..1.2.40x75\TACG\42x_TACG_40x75

   – In the Run Name field, type in a specific name for this Run and enter appropriate barcode, size, and layout of the PicoTiterPlate™.

   – Under Data Analysis File, select the Data Analysis Configuration file to be used for this Run.

   – Select the appropriate user and user group and under Backup Class, select Backup.
3. Click on the **Next** button at the bottom of the Settings window to proceed.

4. The **Run Name Confirmation** window will open, click **Yes** for a name that is specified correctly (as designated by the user), or **NO** if the name is not correct (in this case the window will close and return to the Settings window).

5. The **Requirements** window will appear. If there is NOT sufficient disk space available, make it available by transferring or deleting files from the disk.

3.8.3. Final PicoTiterPlate™ Rinse

1. After centrifugation of the bead layers is complete, remove the BDD from the centrifuge, and remove and discard the Port Seals (or MicroSeal A strips).

2. Gently draw out and discard the supernatant from the centrifuged bead layer.

3. Remove the PicoTiterPlate™ from the Bead Deposition Device (BDD), as follows:
   a. Rotate down the latches of the BDD.
   b. Carefully remove the BDD top.
   c. Gently lift off the gasket.

4. Remove the PicoTiterPlate™, being careful to handle it only by the edges.

5. Submerge the PicoTiterPlate™ into the shipping tray filled with Bead Buffer 2.

3.8.4. Insert the PicoTiterPlate™ and Launch the Sequencing Run

1. Move to the camera cover. If the PicoTiterPlate™ frame is open, close it (NOT the camera cover).

2. Install the cartridge seal: verify that the square ridge on the seal is facing up, and drop the seal in the cartridge groove. If necessary, gently tap the seal into place with a gloved hand. DO NOT wipe the seal with anything.

3. Press the PicoTiterPlate™ cartridge spring latch to lift the PicoTiterPlate™ frame from the cartridge. Remove the fully prepared PicoTiterPlate™ from the tray, being careful to only handle it by the edges.

4. Slide the PicoTiterPlate™ into the frame, making sure that the notch is on the lower right-hand corner and close the frame making sure it is caught by the latch.

5. Wipe the backside of the PicoTiterPlate™ with a Kimwipe and close the camera cover.

6. Click on the **Finish** button at the bottom of the Requirements window to exit it and start the Run.
After a 454-run completes, data are deposited into a run folder. This run folder will contain a subdirectory named “SFF”, which itself will contain one or more SFF (Standard Flowgram Format) files. One SFF file will initially be generated per 454 region, and they can be freely concatenated together to represent an entire run in one SFF file if desired. These SFF files are what will be used to extract high-quality basecalls from, using the high-quality determination criteria made by the 454 software. Note that prior to version 1.0.52 of the 454 software, the SFF files were not automatically created during the run, but instead were created from other data files deposited in the analysis directory using the script “sffcall” after the run completed.

SFF is a file format designed in collaboration by 454 Life Sciences, the Whitehead Institute for Biomedical Research, and the Sanger Institute that is meant to hold flowgrams (the “traces” produced by the 454-sequencing machine) in a manner similar to the traditional SCF trace format. An SFF file can hold 1 or more sequences (typically it will consist of all the reads of a single 454 run, which may number in the hundreds of thousands). Basecalls and quality values (which use a phred-like scoring system) can be extracted on a per-read basis from an SFF file analogous to how similar data would be extracted from a standard SCF trace file. The data themselves differ from a standard SCF trace file in that the 454 data do not provide individual base measurements from which basecalls can be derived. Instead they provide measurements that estimate the length of the next homopolymer stretch in the sequence (i.e., “AAATGG”, “AAA” is a 3-mer of A’s, “T” is a 1-mer of T’s, and “GG” is a 2-mer stretch of G’s). A basecalled sequence is then derived by converting each estimate into a homopolymer stretch of that length, and concatenating the homopolymers. As of version 1.0.52 of the software, quality values are only “phred-like”, and should not be taken as true phred quality scores. In our experience, these values are based on the confidence of the length estimate for each homopolymer. Bases further toward the end of long homopolymers typically have lower quality values than the bases near the beginning of the same stretch.

Once the SFF file is transferred to a working directory, extract high-quality basecalls (as determined by the 454 basecalling software) for each sequence using the program “sffinfo”. Sffinfo comes with the 454 software package (as of v1.0.52, which is the version our center is currently using) and is easy to use. To extract the basecalls for the high-quality region of all traces described by a given SFF file:

```
sffinfo -s <SFF file filename> > <fasta output file>
```

This will deposit fasta nucleotide sequence into <fasta output file> representing what the 454 software believes to be the high-quality region of each individual read. Options understood by the “sffinfo” script are:

1. **-s**
   - Uses standard output
2. **-f**
   - Uses standard input
-a or -accno  Output just the accessions
-s or -seq  Output just the sequences
-q or -qual  Output just the quality scores
-f or -flow  Output just the flowgrams
-t or -tab  Output the seq/qual/flow as tab-delimited lines
-n or -notrim  Output the untrimmed sequence or quality scores
-m or -mft  Output the manifest text

Note that as long as the -notrim (or -n) switch is not used, the output sequence will already be trimmed down to what the software considers the high-quality region. For the purposes of EST submissions only the quality-trimmed, nucleotide fasta data will be needed. So most likely only using the -s switch will be necessary.

3.9.2. Trim Linkers

This step can be done in the same manner as that described for 3730 data. See Section 3.2 and Chapter 9.

In place of the blastn line suggested earlier for the 3730 process, we would suggest this alternative (Note 17):

```
blastn <vector db> <cleaned sequence> S=80 M=5 N=-8 R=11 Q=11
```

As with any blast parameter, it is important that the user experiment and make sure they are able to capture the alignments they are hunting for. The parameters above are merely a starting point. Create a sequence that you know should fail (i.e., be found by this blast) and tweak the parameters until you can reliably find that sequence, while passing other sequence that is free of vector.

3.9.3. Screen for Contaminants

This step can be done in the same manner as that described for 3730 data. See Section 3.2 and Chapter 9.

In place of the 2 blastn lines we suggested for the 3730 process, we suggest these as replacements for short-454 reads (Note 18):

For a less stringent blast:

```
blastn <screening db> <cleaned sequence> S=80 S2=80 M=5 N=-11 Q=11 R=11 W=10 E2=0.5
```

For a more stringent blast:

```
blastn <screening db> <cleaned sequence> S=120 gapS2=100 M=5 N=-11 Q=11 R=11 W=10 E2=0.5
```

As with any blast parameter, it is important that the user experiment and make sure they are able to capture the alignments they seek. The parameters above are merely a starting point. Create a sequence that you know should fail (i.e., be found by these blasts) and tweak the parameters until you can reliably find that sequence, while passing other sequence that is free of contaminants.
3.9.4. Screen for “Low Complexity” Sequence

This step can be done in the same manner as that described for 3730 data. See Section 3.2 and Chapter 9.

3.9.5. Derive Putative IDs

This step can be done in the same manner as that described for 3730 data. See Section 3.2 and Chapter 9.

Note: The blastx described in the 3730 process should work fine with the shorter reads generated by the 454 instrument. None of the suggested parameters are as heavily dependant on the alignment length as was so in the contamination and vector screens. The user may wish to alter the minimum e-value threshold from 1e-10 to 1e-05. As always, test and tweak your parameters until they meet your needs.

3.9.6. Format EST File for Submission to NCBI’s GenBank dbEST Division

This step can be done in the same manner as that described for 3730 data. See Section 3.3.

3.9.7. Submit Flowgrams to the NCBI Trace Archive (Note 19)

The first thing to do is to contact trace@ncbi.nlm.nih.gov to obtain a secure FTP account, as described earlier. This secure FTP site is where you will eventually upload your submission tar format file. Note that submissions to the NCBI trace archive are automatically picked up by Ensembl as well (and vice versa).

You should prepare one submission tarfile per 454 region (which will correlate to one SFF file). Each submission consists of a single, compressed UNIX tarfile. When extracted, that tarfile should expand into a directory containing the following things (Note: do NOT put traces in the top level of this directory tree):

1. TRACEINFO.xml – this is the main file describing the submission;
2. SFF file – the SFF file aggregating all the flowgrams for a single 454 region;
3. README – free text file describing the submission and contents of this directory;
4. MD5 – this is an MD5 footprint of this directory, for verification purposes.

For more detail on preparing these files, please visit the website (Note 20): http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=submit&m=doc&s=submit

1. TRACEINFO.xml – this is the main file describing the submission.

Included with the 454 software package as of the current release (v1.0.52) is an application named “sffvolume”. The TRACEINFO.xml file is generated simply by running the “sffvolume” program. Note that “sffvolume” also prepares the SFF file itself for submission, creating a version in the specified output directory (see below).
Here is a typical sffvolume command line:

```
sffvolume -o <output directory/file> -center <center name annotation>
-project <center project annotation> -seqlib <seq_lid_id annotation>
-species <species annotation> -source NONGENOMIC
-strategy EST -subtype NEW <path to SFF file>.
```

- `<output directory/file>`: path to where you want the submission directory created;
- `<center name annotation>`: CENTER_NAME annotation (you need to have already arranged for a center name acronym before submission – see above);
- `<center project annotation>`: your center’s project name designation for this specific sequencing project;
- `<seq_lid_id annotation>`: center identifier for the clone that is actually sequenced in this experiment. This value will be unique within a given center;
- `<species annotation>`: species name;
- `<path to SFF file>`: path to the SFF file associated with volume.


The sffvolume command will create a submission directory, and inside build the required TRACEINFO.xml file, and the required SFF file.

2. SFF file - the SFF file aggregating all the flowgrams for a single 454 region.
This file will be formatted and copied into the submission directory by the sffvolume program (see above).

3. README - free text file describing the submission and contents of this directory.
This file is simply a short text file describing the current submission. README files should just mention the number of reads being submitted in the current submission directory and list the date the directory was created. Example:

This submission contains 215,225 sequences.
This volume was created on 8/25/2006.

4. MD5 – this is an MD5 footprint of this directory, for verification purposes.
This is a verification footprint generated by the MD5 algorithm for each file included in the submission directory.

MD5 example:
728018368a7820c50cbaad633bc608a1 TRACEINFO.xml
0cbaad633bc608a1728018368a7820c5 454Reads.sff
1c877d636cd69aa54b8c33259a12309d README.

Once the above submission directory is created, tar and compress it and then upload it to the NCBI Trace repository using the secure FTP account they will have provided (see above).

Each submission can be tracked by watching the corresponding secure FTP site. After each submission has been processed, log files documenting the upload are placed on the FTP site detailing any problems that may have occurred. Check inside this log file to determine the fate of your submission.

4. Notes

1. With this machine one could process ~20,000 ESTs in ~3 days. The majority of that time is spent doing the BLASTX putative id search. Disk space needed is dependent upon the amount of data one wants to process. An average SCF format trace takes ~30–50 Kb of space, and the additional generated files (such as the STADEN experiment format files, and dbEST format files) require a few additional Kilobytes. Therefore, an average EST project of ~20,000 ESTs (x 50 Kb each) would take up about 1 Gb of hard disk space.

2. Since picking is done manually, picking guards are utilized to mitigate picking error and prevent cross-contamination of colonies into surrounding quadrants. Quadrants 1–4 should be labeled in consecutive order from left to right. Using a sterile toothpick, scrape an individual bacterial colony from a premade agar plate and inoculate each well with a single colony. Agitate by lightly tapping the toothpick up and down in the well. Pick only colonies without surrounding satellite colonies as they may result in poor sequencing template. Be sure to pick into the appropriate quadrant(s) of the 384-well plate.

3. Verification test on the Bio-Tek Microplate Scanning Spectrophotometer – needed material: Universal Test Plate from Bio-Tek; Bio-Tek QC check solution (yellow liquid); ddH2O.
Procedure:

a. **Optics Test**
   1. Open the KC4 application.
   2. From the menu select [System -> Diagnostics -> Run Optics Test].
   3. Enter the reader serial number [spec #1 – 152293, spec #2 - 158056].
   4. Click Start.
   5. A box will appear with the message, “Processing Optics Test...”.
   6. When identification box pops up, click OK.
   7. When the results are displayed (Bio_diag – Notepad) select [File -> Print] to print out the report.
   8. Look for “SYSTEM TEST PASS” on the last page of print-out before filing the report in the SPEC MONTHLY TEST LOG BOOK. If the system fails for any reason, e-mail the repair email alias.
   9. Hit the CLOSE button to exit out of the Optics test.

b. **Universal Plate Test**
   1. From the KC4 menu, select [System -> Diagnostics -> Run Universal Plate Test].
   2. Enter reader serial number [spec #1 – 152293, spec #2 - 158056].
   3. Choose Universal Test Plate and enter your name in the Operator box.
   4. Insert the Universal Test Plate into the reader with the A1 corner of the plate in the upper left corner of spec #1 or the A1 corner of the plate in the upper right corner of spec #2 and click OK.
   5. Click Run Test.
   6. Tests will run on all predefined wavelengths, then a message will appear asking for the Universal plate to be rotated 180 degrees. After rotating, click OK and all of the tests will be run again.
   7. When the results appear, select [File -> Print] to print the report, then file the report in the SPEC MONTHLY TEST LOG BOOK. *(report will be saved to CALIB.MDB in the KC4 folder)*.

c. **Liquid Test 1 to Test Channel-to-Channel Variation**
   1. Prepare test plate
      (a) Mix 5 mL of QC check solution with 95mL of ddH₂O.
      (b) Dispense 200 uL of this solution into the first column of a 96-well tray.
(c) Make a 1:2 dilution of the QC mixture and dispense into the second column of the 96 well tray.

2. Read the test plate 5 times at 405 nm using normal mode, single wavelength, and no blanking.

3. Turn the microplate around 180 degrees so that the A1 is in the H12 position.

4. Read the test plate 5 more times.

5. Export data to an Excel spreadsheet.

(a) Calculate the mean of the wells in columns 1 and 2 in normal and turnaround position.

(b) Compare the mean for A1 and H12, B1 and G12, etc.

(c) The difference in the values for any two corresponding wells should be within the accuracy specs for the instrument.

\[ +/– 1\% \pm 0.010 \text{ Abs from 0-2.0 Abs} \]

\[ +/– 3\% \pm 0.010 \text{ Abs from 2.0-3.0 Abs.} \]

4. When a smaller number of samples need to be prepped, use one of a high-purity plasmid miniprep kits (e.g. QIAprep Spin Miniprep Kit, QIAGEN cat # 27104).

5. Our experience is that 5’ end sequencing has a higher pass rate than the 3’ end sequencing. Unless the library has been specifically constructed to eliminate a large portion of the 3’ poly(A) tails, the polymerase can slip on the long poly(A) template and give poor or erroneous reads from that end. In addition, one can choose to sequence the 5’ end as a more informative terminus since the shorter 5’ UTR sequence usually results in a larger proportion of coding sequence.

6. The described instructions are specific to the touch screen interface and will need modifying for other LIMS setups.

7. The number of cycles and the thermal-cycles temperatures might need to be optimized. These are the conditions that worked well with most of our sequencing reactions.

8. General program information:

**Batmobile** – Batmobile is meant to replace bc3730 for both Mutational Profiling and Production pipeline trays. Basically, the program serves as a load vehicle for the Batman software in instances where run module or polymer bottle changes are necessary. If a tray is not full, quadrants may be removed with the Alfred program prior to creating/importing sample sheets with Batmobile. Note that this is very specific to our LIMS, and you may consider using commercially available programs that can be easily incorporated in your own LIMS setup.
**Alfred**—This program allows for loading a subset of the quadrants of a 384-well plate and allows quadrants/runs to be reloaded and/or rescheduled. This is useful, for example, when a testing tray must be reinjected for multiple runs. Alfred is sophisticated enough to determine if the quadrants have been run or scheduled elsewhere and resolve the data conflicts. It is important to know that Alfred will not remove the plates from Data Collection or the run file from the sample sheets folder. Manual removal from Data Collection and the sample sheets folder is still required. This tool is a plate set-up modifier and is used before Batman or Batmobile in instances of partial plates or if a plate needs to be flagged as testing.

9. `-st xbap` outputs the sequence data in xbap format. One would probably prefer `-st fasta` for fasta outputs. Phred documentation is available online at: [http://bozeman.mbt.washington.edu/phrap.docs/phred.html](http://bozeman.mbt.washington.edu/phrap.docs/phred.html).

New users of Phred may prefer to produce “-st fasta” format sequence data. That will produce a less obscure looking, standard fasta format sequence output. The same information is produced, but the sequence data do not need to be stitched together to form the full, raw (untrimmed) sequence. Note that the header still contains the number of leading, low quality bases, and the number of bases in the following region considered “high quality”. The header also provides the total number of bases in the raw sequence.

10. It is valuable to save the entire basecalled sequence, including data outside the high-quality region. One can imagine instances when a gene alignment or other feature may run beyond the region of high quality and it would be helpful to watch the alignment extend into the more questionable region of low-confidence basecalls, in case it turns out to be informative. For storing this information in a format that preserves the entire raw sequence, one would need to stitch together the 3 “sections” (low-quality region trimmed from before the high-quality region PLUS the high-quality region PLUS the low-quality region trimmed from after the high-quality region). Alternatively, one can store all of the raw basecalls, and then just store the boundary positions of the high-quality region.

11. Furthermore, one may want to use Phred to create SCF format trace files from the original AB1 format files (some software packages accept only SCF traces and not the AB1 format). The basic command switch to produce an SCF format trace file using Phred is “-c”. There are a number of alternate ways to call this parameter to allow things such as
defining whether SCF version 2 or SCF version 3 format is desired, to rename the trace, etc. See http://bozeman.mbt.washington.edu/phrap.docs/phred.html for full details.

There are trace conversion programs available for downstream use as well (for example: the STADEN package’s makeSCF program.)

12. After assembling fasta files for the specific databases to be screened and formatting those databases into blastable entities using xformat, simply run whatever type of BLAST is appropriate for each database (e.g., blastn for nucleotide query vs. nucleotide subject, blastx for nucleotide query vs. amino acid subject, etc.). The query sequence will be the “cleaned” sequence of all reads (e.g., only the portion of the read that is high quality as determined by Phred, and also not including anything determined to be vector by either vector_clip or the BLAST-based vector detection step.)

A typical blastn command line to use for screening would be:

```
blastn <screening db> <cleaned sequence> S=133 S2=133 M=5 N=-11 Q=11 R=11 W=10 E2=0.5
```

One could chose simply to exclude ANY hits using the recommended BLAST parameters, or could set a threshold BLAST score, or p-value, or % identity + length of alignment as a threshold for deciding if a read is a contaminant, or is just occurring by chance. Note that some contaminant databases may contain genes that could have homologs to the species of interest. In these cases, set a stringent threshold and only fail reads that are nearly identical to sequences in the contaminant db. In such cases it is safe to set a threshold, such as requiring the alignment to be 96% identical at the nucleotide level along 90% of the length of the read.

13. **Figure 8.3** provides an outline of the process that is used to generate full-length cDNAs for sequencing on 454 GS20 sequencer.

Fig. 8.3. Generation of full-length cDNAs for sequencing on 454 GS20 sequencer. The top graphic indicates the steps aimed at producing full-length cDNA through first-strand cDNA synthesis using modified SMART. In the bottom graphic the arrows indicate the primers for the first-strand cDNA sequencing.
14. The optimal number of cycles will have a bright smear of >300 bp and < 4-6 kb. The smear should not reach the bottom of the well on the agarose gel, as this is diagnostic of overcycling. Your goal is to perform sufficient cycles to remain within the exponential phase of PCR and no more (Fig. 8.1).

15. Note that MmeI is an unusual enzyme. It is fully active at 50°C and on ice! The enzyme will only cut properly if the reaction includes SAM. Our best success has been cutting 1 µg of PCR product with 1 µl (2U) of MmeI. The reaction is supposed to be 100% complete in 15 min so waiting for 30 min may be overkill. Do not heat-kill the enzyme, but purify the cDNA using QIAquick columns.

16. Change the cartridge on GS20:
   - Turn the four PicoTiterPlate™ cartridge latches to release the cartridge.
   - Gently lift the cartridge out of the camera cover about 1.5 inches (3–4 cm). Rotate the cartridge slightly to see the PEEK nuts screwed into the back of the cartridge.
   - Unscrew the inlet and outlet PEEK nuts from the cartridge, and set the cartridge aside.
   - Install the cartridge needed for your Run: For a 40 × 75 cartridge, connect the short PEEK nuts to the cartridge bypass tube (if not already connected). Place the bypass tube into its slot in the camera cover and press it down until held securely. Screw the long PEEK nuts into the fluidics ports in the back of the cartridge. For a 70 × 75 cartridge, disconnect the short PEEK nuts from the cartridge bypass tube (if not already disconnected), and store the bypass tube away. Screw the short PEEK nuts into the left-hand side fluidics ports in the back of the cartridge. Screw the long PEEK nuts into the right-hand side fluidics ports.
   - Place the new cartridge down into the camera cover.
   - Turn the cartridge latches to secure the cartridge in the camera cover.
   - Squirt 50% ethanol on the newly installed cartridge and wipe it with a Kimwipe. Allow to air dry fully.

17. Due to the shorter length of 454 reads as compared to 3730 reads, when performing the blast-based vector screen, you may wish to use a less restrictive value for “S”. The maximum achievable scores for an alignment are based in part on the length of the alignment. And since 454 reads may be only 40 bp in length, high scores may not be possible.

18. Due to the shorter length of 454 reads as compared to 3730 reads, when performing the blast-based contamination screen, you may wish to use less restrictive “S”, “S2”, and
“gapS2” score thresholds. The maximum achievable scores for an alignment are based in part on the length of the alignment. And since 454 reads may be only 40 bp in length, high scores may not be possible.

19. The procedure for submitted flowgrams (files from which the fasta sequences were derived) to the NCBI Trace repository is still very young and susceptible to change. The original SFF file used to extract basecalls is a collection of all the flowgrams representing a 454 region. These SFF files are the equivalent of the SCF trace files from ABI sequencing instruments.

20. While the above link does summarize the basics of submitting to the trace repository, keep in mind that it was written specifically for submissions of SCF trace files. The submission of SFF flowgram files is relatively new as of this writing and is not well documented at the NCBI trace repository. These details specific to submitting SFF files come mainly from our experience with submitting these files, and may change as the trace repository puts more rules in place regarding SFF file submissions. Please check the NCBI trace repository regularly for updated notes on submitting SFF files.

Acknowledgements

The authors would like to thank the Genome Sequencing Center’s Sequence Production and Technology Development Groups. The authors are supported by grants from the National Human Genome Research Institute U54 HG003079 and the National Institute of Allergy and Infectious Disease AI 46593.

References


Chapter 9

EST Processing: From Trace to Sequence

Ralf Schmid and Mark Blaxter

Abstract

A common task in EST projects is the conversion of sequence chromatograms originating from gel-based or capillary sequencers into annotated sequence objects. Here we describe the usage of a software pipeline (available from http://www.nematodes.org/bioinformatics/), which has been developed to make the most of EST datasets. This modular software solution is targeted toward small- to medium-sized EST projects and comprises a series of Perl scripts. The software design is based on our experience during EST projects for parasitic nematodes and other species. The trace2dbest module processes sequence trace files and prepares the text files necessary for the submission of the sequences to the public repository dbEST. PartiGene provides facilities for clustering and assembling the ESTs into putative gene objects or unigenes and organizes the data in a relational database. Additional tools are available for annotation and for making the data accessible via the World Wide Web.

Key words: EST, expressed sequence tags, dbEST, Bioinformatics, PartiGene, trace2dbest.

1. Introduction

1.1. Expressed Sequence Tags

Expressed sequence tags (ESTs) are obtained through single-pass read sequencing of cDNA clones derived from mRNA (1, 2). ESTs are widely used for gene discovery, for genome mapping and annotation, and for generation of custom microarrays. Most EST projects create datasets of a size that is beyond the scope of manual processing. Thus analysis of EST data is an ideal field for the development and application of bioinformatics and computational methods. Computational challenges within EST projects are the actual processing of datasets, the rational reduction of data complexity, the organization of data in a meaningful way, and the extraction of informative content.
ESTs pose a few specific challenges that need to be addressed. As ESTs come from samples of mRNA taken from an organism (or part of it) at a specific time, variation in expression levels for particular genes will be reflected in the dataset. Thus, a cDNA library and the ESTs derived from it may contain many sequences from the same gene, while other genes may be represented by a single sequence or not at all. ESTs for highly expressed genes such as some mitochondrially encoded transcripts can be highly redundant. Although this can be partly overcome by presequencing normalization of libraries, it needs to be computationally addressed by the clustering of sequences. Ideally, sequences that are derived from transcripts of the same gene will end up in a single cluster. Therefore, clustering will drastically reduce the level of redundancy and complexity of the dataset.

Furthermore, as ESTs are single-pass reads the resulting sequence does not have the same quality as that derived from multiple reads. Reading errors can occur, and parts of the sequence, particularly at the beginning or end of the read, may be of low quality. Again clustering comes to the rescue as information coming from multiple ESTs belonging to a single cluster can be used to resolve ambiguities. Also a single EST will normally not cover the full length of a transcript. Sequences assembled from multiple ESTs may extend coverage.

1.2. dbEST – A Public Repository for EST Sequencing

The NCBI established dbEST as the public resource to store EST data (3). To date (07/2007) dbEST holds more than 44,000,000 entries from more than 1000 species (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). We strongly encourage the submission of data to this resource, as increased sequence and species coverage is of great value for comparative studies. dbEST simplifies the batch submission of sequences by splitting the information across a set of four types of linked files (for full information see http://www.ncbi.nlm.nih.gov). Each individual EST is represented by an EST file. The EST file holds sequence and basic metadata such as the name of the library, name of any publication existing or planned describing the dataset, and the name of the person who can be contacted for more details. Rather than repeating this information in each and every EST file, dbEST holds data in linked files called ‘Lib’ for library information, ‘Pub’ for publication information, and ‘Cont’ for contact information. To submit a set of ESTs to dbEST, one has to submit these three files along with the EST files holding the sequences. The software described below assists the user in the process of creating and submitting these files.

1.3. trace2dbest and PartiGene

EST processing is a multistep procedure. As ESTs have been in use for more than 15 years, a large number of software solutions for EST processing are available (4–9). During the development of trace2dbest and PartiGene (6), the main focus has been on the
creation of a freely available solution which is robust and flexible, enables cross-species analysis, and keeps identifiers stable during updates. The software has been tested in the processing of many EST projects such as NEMBASE (http://www.nematodes.org) (10), LumbriBase (http://www.earthworms.org) (11), and TardiBase (http://www.tardigrades.org). PartiGene was also used to create PartiGeneDB (http://www.PartiGeneDB.org), a database of all significant EST datasets in dbEST (12). Extensive user guides and tutorials are part of the distribution and material from training workshops is available on request (nematode.bioinf@ed.ac.uk). While trace2dbest and PartiGene form the stable core of our EST analysis pipeline, we have added significant enhancements recently. Translation of ESTs to generate putative peptides is a difficult problem, largely solved by prot4EST (13) (see Chapter 10 of this book). For systems analysis, we have also developed a series of annotation tools (the annot8r series) to add additional functional inference to PartiGene datasets (14). wwwPartiGene is a script that assists in the setup of a Web-based interface to PartiGene databases (15). All these tools are available for download via http://www.nematodes.org/bioinformatics/ (see Fig. 9.1 for an overview).

**Fig. 9.1. Overview of the PartiGene EST-processing pipeline**

The rectangular boxes correspond to software components of the pipeline, the rounded boxes to data stored in databases, and the ellipse to raw data. Raw chromatograms are processed and submitted to dbEST using trace2dbest. PartiGene retrieves data from both trace2dbest and dbEST and clusters and assembles these data. The resulting dataset is stored in a relational database, termed partial genome database. prot4EST translates the clustered and assembled dataset provided by PartiGene; the resulting peptide translations are also stored in the partial genome database. The translated peptides can be annotated using annot8r. wwwPartiGene provides an interface to query the partial genome database via the World Wide Web.
trace2dbest is a software tool, which takes sequencing trace files from EST projects, as produced by gel-based and capillary fluorescent detection sequencers, and processes them in a stepwise fashion into quality-checked sequences. The basecalling program phred (16, 17) is used by trace2dbest to transform trace files to raw sequences. phred also assigns a quality score to each of the bases it calls based on the strength of the signal, the shape of the peak, and the local environment of the peak. trace2dbest then uses cross_match (18) to identify and trim vector sequence and, optionally, *E. coli* or any other contamination. In addition, trace2dbest offers the option to trim off adapter sequence, poly(A) tails, and low-quality sequence. The processed sequences can then be decorated with preliminary annotations based on BLAST (19) searches. Once the sequences have been processed, trace2dbest will also assist in the creation of “Pub,” “Lib,” “Cont,” and “EST” files.

The PartiGene process starts with the download of all ESTs for a particular species from EMBL/GenBank. The user can also supply “local” sequences. These ESTs are then clustered using the BLAST-based algorithm CLOBB (20). The resulting clusters are assembled using the program phrap (18). If the original trace files have been processed by trace2dbest, PartiGene can easily access and use their corresponding quality files during the assembly process. Once the cluster consensus sequences have been generated, PartiGene assists the user in running BLAST searches against locally installed sequence databases for preliminary annotation and comparative analysis. For smaller datasets, the results can be visualized using HTML tables. For larger datasets, where such flat-file, static representation is insufficient, PartiGene stores the data in a relational database. Additional options within PartiGene allow the user to add library- and species-specific information, as well as further annotation, to the database.

2. Materials

In the following sections, we use a few conventions to make the text more readable. We use *bold italics* to highlight file names or directories in the text and we use the Courier font for text to be entered by the user in a terminal window. For the user’s home directory we use `your_username`. When running the program, you have to replace this with your real user name.

2.1. Hardware and Operating System

trace2dbest and PartiGene, the software tools introduced in this chapter, can be run on any standard PC having an ethernet connection and using the Linux operating system. We have used and tested the software on several “flavors” of Linux (Fedora, RedHat,
debian, Suse) as well as on Mac OSX 10.3 and above. trace2dbest and PartiGene are preinstalled on Bio-Linux, a debian-based Linux distribution, which has a large number of useful Bioinformatics software readily installed. Bio-Linux is available free of charge from the NEBC (http://environ.nox.ac.uk/biolinux.html) and provides the user with a ready-to-go Bioinformatics workstation.

2.2. Software Installation

Unless you already have a computer fully set up as a bioinformatics workstation (such as one with the Bio-Linux distribution in place), you may need to install some or even all of the following software packages. We recommend that you install software packages in a specific directory where you keep locally installed software, for example the directory /usr/local/software. When installing software on a Linux/UNIX/OSX Darwin operating system, it is helpful to keep executables in a directory which is in the user’s PATH (Note 1). The PATH variable will tell the operating system where to look for executable programs. A typical location for executables of user-installed software is the directory /usr/local/bin and we will assume this convention below.

2.2.1. Ensuring that Your Perl Installation Is Compatible with trace2dbest and PartiGene

trace2dbest and PartiGene are written in Perl. While Perl is usually distributed with any Linux system, there are a few additional Perl modules, which either are crucial for trace2dbest and PartiGene to function or will enhance their performance. If any of the Perl modules described below need to be installed on your system see Note 2.

BioPerl (21) is an integrated collection of useful routines for bioinformatics tasks, and trace2dbest and PartiGene take advantage of it. BioPerl can be downloaded from the BioPerl project Web page (http://www.bioperl.org).

Using trace2dbest and PartiGene requires the user to enter relevant information interactively via a terminal window. The Term::Readline::Gnu module assists with this process as it allows some convenient features such as filename completion and command history. The module is available from the CPAN Perl module repository http://search.cpan.org/dist/Term-ReadLine-Gnu/.

The e-mail-submission of files to dbEST is handled by trace2dbest using the Mail::Mailer module. This module is not shipped with all distributions of Perl. If it is not installed on your system already, it can be downloaded from the CPAN repository http://search.cpan.org/dist/MailTools/Mail/Mailer.pm.

For Macintosh OSX 10.4 users, a version of trace2dbest and PartiGene is available (from http://www.nematodes.org/bioinformatics/) that assists in installation of the package, including the above-noted Perl modules.

2.2.2. Installing trace2dbest and PartiGene

trace2dbest and PartiGene can be downloaded from the PartiGene project Web page http://www.nematodes.org/bioinformatics/, or alternatively from http://bioinformatics.org/PartiGene. At the
time of writing, both programs are at version 3.01 and under continuing development. The description of software behavior and operation below is for versions 3.01. As in future releases the mode of operation might change, please always consult the documentation for the latest releases of the software.

1. After downloading and unpacking trace2dbest (Note 3), copy the unpacked package into the /usr/local/software directory:

   cp -r trace2dbest_v3.0.1 /usr/local/software

   If /usr/local/software does not exist on your system, you can create it using the command:

   mkdir /usr/local/software

   Please note this operation may require “root access.” If you do not have root access, you need to contact your local system administrator.

2. Now copy the executables trace2dbest.pl and rename_file.pl into the directory /usr/local/bin by entering the following two commands in a terminal window:

   cp trace2dbest_v3.0.1/trace2dbest.pl /usr/local/bin
   cp trace2dbest_v3.0.1/rename_file.pl /usr/local/bin

3. The installation for PartiGene follows essentially the same scheme. First download and unpack the PartiGene package (Note 3). Then copy the PartiGene package into the /usr/local/software directory using:

   cp -r PartiGene_v3.01 /usr/local/software

4. To copy all PartiGene executables at once into the /usr/local/bin directory type the command:

   cp PartiGene_v3.01/*.pl /usr/local/bin

2.2.3. BLAST

BLAST (19) is the standard tool for comparing sequences. Both, preliminary annotation based on pairwise sequence similarity and the clustering program CLOBB (20), which is used by PartiGene (6) rely on BLAST searches.

1. For installation of BLAST go to the BLAST download page http://www.ncbi.nlm.nih.gov/BLAST/download.shtml and download the precompiled “blast” and “netblast” suites for your platform (for example, this is linux-ia32 for standard
Linux installs). Unpack the two files netblast-2.2.13-ia32-linux.tar.gz and blast-2.2.13-ia32-linux.tar.gz (version numbers may be different) (Note 3).

2. Then copy the executables for both installations into the /usr/local/bin directory:

```bash
cp blast-2.2.13/bin/* /usr/local/bin
cp netblast-2.2.13/bin/* /usr/local/bin
```

3. After this move the install directories to /usr/local/software:

```bash
mv blast-2.2.13/ /usr/local/software
mv netblast-2.2.13/ /usr/local/software
```

4. BLAST requires two environmental variables to be set. In our example, BLASTMAT needs to point to /usr/local/software/blast-2.2.13/data where the BLAST matrices have been installed. BLASTDB points to the directory where you will store BLAST databases. In the example given, this is /home/db/blastdb (Note 4).

2.2.4. phred, phrap, Cross_match

trace2dbest and PartiGene rely on the basecalling program phred\(^{(16, 17)}\), the vector sequence matching software cross_match\(^{(18)}\), and the assembly program phrap\(^{(18)}\).

1. All these programs have been developed in Phil Green’s Lab at the University of Washington, and they are available free of charge for academic laboratories from http://www.phrap.org. The software is sent out via e-mail after filling in a license agreement form and comes with instructions for installation.

2. Again we recommend that you store the executables phred, phrap, and cross_match in /usr/local/bin/:

```bash
cp phred /usr/local/bin
cp phrap /usr/local/bin
cp cross_match /usr/local/bin
```

3. In addition, to use phred you need to set the environmental variable PHRED_PARAMETER_FILE (Note 4).

2.2.5. PostgreSQL

1. To start the postgreSQL service (check with your system administrator whether it is running already), you have to log on as user “postgres” (‘su postgres’) and issue the command (though this command may vary slightly for different installations, in this case consult the installation documentation for more info):

```
/etc/init.d/postgresql start
```

(Note that starting the postgreSQL server under Mac OSX Darwin is slightly different.)

2. Every user of PartiGene requires permission to create databases. A postgreSQL user account needs to be created for each PartiGene user. This is done by issuing the command (again as user ‘postgres’):

```
createuser your_username
```

You will be asked whether the new user should be allowed to create databases and new users. The user must be given permissions to create databases. Whether the user is permitted to create other database user accounts as well is in your, or the system administrator’s, discretion.

3. Methods

This chapter describes how to proceed from raw trace files delivered from the sequencing instrument to a set of annotated sequences using trace2dbest and PartiGene. As an example to illustrate how a set of trace files can be processed, we start with some trace files from the spider *Cupiennius salei* (Martin Jones and Mark Blaxter, unpublished; these trace files are available for download from [http://www.nematodes.org/bioinformatics/](http://www.nematodes.org/bioinformatics/)). trace2dbest is employed for processing the sequence trace files and to prepare the submission files for dbEST. In a second step these sequences will be analyzed using PartiGene. This will result in a set of HTML pages containing BLAST annotated, clustered, and assembled ESTs and a relational database, which can include additional information and can be made searchable via a Web interface. We recommend using this chapter as a hands-on guide and companion while actually running the software on a computer.

3.1. From Trace Files to dbEST Submission

trace2dbest uses a set of programs and routines to create quality-checked sequences ready for submission to dbEST and the corresponding submission files from sequence trace files.
3.1.1. Starting trace2dbest

We recommend running trace2dbest from an empty directory, for example /home/your_username/ESTproject. This is where trace2dbest will initially write its output files.

1. You can create this directory by entering the following command in a terminal window:

```bash
mkdir /home/your_username/ESTproject
```

To change your working directory to this directory, enter:

```bash
cd /home/your_username/ESTproject
```

2. trace2dbest is started by launching the command `trace2dbest.pl` in a terminal window. This will display the start screen with seven options to proceed (Fig. 9.2). When running the program for the first time, we recommend that you run option 1 “Setup configuration” and option 2 “Checks and tests,” before starting the actual trace processing.

3. trace2dbest needs to know the locations of a few files that are accessed during the trace processing. This information is stored in the configuration file `.trace2best.conf`, which is saved in the user’s home directory. Starting the “Setup configuration” option allows the user to modify the entries specifying these file locations. The default values comply with the Bio-Linux setup. The “Setup configuration” option guides the user through the entries in the configuration file in a stepwise fashion.

---

![Screenshot of the trace2dbest starting page](image)

Enter the number corresponding to the part of the trace2dbEST process you want to perform:

1. Setup configuration.
2. Checks and tests.
5. Create or view submission information.
6. Prepare dbEST submission files.
7. Exit.

---

Fig. 9.2. **Screenshot of the trace2dbest starting page**
Starting trace2dbest opens the main page and gives the user an interface with six menu points and an exit option to choose from. The program is menu driven and it is recommended to run the six menu points in a sequential order.
For the screening for vector sequence and \textit{E. coli} contamination, trace2dbest needs to be told where to find the respective FASTA files \texttt{vector.seq} and \texttt{ecoli.seq}. In the example given, trace2dbest is installed in \texttt{/usr/local/software}, therefore the respective entries for the \texttt{vector.seq} file and the \texttt{ecoli.seq} file are as follows:

\begin{verbatim}
/usr/local/software/trace2dbest_v3.0.1/
  vector.seq
/usr/local/software/trace2dbest_v3.0.1/
  ecoli.seq
\end{verbatim}

4. Submission of sequences to dbEST requires some associated information. trace2dbest ensures that you do not have to reenter information for dbEST submissions multiple times, and stores these data locally in a set of files: \texttt{Libfile.db}, \texttt{Pubfile.db}, \texttt{Contfile.db}, and \texttt{ESTfile.db}. In the example given, these files are kept in the user’s home directory, therefore the respective entries have to be set to:

\begin{verbatim}
/home/your_username/Libfile.db
/home/your_username/Pubfile.db
/home/your_username/Contfile.db
/home/your_username/ESTfile.db
\end{verbatim}

5. The option “Checks and tests” accesses the required files and software tools and tries to highlight potential problems at an early stage. We strongly recommend running it before starting to process your trace files. First this option checks for the availability of programs such as phred, cross_match etc. A common error is that these programs are either not installed or the executables are not in the user’s path and therefore cannot be found (\textbf{Note 1}). Next trace2dbest looks whether the PHRED\_PARAMETER\_FILE environmental variable is set (\textbf{Note 4}) and for the existence of the files defined in the configuration file.

\subsection*{3.1.2. Naming Schemes}

A controlled, consistent naming scheme for sequence trace files allows trace2dbest and PartiGene to extract information from the file name. For example, trace2dbest can extract microtiter plate number and coordinates from each file name and uses this information for the preparation of the dbEST submission file. At the moment trace2dbest supports two naming schemes, the NERC environmental genomics, or Edinburgh (EG) scheme and the STRESSGENES scheme. A trace file named using the NERC EG scheme would look like “Cs_wim_001A05,” while one using the full STRESSGENES naming scheme would look like
“CcLL03b01a02f2_AbaRb.” The NERC scheme comprises three parts separated by underscore characters, with each part carrying different information. For example in “Cs_wim_001A05” the “Cs” indicates the species (*Cupiennius salei*), “wim” a library-specific mnemonic (in this case the name of the originating laboratory), and “001A05” the 96-well or 384-well plate coordinates (plate 001, row A, column 05). Two digit codes for plates (such as plate 01) can also be used. For details of the coding in the STRESSGENES scheme, see http://legr.liv.ac.uk.

When traces are generated by the ABI3730, they are produced with a three-digit identifier; other sequencers may use different naming formats. To convert these to the NERC EG format, we have developed the script rename_file.pl, which is part of the trace2dbest distribution. The rename_file.pl script takes arguments to modify filenames, for example:

```
rename_file.pl –dir spidertraces –txt tracesspider –sub Cs_wim_
```

would rename all trace files in the directory “*spidertraces*” and substitute every occurrence of the text string “tracesspider” in the file names (such as “tracesspider01A05”) to the correct one (“Cs_wim_01A05”). Other formats may require the use of additional options and arguments. A description of the full set of available options can be accessed via the command:

```
rename_file.pl –help
```

3.1.3. Processing the Trace Files

trace2dbest employs phred (16, 17) to translate fluorescence chromatograms into sequence data. phred is regarded as the “industry standard” for this process. Each nucleotide basecall made by phred is given an associated quality score, which can subsequently be used to trim off low-quality sequence using user-defined cutoffs. Phred deals with a range of different sequencing chemistries (and fluorescent dyes) and sequencer instruments, though newly developed dyes can cause the “unknown chemistry” problem (Note 5).

Raw EST sequence files usually contain a small segment of vector sequence, and there may be adapters used in the construction of the cDNA library at the interface between vector and insert. During the construction of cDNA libraries, there is a risk of contamination, particularly by *E. coli* and bacteriophages. Vector sequence, adapters, and contaminants are removed by trace2dbest using the cross_match (18) tool.

trace2dbest also assists in trimming of poly(A) tails. It will scan all but the first 150 bases of the sequence for continuous stretches of As. It will also scan the reverse sequence for stretches of Ts.
If found, the poly(A) or poly(T) tail and all sequence after it [or before it in the case of poly(T)] is trimmed and this event is recorded in the POLYA field of the EST file.

1. Before processing trace files, the user must ensure that all the trace files to be processed are in a single directory that contains no other files, and that the trace files are named according to one of the naming schemes described in Section 3.1.2. In the following we assume the trace files have been stored in the directory `/home/your_username/spidertraces` and have been named according to the NERC EG naming scheme.

2. The actual processing is started by selecting option 3, “Process traces“, from the trace2dbest interface. Two different levels for the processing are available: “Standard” and “Advanced”. The standard level uses a set of standard parameters and is very useful for a quick screening of trace files. The advanced level gives the user control over several of the processing steps and parameters and allows therefore for a more thorough processing. We will use the advanced option in the following.

3. After indicating that the trace files comply with, for example, the NERC EG naming scheme, the first information to be entered is the directory where the trace files are stored. In the example given, the trace files have been stored in the directory:

   `/home/your_username/spidertraces`

   trace2dbest checks this location for the number of files matching the selected format and asks the user to confirm whether this is the expected number.

4. Next, trace2dbest offers trimming of adapter sequence. If you want to do so simply enter the nucleotide sequence of the adapter. Otherwise just hit the “enter” key. If the adapter sequence is found, it will be removed, along with all residues upstream (to 5').

5. Before vector trimming, trace2dbest allows the user to review the list of the names of the vector sequences available in the `vector.seq` file. If the relevant vector is present, you should accept this setting. If the cloning vector you have been using is not in the list, you can add its sequence to the `vector.seq` FASTA file (Note 6).

6. Next you are asked whether you want to screen for *E. coli* contamination. Accept if you wish to do so. If you suspect a different kind of sequence contamination (for example from other common contaminants or commensals for which genome sequence is available), add the relevant sequences in FASTA format to the `ecoli.seq` file (Note 6).
7. A cut-off value for the minimum number of high-quality bases required for accepting a sequence has to be set. The default value is 150. All sequences having fewer residues will be disregarded. Change this value to a lower number if you want to keep very short sequences.

8. The user also can modify the default parameters for cross_match screening for vector sequence. “minmatch” is the minimum length of a match to nucleate comparison, “minscore” the minimum score to accept a match. Lowering both values would make the search more sensitive, but increases the risk of losing sequences due to false positives.

9. Finally, the user is asked for the minimum numbers of “A”s constituting a poly(A) tail. Normally it is best to accept the default setting of 12 consecutive A bases.

10. There is an option to trim “trans-spliced leader” sequences from EST sequences (for the generation of SL-libraries see Chapter 7). The C. elegans spliced leader 1 (SL1) sequence is preloaded, and to use it enter “yes”. To use any other spliced leader sequence, just enter its sequence.

Now all the necessary information has been gathered and the actual processing starts. As this can take some time, comments informing you of progress will appear on the screen. You will be given information about rejected sequences (due to low quality or length criteria) and a summary listing the number of trace files processed, the number of good-quality trace files, the number of submittable sequences after trimming, as well as the average length and average number of high-quality bases. To continue with processing, go back to the main menu.

3.1.4. Preliminary Annotation Using BLAST

dBEST permits inclusion of preliminary annotation within an EST submission using the “similar to” tag. trace2dbest option 4 “Blasts for preliminary annotation” assists the user in gathering and processing this information. Within this option, you can choose between local BLAST searches [this assumes the required BLAST databases are installed locally (Note 7)], or remote BLAST searches using the NCBI “netblast” system. Please note that for remote BLAST searches, NCBI restricts individual users to 400 searches a day. If you intend to use a locally installed database, we recommend the uniprot or uniref databases from EMBL EBI (22).

1. As the example dataset comprises only a small number of sequences, we will use NCBI’s remote BLAST option.

2. To set the level of stringency, i.e., the level of similarity required to annotate your sequence with the “similar to” tag, the user is asked to set a BLAST bit score cut-off value. The default value is 55, which corresponds, very approximately, to an expect value of 1e-10 when searching the uniref100 database. In this
example we stick with this default value. Higher values result in more stringent annotation. As soon as a bit score cut-off is entered, the BLAST similarity searches are started. Depending on the number of sequences, this can take some time, but you will be continuously updated on progress.

3.1.5. Preparing
Submission Information

The submission of ESTs to dbEST requires not only the sequences, but also additional information. trace2dbest stores this information locally in four files: Libfile.db, Pubfile.db, Contfile.db, and ESTfile.db. Using these files, information that has been entered once can be reused again later for future submissions. The text file Libfile.db stores entries for each cDNA library. It stores the name of the library, the name and strain of the organism sequenced, the cultivar and more specific information such as sex, organ, tissue, stage, cell type, and cell line of the material sequenced. It also stores information on the vector and restriction enzymes used in cloning as well as a general library description. The text file Contfile.db contains name, address, and contact details for the person submitting the sequences. The text file Pubfile.db stores details of any associated publications. Details of sequencing primers that have been used and the date for the release of the data to the public are saved in the text file ESTfile.db.

To feed this information into these files, option 5 “Create or view submission information” is used. This option is subdivided so that each file can be modified individually. Each subdivision follows a questionnaire format, and answers are stored for further use. When running trace2dbest for the first time, the user needs to provide at least one entry for each of these subdivisions. Some of the fields (such as the provision of a library name) are compulsory while others are voluntary as indicated by the interface. As soon as there is some content in these files, the option “create or view submission information” can also be used to simply view the entries stored in the local files.

3.1.6. Submission

In trace2dbest option 6 “prepare dbEST submission files,” the information stored previously is used to create the final submission files. trace2dbest accesses the data stored locally in Library.db, Pubfile.db, etc., and allows the user to select the relevant entries from each of these files. After the user has made a selection, trace2dbest combines this information with the processed sequences to create the submission file. This file is stored under the name dbEST_submission.txt. Next, users are asked whether they want to submit the submission file immediately to dbEST. (In the example given immediate submission is not wanted as these sequences have already been submitted). If “yes” is not entered, the file is not submitted, but instead stored locally for possible later manual submission. Manual submission can be achieved by e-mailing the
text file as an attachment to batch-sub@ncbi.nlm.nih.gov. The submission file and all the other trace2dbest outputs are then moved to a project directory (see Sections 3.1.7 and 3.1.8).

3.1.7. Directory Structure

At this stage trace2dbest tidies up the working directory and moves all the files related to the trace2dbest process to a specifically created, standard directory structure (Fig. 9.3). A well-organized directory structure has many advantages. In particular, it allows different programs to access files that have been stored without the user needing to specify file locations every time. The parent directory of the trace2dbest/PartiGene EST-pipeline directory structure is est_solutions, which is stored as a subdirectory of a user’s home directory. In our example, this is /home/your_username/est_solutions. This directory is organized in a hierarchy of four sublevels. Each species to be processed has its own directory. For example, all data related to the spider Cupiennius salei would be in /home/your_username/est_solutions/Cupiennius_salei. In each of the species directories there is space for the “tools” level of directories. Each “tool”, such as trace2dbest or PartiGene, has its own directory. An additional layer is the “event” level, which for example allows the user to distinguish between different batches of sequences processed. So the full name for the directory where

![Fig. 9.3. The trace2dbest directory structure](image)

The recommended directory structure for trace2dbest comprises five levels. The top level is the /home/user_name/est_solutions directory. The next level, “species”, allows the user to work on more than one species in parallel. Each of the software tools of the EST pipeline occurs as a sublevel for each species level. Each “software tools level” can have the sublevels “event” and “output”. The “event” level allows trace2dbest to store processed data in parallel, for example when exploring the processing of a particular dataset with different parameters.
the trace2best output is stored would be similar to /home/
your_username/est_solutions/Cupiennius_salei/trace2dbest/
05_22_06_17-45-12unsubmitted/. The “event” level is marked
with a date and time stamp such that the output is stored
separately for each event.

3.1.8. trace2dbest Outputs

trace2dbest creates comprehensive output consisting of ten output
directories and two distinct output files. These are saved in a
project directory such as /home/your_username/est_solutions/
Cupiennius_salei/trace2dbest/05_22_06_17-45-12unsubmitted/.
To list them, use the “ls” command on this directory. The output is
described below:

1. **dbEST_submission.txt**: This file contains the merged dbEST
   submission files. Unless you have submitted from within the
   trace2dbest program, this file should be emailed to dbEST
   (batch-sub@ncbi.nlm.nih.gov) as soon as you are ready to
   submit your sequences.

2. **logfile**: The logfile stores progress information from various
   parts of the trace2dbest process. Normally there is no need to
   look into this file, but it may provide vital clues if things go
   wrong.

3. **blast_reports**: If you have chosen to perform BLAST searches,
   this directory contains two files: blast_full which keeps the
   full BLAST reports for each sequence, and blast_tophit
   containing just the top hits.

4. **fasta**: This is a directory storing two types of sequence files.
   “.seq” files are raw FASTA formatted files created by phred,
   “.seqsc” files have vector sequence and *E. coli* contamination
   replaced by Zs and Xs, respectively.

5. **fastafiles**: This is the directory in which the final fully processed
   sequence files are saved in FASTA format. These are the files to
   be used for further analysis, for example in PartiGene.

6. **partigene**: This directory contains “.seq” and “.qual” files.
   These files can be automatically accessed by PartiGene
   when looking up sequence quality information (see Section
   3.2.2.4).

7. **phd_dir**: The phd_dir directory contains the “.phd” files
   produced by phred.

8. **process**: In this directory, files which describe in detail the
   different trimming steps that have been performed are kept.
   There is also a file, statistics, which gives summary statistics.

9. **qual**: This is the directory where “.qual” files (quality score
   matrix files created by phred) are stored.

10. **raw_traces**: In this directory, copies of the original unpro-
    cessed trace files are kept.
11. **scf:** This directory contains the standard chromatogram format files produced by phred.

12. **subfiles:** The **subfiles** directory stores the individual EST submission file and processed sequence file for each trace file processed.

### 3.2. From Sequences to Partial Genomes

PartiGene is a menu-driven, multistep program that processes ESTs and creates a set of sequence objects (putative genes), which we term a partial genome. In using PartiGene, we recommend that you create a specific directory and run PartiGene from within that directory. We further recommend that you use the directory scheme that has been created by trace2dbest, as this is recognized by PartiGene automatically.

1. The PartiGene “tools” level directory is created by typing the following command in a terminal window:

   ```bash
   mkdir
   /home/my_username/est_solutions/Cupiennius_salei/PartiGene
   ```
   (all in one line)

2. To change your working directory to the directory that has just been created use:

   ```bash
   cd /home/my_username/est_solutions/Cupiennius_salei/PartiGene
   ```

3. The program is started by the command:

   ```bash
   PartiGene.pl
   ```

   Starting PartiGene will display the PartiGene main menu and allows the user to select between the four main subsections of the program “PartiGene – setup”, “PartiGene – processing sequences”, “PartiGene – databasing” and “PartiGene – post database analysis”.

### 3.2.1. PartiGene Setup

Before actually using PartiGene a few parameters need to be set. This is done by launching the menu option “PartiGene – setup,” which offers three options:

1. PartiGene requires two specific Perl libraries, PG_basics.pm and PG_db.pm, which need to be installed in the user’s home directory. These libraries are part of the PartiGene distribution and can easily be installed by selecting the option “Install or update libraries”. Enter either the path to a local copy of these files or, if unsure, just hit the “return” key to use the automated download as this will take care of everything.
2. PartiGene also needs to know the whereabouts of a few files and directories. It stores these file locations and a few parameters in the configuration file `.partigene.conf`. This file can be created and updated via the “Create or update configuration file” option. The defaults are set to match a typical Bio-Linux setup. The user will be asked, for each item, whether it should be kept as is, or changed to suit the local installation.

The first entry defines the directory where PartiGene looks for local BLAST databases:

```
/home/db/blast
```

PartiGene uses the `vector.seq` file to screen for vector sequence. The entry needs to be modified so that it points to the location of this file:

```
/usr/local/software/trace2dbest_v3.0.1/
vector.seq
```

PartiGene then asks the user for a name of a database. This database will be the default PartiGene database where the results of the data processing will later be stored. For the example given, “spiderbase” would be a descriptive database name:

```
spiderbase
```

There is usually no need to modify any of the other entries: see the PartiGene documentation for more details. You can also modify the `.partigene.conf` file using a text editor.

3. A further option within the “PartiGene – setup” menu is “Tests”. This option runs a set of tests to ensure that all the necessary programs are installed; it also checks for the presence of essential files and whether the necessary environmental variables are defined. The user is encouraged to run this option before starting to process sequences as it will highlight potential problems at an early stage.

### 3.2.2. Processing Sequences

3.2.2.1. Downloading Sequences

Selecting “PartiGene – processing sequences” from the main menu offers six further options: “Download sequences from EBI for analysis,” “Pre-process sequences,” “Cluster sequences,” “Assemble clusters,” “Perform BLASTs,” and “Additional options.” These options are normally run in a sequential order.

The input sequences for PartiGene are stored as individual files in FASTA format in a directory called `sequences`. This directory is created automatically when running the option “Download
sequences from EBI for analysis.” It is highly recommended to use this option, but see **Note 8** for alternatives and **Note 9** for using PartiGene with data generated by 454 sequencers.

1. To download sequences simply enter the species name or the NCBI taxid (see http://www.ncbi.nlm.nih.gov/Taxonomy.) For our current example, this is

   Cupiennius slei

   PartiGene will contact the European Bioinformatics Institute SRS service and download sequences for *Cupiennius salei*. On submission to dbEST, ESTs obtain an “official” reference database identifier or accession number from the NCBI. These accession numbers permit simple cross referencing between laboratories, and facilitate updating, particularly if many separate laboratories are generating and submitting ESTs for your chosen species.

2. The default is to just retrieve ESTs. Adding fully sequenced mRNAs to the EST dataset can assist in assembly of the sequences to generate clusters, and in some cases, addition of other DNA sequences can also assist. Before the actual download takes place, the program accesses the EBI database, extracts and displays the number of sequences, and asks the user whether he wants to proceed with the download. Accepting this starts the download process.

### 3.2.2.2. Preprocessing Sequences

At this stage there is a set of sequences stored in the `sequences` directory. Since one cannot necessarily assume that all sequences submitted to dbEST have been processed to the high standards of trace2dbest, a “Pre-process sequences” option is offered. This option is a three-step process.

1. PartiGene will screen for and remove vector sequence.
2. PartiGene removes poly(A) and poly(T) tails.
3. The user is asked for the minimum length of sequence to include in the analysis. After entering a number, or accepting the default of 100, PartiGene removes sequences too short for meaningful analysis.

### 3.2.2.3. Clustering Sequences

A cluster in PartiGene is a set of sequences that are putatively derived from the same gene. PartiGene uses simple-cluster identifiers to uniquely tag each cluster generated. A cluster identifier is a three-letter string, where, typically, the first two letters represent the genus and species followed by C for cluster. For the spider, *Cupiennius salei*, the suggested identifier is CSC. Individual sequence clusters are identified by this three-letter identifier followed by a five-digit number. Distinctive cluster identifiers for individual species allow the storing of datasets from more than one species in a single database for comparative analysis.
PartiGene uses CLOBB2.pl to cluster sequences. CLOBB2.pl is based on all against all similarity searches using the rapid megablast algorithm that is part of the NCBI BLAST distribution, and importantly allows for incremental updates. Details of the original CLOBB clustering algorithm are given in Ref. (20). CLOBB2 (A. Yam and J. Parkinson, unpublished) is enhanced by the use of megablast and other refinements that significantly improve speed performance for large datasets (>10,000 sequences). Comparison between CLOBB and other EST clustering options, such as TIGRCLUST, suggests that CLOBB is equivalent in terms of the clustering produced: where there are differences they are likely to arise from chimaeric or otherwise “genuinely” confusing sequences (20). CLOBB’s use of unique and incrementally updatable cluster identifiers makes it suitable for multicentre, ongoing projects.

1. To start the clustering process, select the option “Cluster sequences” and simply enter a three-letter cluster ID:

   CSC

   Depending on the number of sequences, the clustering can take a substantial amount of time. During the clustering process, the user is updated on progress. At the end of the clustering process, PartiGene displays a summary that lists the number of sequences, the total number of clusters comprising clusters with one member (singletons), and clusters with more than one member. This information is also stored in a date- and time-stamped file in the OUT directory (in the example, this will be similar to OUT/CLOBB_CSC_05-30-06+19:25.txt.)

2. The results of the clustering process are saved in the directory Clus. After exiting the program, the command:

   ls Clus

   can be used to display the results. FASTA format text files are generated for each cluster with more than one member (in the example these will be called CSC#####, where ##### is a five-digit number), while all singletons are combined in the file singletons.fasta. Two files are stored directly in the working directory: the file superclusters contains a list of clusters that are potentially related, the file merge keeps a note on clusters that have been merged into one.

   3.2.2.4. Sequence Assembly

   Now all sequences in multimember clusters can be assembled into a consensus sequence. To derive consensus sequences PartiGene uses the program phrap (18). If the original quality scores for sequences are available, phrap can take advantage of them. One benefit of phrap compared to alternative assembly solutions is that it extends the consensus building using 5’ and 3’ overhangs from a single sequence. This gives longer consensus sequences, a feature which is often helpful for downstream processing. As the
stringencies of cluster definition used by CLOBB2 and phrap are somewhat different, the assembly process can result in a cluster being split into two (or sometimes more) consensus sequences. These splits may indicate alternative splicing or examples of allelic variation that were not detected by CLOBB2.

1. After starting the “Assemble clusters” option, the number of clusters to be assembled is displayed on the screen. PartiGene offers three modes to take quality information into account: “for all clusters,” “only for clusters containing two sequences,” or “ignoring them.” If quality files are available, we recommend using them only for clusters containing two sequences as this in our experience reduces the amount of multiple consensuses.

2. For sequence trace files processed with trace2dbest, quality scores are readily available in the prespecified directory structure (see Section 3.1.7). PartiGene will derive the names of the trace files from the FASTA description of the sequence files and semiautomatically access the quality scores. One piece of information PartiGene requires at this step is the trace file identifier. This is everything in the trace name before the first underscore. For trace files such as “Cs_wim_001A05” used in the example given, one would simply enter “Cs”.

3. After the assembly process is finished a date- and time-stamped summary is saved in the OUT directory. In the example used, you should expect something similar to \textit{OUT.phrap\_CSC\_05-30-06\_19:52.txt}.

The results for each cluster are stored in the directory \textit{phrap}. For example, the following command lists the results for cluster CSC00084:

\texttt{ls phrap/CSC00084*}

This will list a set of files:

- \texttt{CSC00084} the original CLOBB cluster
- \texttt{CSC00084.ace} the main phrap output file
- \texttt{CSC00084.qual} the original quality scores for the constituent sequences
- \texttt{CSC00084.contigs} a FASTA file containing the consensus sequence(s)
- \texttt{CSC00084.contigs.qual} quality scores for this file
- \texttt{CSC00084.singlets} any sequence not used for building the consensus sequence
- \texttt{CSC00084.log} the logfile
In some cases there may be additional files: consult the PartiGene and phrap documentation for more information on this.

4. During the assembly process, PartiGene also creates files used for protein translation by prot4EST (see Chapter 10). Sequence (“.seq”) and quality (“.qual”) files for each consensus sequence are stored in the directory protein. A concatenated input file for prot4EST is saved in the directory files_stored as prot4EST_input_CSC.fsa.

3.2.2.5. BLAST Searches

PartiGene can perform BLAST searches against a series of databases. These databases have to be locally installed in the directory specified as BLASTDB in the configuration file .partigene.conf (Section 3.2.1). Particularly useful for annotations is a search against a well-annotated data resource such as uniprot or uniref. For a comparative analysis, in the example given, we will run similarity searches against the Caenorhabditis elegans proteome database ‘wormpep’ (available from http://www.wormbase.org/) and the Drosophila melanogaster proteome database (which we call ‘flypep’ below; available from http://flybase.bio.indiana.edu/) to find out whether a particular gene is also present in insects, or nematodes, or both (Note 7).

1. Launching the “BLAST annotation” option and selecting the suboption “Prepare files for BLAST searches” creates a new directory, blast, in which the input files necessary for BLAST searches are prepared. This option also creates a concatenated FASTA file, which is suitable for external BLAST searches, for example on “BLAST farms”. This file, blast_inputCSC.txt, is saved in the directory files_stored.

2. In the next step “Run BLASTs,” PartiGene will run BLAST searches locally. The user can select up to five locally installed databases. The selection is done interactively by following the instructions given on the screen. In our example, we select “uniprot,” “wormpep,” and “flypep.” After the selection is completed, the BLAST similarity searches start automatically. These searches can take a substantial amount of time, depending on the size of databases being compared and the speed of the computer processor. PartiGene keeps the user updated on progress. After the BLAST searches are finished, the results are stored in subdirectories in the blast directory. To list them exit PartiGene and enter the following command in a terminal window:

   ls blast

   This lists a directory named passed and additional directories specific for each BLAST database, which has been searched against. In the example given, this is uniprot, wormpep, and flypep.
3. The additional option “Import BLAST results from external sources” is only relevant when BLAST searches are performed externally. An example of a BLAST command to use within external resources would be: “blastall -p blastx -i blast_input_CSC.txt -d uniprot -b 20 -v 20 -e 0.001 -o CSC_uniprot.out.” To integrate the external BLAST results into the PartiGene process, one would place the output file **CSC_uniprot.out** in the directory **external_blast** and use the “Import BLAST results from external sources” option.

3.2.2.6. Further Options

Selecting “Additional options” gives access to three further options.

1. When starting the “Creating HTML Tables” option, the PartiGene interface allows the user to select the BLAST searches to be included in the HTML tables. After the selection the relevant files are created and stored in the directory **html**.

2. If PartiGene finds a Web-browser on the computer, the results can be viewed directly (Fig. 9.4).

3. On any system running a Web server, these pages can be made accessible to the world by copying the directories **html**, **blast**, and **phrap** into a Web-accessible directory. Under a Linux/Apache web service, this would usually be /home/your_username/public_html, while on a Mac OSX machine it will be /Library/Webserver/Documents/.

Access to library information is very useful for the analysis of specific expression when comparing EST datasets compiled from more than one library. The option “Retrieve library information” first downloads the file **library.report** from the NCBI. This file contains the “Lib” information submitted with each set of ESTs (see Section 3.1.1). The user is then asked to reenter the name of the species of interest. PartiGene parses the **library.report** file and collects relevant information for all libraries available for this species. In a second step, PartiGene tags all ESTs in the dataset with the name of the library from which they originated. cDNA library information is stored in the directory **library**.

The option “Add species information” asks the user for some species-specific information. This information is stored in the file **species/species.txt** and can be integrated into the relational database at a later stage. Again this feature can be useful for comparative analysis.
3.2.3. Databasing

In particular for bigger datasets and comparative analysis, it is advisable to build a local, relational database. PartiGene allows the user to build a database from the results via the “PartiGene – databasing” menu. Details of the database scheme can be found in the PartiGene manual.

After starting the menu “PartiGene – databasing” an empty database with the name defined in the configuration file (Section 3.2.1, "spiderbase" in the example) is created. Then PartiGene offers the user a selection of six options to feed data into the database. These options hardly require any user input, but we recommend they are run sequentially. Option 1 “ESTs and clusters” inserts the core data, information on clusters and ESTs into the database. In the option “BLAST results,” the user is asked to select which BLAST results to include in the database. In the C. salei example, we select all available: uniprot, flybase, and wormbase. Option 3 “Clone names” uses the NERC EG naming scheme to derive clone names from the FASTA header of the EST entries and inserts them in the database. The options 4 and 5, “Library information” and “Species specific information,” access and add automatically the results from Section 3.2.2.6 to the database.
database. Option 6 “prot4EST results” is an additional option to allow the output of prot4EST to be added to the database (see Chapter 10).

3.2.4. Postdatabase Analysis

The menu “PartiGene – post database analysis” offers three options for data output from the PartiGene dataset.

To create a BLAST-searchable database from the consensus sequences of the *Cupiennius salei* EST dataset we select option 1 “Create BLASTable databases.” This option will retrieve all consensus sequences of the *Cupiennius salei* EST dataset in FASTA format and generate the files that are required for BLAST searches against this particular set of sequences. BLAST-searchable databases are more useful if the constituting sequences contain some annotation in their FASTA header. For this purpose, PartiGene offers the option to add preliminary annotation to the FASTA header. This annotation is based on one of the BLAST results that have been generated in Section 3.2.2.5. Each consensus sequence will be annotated with the “similar to” tag followed by identifier, description, score, and e-value of the best hit of the BLAST results the user has selected. Generally, we recommend choosing “uniprot” or “uniref” for this purpose. The formatted BLAST-searchable database is saved in the directory `files_stored` under the name `CSC.tar.gz` (Note 3). This file provides a resource to search with any sequence for similar sequences in the *Cupiennius salei* dataset in the context of a local BLAST installation.

Option 2 “Annotated EST-files” creates tab-delimited files, which are suitable as input for other software. The user is asked to select one or more of the available BLAST results to be included. In the example given, the resulting file is stored as `ESTs_CSC.txt` in the directory `files_stored`. Option 3 “wwwPartiGene” is at the moment only available as a standalone program from http://www.nematodes.org/bioinformatics/ (see Section 3.3).

3.3. Additional Tools

There are additional tools available from www.nematodes.org to make more of your PartiGene dataset. In particular we recommend using prot4EST (see Chapter 10) to translate consensus sequences into putative peptides. Putative peptides are more robust templates for advanced similarity searches, definition of protein domains, etc. The annot8r series of programs (14) provides tools for annotation using GO-terms (23), EC-numbers (24), and KEGG biochemical pathways (25) as well as domain annotation based on interpro (26). wwwPartiGene (15) assists the user in setting up a Web-based interface to query a PartiGene-derived database similar in style to NEMBASE2 (10). This interface permits text searches of the BLAST results and displays clusters in context of their constituent sequences and annotation. For the *Cupiennius salei* dataset presented via wwwPartigene, see http://www.nematodes.org/NeglectedGenomes/ARTHROPODA/Cupiennius.html.
4. Notes

**Note 1 Adding directories to a user’s PATH**

This note is not intended to replace any introduction to Linux or UNIX where one will find much more complete information, but to assist the less-experienced Linux/UNIX/Darwin user in overcoming issues related to the setting of the user’s PATH. Linux/UNIX/Darwin systems need to be told which directories the operating system should look for executables. The collection of these directories is defined in the variable PATH. In the following example, we show how to include the directory `/usr/local/bin` in the user’s path.

1. The way to set or modify this variable depends on the shell you are using, therefore you first need to find out what shell you are using. Simply type:

   ```bash
echo $SHELL
   ```

2. (a) If you are using a c-type shell (csh, tcsh) open your `~/.cshrc` file and modify (or add) the following line to include the directory `/usr/local/bin`:

   ```bash
   setenv PATH "$PATH:/usr/software/exec/:
   /usr/local/bin"
   ```

   (b) Alternatively, for a bash-type shell (bash, sh) open `~/.bashrc` in a text editor and modify or add the PATH line accordingly:

   ```bash
   export PATH=$PATH:/usr/software/exec:/usr/
   local/bin;
   ```

3. To make the change active, you need to “source” the respective file by running:

   ```bash
   source ~/.cshrc or source ~/.bashrc respectively.
   ```

   Please note, it is important that the “export PATH” and “setenv PATH” entries in `~/.bashrc` and `~/.cshrc` are in one line, i.e., do not contain an “end of line” character.

**Note 2 Installing Perl modules**

We will demonstrate the installation of Perl modules using the Bioperl module as an example. Other modules can be installed in analogy, but consult the respective README and INSTALL files in any case. For Mac OSX users, the installation instructions provided with PartiGene for Mac OSX give Darwin-specific instructions for this process.
1. First unpack the downloaded file as described in (Note 3).

2. Then change your working directory to the directory “bioperl-1.2,” which has been created by the tar command:
   
   ```
   cd bioperl-1.2
   ```

3. Now have a quick look at the README and INSTALL files for details of the module and installation (in some cases there will be only one of these files):
   
   ```
   more README
   more INSTALL
   ```

4. Then, unless the instructions in README or INSTALL tell you differently, issue the following commands:
   
   ```
   perl Makefile.PL
   make
   make test
   ```
   
   It is often not a problem if a few (sub)-tests fail.

5. After the tests login as root (or use the “su” [Linux/UNIX] or “sudo” [OSX Darwin] command) and run the command for placing the module in the correct place:
   
   ```
   make install
   ```
   
   Now everything should be in place.

**Note 3 Unpacking software**

A common way to pack Linux/UNIX/Darwin software and collections of files is the “tar.gz” format. Files in this format can be unpacked in two steps. The following example uses the BioPerl module to illustrate this, the steps for any other “tar.gz” file are analogous:

1. First “unzip” the downloaded file:
   
   ```
   gunzip bioperl-1.2.tar.gz
   ```

2. The second step is to “untar” the unzipped file. This will normally place a series of files into a newly created directory:
   
   ```
   tar xvf bioperl-1.2.tar
   ```

3. In this example a directory called “bioperl-1.2” has been created. To list the content of this directory use the “ls” command:
   
   ```
   ls bioperl-1.2
   ```
   
   Instructions for program installation and important information are normally found in the INSTALL and README files.
**Note 4 Environmental Variables**

Some programs under Linux/UNIX/Darwin require environmental variables to be set. These variables specify file locations, directory locations, or simply values to be accessed by a certain program. Setting these variables depends on the type of shell you are using (Note 1). We will demonstrate the procedure for both shell types using the PHRED_PARAMETER_FILE variable as an example.

1. (a) For a c-type shell open the file `~/.cshrc` and add a line:

   ```
   setenv PHRED_PARAMETER_FILE "/usr/local/phred/phredpar.dat"
   ```

   This assumes that you have installed the file `phredpar.dat` in the directory `/usr/local/phred`, otherwise modify the entry accordingly.

   (b) For a bash shell you would modify the `~/.bashrc` file by adding the line:

   ```
   export PHRED_PARAMETER_FILE=/usr/local/phred/phredpar.dat;
   ```

2. Again you have to run the source command:

   ```
   source ~/.cshrc or
   source ~/.bashrc respectively, to make the changes effective.
   ```

**Note 5 Unknown chemistry problem**

Different fluorescent-sequencing platforms use different dye chemistries and collection routines, and thus store data differently. Every time a new dye/chemistry is used, phred and therefore trace2dbest will encounter the “unknown chemistry” problem. When reading a trace file, phred extracts the chemistry (primer or terminator), the dye type (e.g., rhodamine or big-dye), and the sequencer (e.g. ABI377 or ABI3700) and compares this information with the entries in the `phredpar.dat` file. If no matching entry is found phred will terminate with the “unknown chemistry” error message. To fix this a new entry needs to be added to the `phredpar.dat` file following the instructions given at the end of this file. An example for an entry we have added recently is given below.

```
# additional lines of machine definitions
"DT3730POP7{BDv3}.mob" terminator big-dye
ABI_3700
```
**Note 6 FASTA files**

FASTA is a commonly used simple sequence format. The basic components are “>,” “identifier,” “description,” and “sequence.” The “>” indicates a line which contains “identifier” and “description,” though “description” is not mandatory. Everything between “>” and the first “white space” is the “identifier,” everything after this in this line is the “description.” Any line not starting with “>” is “sequence,” in these lines “white space” is ignored. See the following valid example of FASTA format containing two peptide sequences:

```plaintext
> identifier1 this is the description
THISISSEQXENCEDATA
THISLINEASWELL
> identifier2
AGAIN THIS IS SEQXENCE AS WHITE SPACE IS IGNXRED
```

**Note 7 BLAST databases**

Locally installed BLAST databases are extremely powerful for comparative analysis. We will go in detail through the process of building the uniprot BLAST database. Other BLAST databases can be built in an analogous way.

1. We change our working directory to the BLAST database directory. This is the directory defined in the BLASTDB environmental variable (see Section 2.2.3) and in the PartiGene configuration file (see Section 3.2.1).

```bash
cd /home/db/blast
```

2. Next, we download the uniprot FASTA files. This can either be done via the uniprot Web site (http://www.ebi.uniprot.org), or directly by launching the following two commands in a terminal window (each of them in one line).

```bash
wget ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/uniprot_t_sprot.fasta.gz
wget ftp://ftp.ebi.ac.uk/pub/databases/uniprot_trembl.fasta.gz
```

3. Now we need to unpack the two files

```bash
gunzip uniprot_trembl.fasta.gz
gunzip uniprot_sprot.fasta.gz
```
4. The uniprot FASTA files are split into two. One file is derived from the manually curated swissprot database, the second from translations of embl nucleotide sequences. To have a unified uniprot BLAST database, we merge the two files into the file `uniprot`.

```
cat uniprot_trembl.fasta > uniprot
cat uniprot_sprot.fasta >> uniprot
```

Please note the ‘’`>>`’’ in the second command.

5. All that remains to be done is to “BLAST format” the merged file with the command

```
formatdb -p T -o T -i uniprot
```

A quick note on the parameters used:
- `-p T` specifies protein sequences, you would use `-pF` for nucleotides
- `-i` input file
- `-o T` creates indices for faster database access.

6. Other protein databases used in the example dataset *Cupiennius salei* such as wormpep (ftp://ftp.sanger.ac.uk/pub/databases/wormpep/wormpep) can be processed in the same way.

**Note 8 Using non dbEST sequences in PartiGene**

Though it is possible to process sequences with PartiGene before they are submitted to dbEST and available via the EBI, we recommend that users delay further analysis until they are released in dbEST. This has the advantage that each EST will have a universally recognized identifier. However, if you want to skip the download option and analyze the sequences you have processed using trace2dbest immediately, you can copy the content of the `trace2dbest/event/fastafiles` directory into `PartiGene/sequences` by running the commands (from within the relevant Partigene directory):

```
mkdir sequences

cp /home/your_username/est_solutions/Cupiennius_salei/trace2dbest/05_22_06_17-45-12unsubmitted/fastafiles/* sequences
```

(enter the cp command in one line and modify the timestamped directory according to your setup).

This emulates the output of PartiGene’s “Download sequences from EBI for analysis” option and allows the user to continue with PartiGene’s “Cluster sequences” option.
Note 9 Using data from 454 sequencers

454 sequencing is a massively parallelized high-throughput sequencing technology developed by 454 Life Sciences. PartiGene can be used to further process the FASTA sequence output from 454 sequencers. The first step is to copy the FASTA output files from 454 sequencing into the directory PartiGene/sequences (see Note 8). As next step we recommend to use the vector-trimming option as described in Section 3.2.2.2 to remove vector sequence and then to continue with the clustering step of a standard PartiGene run (Section 3.2.2.3). We note that this step may cause problems when dealing with more than 150,000 sequences due to the scaling of the clustering algorithm.

Acknowledgements

The authors would like to thank all contributors and users of trace2dbest, PartiGene, and other tools of the Edinburgh EST pipeline, in particular Alasdair Anthony, John Parkinson, James Wasmuth, and Ann Hedley. Funding was in part from the NERC Environmental Genomics Thematic Data Program.

References


18. Green, P. phrap unpublished.


Chapter 10

Obtaining Accurate Translations from Expressed Sequence Tags

James Wasmuth and Mark Blaxter

Abstract

The genomes of an increasing number of species are being investigated through the generation of expressed sequence tags (ESTs). However, ESTs are prone to sequencing errors and typically define incomplete transcripts, making downstream annotation difficult. Annotation would be greatly improved with robust polypeptide translations. Many current solutions for EST translation require a large number of full-length gene sequences for training purposes, a resource that is not available for the majority of EST projects. As part of our ongoing EST programs investigating these “neglected” genomes, we have developed a polypeptide prediction pipeline, prot4EST. It incorporates freely available software to produce final translations that are more accurate than those derived from any single method. We describe how this integrated approach goes a long way to overcoming the deficit in training data.

Key words: Expressed sequence tags, ESTs, protein translations, simulated transcriptomes.

1. Introduction

The current most popular uses of expressed sequence tags (ESTs) are:

- assisting in gene finding in a complete genome sequence,
- underpinning large-scale expression studies, either microarray or serial analysis of gene expression (SAGE),
- identifying single nucleotide polymorphisms.

However, ESTs also present an opportunity to perform comparative analyses previously confined to organisms with complete genomes. Many of these studies use protein sequences as the unit for comparison. These polypeptide sequences present a better
template for almost all annotation, including domain determination with Interpro (1) and Pfam (2), as well as the construction of more accurate multiple-sequence alignments, the creation of protein-mass fingerprint libraries for proteomic studies and structure prediction using comparative techniques such as threading or other modelling methods [see (3) for overview]. Using a protein dataset also eases study of metabolic and protein-protein interaction pathways. To robustly use the added value of the polypeptide predictions from an EST dataset, the coding region of each EST must be identified.

1.1. Translating Expressed Sequence Tags

The structure of mature mRNA is consistent throughout the Eukaryota. A typical mRNA can be divided into a 5'-cap, the 5'-untranslated region (5' UTR), the protein-coding region [also known as the coding sequence (CDS) or open reading frame (ORF)], 3' UTR and a poly-adenylated [poly(A)] tail. As an EST does not usually cover the entire mRNA, it is not necessarily obvious where the coding region should start, and in ESTs that derive from only the 5' or 3' UTR regions there may be no coding region at all. The problem of finding the coding region is further compounded by the low quality of ESTs, as they are derived from only a single-pass read. Errors due to low-sequence quality can result in apparent shifts in the reading frame (due to missing or inserted bases) or premature termination of the coding region or errors in the amino acid sequence (due to ambiguous or incorrect nucleotide predictions).

The issue of low quality can be mitigated by clustering of ESTs and prediction of consensus sequences, as detailed elsewhere in this volume. As well as producing a consensus sequence for a cluster that is longer than the constitutive ESTs, assessing the basecalls in the overlap between the constituent sequences goes someway to overcoming the problem of ambiguous nucleotides. However, current clustering methods are agnostic as to frame-shifts, and make no effort to identify the coding region. Additionally, for most EST projects, the majority of ESTs are singletons, and there are thus no sequences to use for comparison. We here use ‘EST contig’ to refer to both the consensus sequence derived from a cluster and to the sequence of a cluster containing a single EST.

Methods for derivation of protein (or polypeptide) sequences from EST contigs can be split into three broad categories: similarity-based, composition-based models and \textit{ab initio} approaches.

1.2. Similarity-Based Methods

A translation from an EST contig that displays statistically significant similarity with a known protein is very likely to represent the correct translation for that EST contig. We have shown that the most accurate approach to identify the coding region in an EST contig is to map the nucleotide sequence onto a protein sequence
with which it shares significant similarity (4). The BLAST (5, 6) and FASTA (7) algorithms are ideal for this type of search and have been used by a number of groups in locally maintained solutions. protEST (NB not to be confused with the prot4EST presented in this chapter) uses a similarity-based approach starting from known protein sequences (8). A protein sequence is compared to an EST database. The phrap program (9, 10) is then used to construct a consensus sequence from the cluster of ESTs detected to have similarity. The consensus is then compared to the original EST using ESTWISE (Birney unpublished, http://www.ebi.ac.uk/Wise2) giving a maximum likelihood position for possible frame-shifts. The system is accurate but not readily adaptable to the high-throughput approach necessary when dealing with very large numbers of ESTs. More crucially, any ESTs that do not show significant similarity to a known protein are ignored and not translated. These similarity-based methods require that the nucleotide sequence shares detectable similarity with a protein in the selected database. Many genes, from both well-studied and neglected genomes, do not share detectable similarity to known proteins. For example, our recent analysis of the Caenorhabditis elegans proteome showed that only ~50% of the 22,000 proteins have matches to the Pfam collection of domain models, and 43% have no detectable similarity (using BLAST) with non-nematode proteins in the UniProt database (11). Therefore, sequence similarity approaches are only applicable for part of most EST collections.

1.3. Composition-Based Models

To overcome the reliance upon sequence similarity, composition-based approaches based on the recognition of potential coding regions within poor-quality sequences have been developed. The programs work by using a training set of known full-length coding regions to populate or estimate parameters for a probabilistic model. To maintain integrity, coding datasets are typically derived exclusively from the species under study. The three most widely used methods are DIANA-EST (12), ESTScan2 (13) and Decoder (14), with each method characterising known coding regions in different ways. DIANA-EST uses artificial neural networks trained on positive and negative examples of coding sequences; the negative training set comprises out-of-frame exons and non-coding regions (12). ESTScan2 uses the frequency of short, contiguous nucleotides, the n-mer (where ‘n’ is the word length), from full-length coding regions of the species under study. The authors found that the use of hexamers resulted in the most accurate predictions (13). Decoder was developed to define start codons and open-reading frames in full-length cDNA sequences, using a rule-based method to identify possible insertions and deletions (indels) in the nucleotide sequence, as well as finding the most
likely initiation site (14). The primary user input for Decoder is a codon bias table for the species from which the EST contigs were generated.

The major drawback with these methods is their reliance upon training sets of full-length coding regions for creating the probabilistic models. The size of the training set required to be representative frequently exceeds the available data for many EST projects. We have extensively tested the effect of using suboptimal training datasets, and find that the accuracy of coding region prediction is indeed affected, but that problematic results are easily detectable. A more detailed review of these methods is available from the prot4EST website (4).

### 1.4. Ab Initio Approaches

One final method, commonly used, requires nothing more than some of the basic principles of biology: the coding region of the mRNA starts with a methionine and terminates with a stop codon. Applied to the problem of EST contig translation, the nucleotide sequence is first translated in all six frames, then the longest open-reading frame – the region between a methionine (or the start of the sequence) and nearest downstream stop codon (or the end of sequence) – is considered to be the putative coding region. This approach is naïve in that it assumes that the nucleotide sequence is error-free.

### 1.5. prot4EST – A Combined Solution

It is clear that no single method is able to overcome all the problems facing the majority of EST analysis projects. This motivated the creation of a solution using the best of several methods to enhance the quality of the coding region predictions, while minimizing the effects of their shortcomings. prot4EST is an EST contig translation pipeline written in Perl, with a user interface (4). In addition to linking together a number of freely available programs, it carries out the retrieval and formatting of files from online databases. Once the preferred algorithm has been used on each sequence, the putative translation is extended both upstream and downstream of the prediction. This process is necessary as similarity-based methods may ignore legitimate coding regions whose sequence is novel to the organism of interest. prot4EST has been designed as a stand-alone tool, but is easily integrated into the PartiGene system described in Chapter 9 (15).

### 1.6. Simulated Transcriptomes

The accuracy of ESTScan is dependent on the training set; it must be representative of the species under study. These training sets, which are full-length coding sequences (CDSs), need to be typically very large, and such resources do not exist for the majority of EST-sequencing projects. Previously, we have shown that using CDSs from a species with similar sequence composition (measured by adenine plus thymine content) give more accurate translations
than do CDSs from a more closely related species with a divergent A+T content (4). Reasoning that the amino acid content (or more strictly the dipeptide content) between related species was likely to be more conserved than the A+T content, we have developed a method that uses ‘simulated transcriptomes’ as the training set. Briefly, the codon-usage table for the species under study is estimated from BLASTX searches of the EST contigs against a well-curated protein database. Regions of the EST contigs that show significant translated peptide similarity with sequences from the protein database are isolated and used to derive a codon-usage table based on (usually) tens of thousands of codons. The distribution of codon-usage for each amino acid is then used to back translate the proteome of a closely related, ‘model’ organism. This generates a simulated transcriptome, comprising a large number of CDSs that have the sequence composition of the species under study, which can then be used to train the models employed by ESTScan. Performing this step prior to running prot4EST results in the generation of more accurate coding region predictions (Wasmuth, Cutter and Blaxter, unpublished observations).

2. Materials

2.1. Conventions Used

In the following sections, we use a few conventions to make the text more readable. We use **bold italics** to highlight file names or directories in the text and we use the Courier font for text to be entered by the user in a terminal window. The command prompt present in your terminal window is represented here with ‘#>’. During installation (see Section 2.3.2) we assume that your current working directory is /tmp/. For the other sections, your current working directory does not affect the running of the programs. However, you should be aware of the directory structure you create and keep it as simple as possible. For the user’s home directory, we use **your_username**. When running the program you should replace this with your real user name. Occasionally, there are command line entries that span more than one line of page text; for clarity we use a backslash (\) character to separate the lines. In practice, you can ignore the blackslash when typing on the command line.

2.2. Computing Hardware

All the software required for prot4EST is capable of running on UNIX/LINUX-based operating systems. The procedure has been tested on the following flavours of LINUX: Red Hat LINUX (version 9), Fedora Core 4 and Ubuntu (versions 5.10, 6.04 and 7.04). You will need to know the root administrator password for your machine.
2.3. Software Compilers

The programs mentioned in this chapter are written in the Perl, C, or Fortran programming languages. It is important to ensure that you have compilers for these three languages installed on your computer. Typically these are included with a LINUX installation, but if not present on your system they are easily installed with your system’s ‘package manager’.

2.4. Software Sources

It is important to take close note of where files, in particular executables (or program files), are kept. It is usual to keep install and keep programs in the ‘/usr/local/bin’ or ‘/usr/bin/’ directories, although use of the ‘/opt/’ directory is becoming more common. Here we have presumed that you are placing the new software tools in /opt/.

2.5. prot4EST Requirements

The following programs are required for the prot4EST pipeline:

- **BioPerl** ([http://www.bioperl.org](http://www.bioperl.org)) – at least version 1.2, although we recommend upgrading to version 1.4
- **EMBOSS** ([http://www.emboss.org](http://www.emboss.org)) – prot4EST only requires the *transeq* program from the EMBOSS package, but we suggest that you install all the EMBOSS programs, as they are extremely useful for subsequent analysis
- **ESTScan_v2.0** ([http://www.isrec.isb-sib.ch/ftp-server/ESTScan/](http://www.isrec.isb-sib.ch/ftp-server/ESTScan/)) – select version 2.0b
- **Decoder** – Decoder is licensed software, and thus you must sign a (free) academic use license before you access the program. The license is available from the programs’ authors: email, genome-ofc@gsc.riken.jp
- **ANSIColor.pm** ([http://www.cpan.org](http://www.cpan.org)) (Note 1). The installation instructions for each of the above programs are available with their distribution.

2.6. prot4EST Installation

From the prot4EST website, you can download the source code of prot4EST, in addition to requesting support from the development team:

```
http://www.nematodes.org/bioinformatics/prot4EST.
```

The download is unpacked with the command:

```
#> tar -zxvf prot4EST_2.3.tgz
```

Change to this directory:

```
#> cd prot4EST_2.3/
```

To install the program:

```
#> perl Makefile.PL
```
Now log on as root/superuser:

```
#> su root
[enter password when requested]
#> make install
```

After you have installed ESTScan (see Section 2.3) you will need to run a script that allows it to be integrated with prot4EST. You should run this as root/superuser:

```
#> ./ editESTScan.pl
```

If the install is successful then the main program (`prot4EST_2.3.pl`) will be found, by default, in `/usr/bin/perl` and perl library files would have been moved into your Perl path. A set of gzipped files containing databases of rRNA genes and mitochondrial proteins will be in the directory `prot4EST_2.3/`. You should move these to the location in which you usually store your BLAST databases (such as `/opt/blast-2.2.13/db/`). To unpack these files you can use:

```
#> tar -zxvf mitoMetazoan90.bldb.tar.gz
#> tar -zxvf rRNAeuk.bldb.tar.gz
```

Here we use the metazoan mitochondrial database, which is appropriate for the *Heliconius melpomene* dataset explored within this chapter. Users analyzing other taxa may wish to construct their own mitochondrial sequence datasets.

You may need to refresh your operating system’s list of executables. You will then be able to run the program.

```
#> rehash
#> prot4EST_2.3.pl
```

This will cause some pre-run checks to be carried out. If you have not yet installed the external programs, you will receive some warning messages.

To aid the smooth running of prot4EST and the other programs used it is necessary to set up a number of environment variables. These are recorded in a special file that is read when you login or open a new terminal window. This file will need to be edited to define three variables. The file name depends on the shell you use. If you are using csh or tcsh, then you will need to edit your `~/.cshrc` file. If you are using bash, then you will need to edit your `~/.bash_profile` file. There are three variables that must be set, one for ESTScan and two for BLAST:

- ‘ESTSCANDIR’ must be the path to where you installed ESTScan (e.g. `/opt/ESTScan/`):

  ```
  setenv ESTSCANDIR /opt/ESTScan/ for ~/.cshrc
  export ESTSCANDIR=/opt/ESTScan/ for ~/.bash_profile
  ```
‘BLASTMAT’ points to the location of the substitution matrices, some of which are available with the NCBI BLAST download (e.g. BLOSUM62):

```bash
setenv BLASTMAT /opt/blast-2.2.13/data/ for ~/.cshrc
export BLASTMAT=/opt/blast-2.2.13/data/ for ~/.bash_profile
```

‘BLASTDB’ is the location of the BLAST formatted databases available on your computer. For more information consult the BLAST documentation:

```bash
setenv BLASTDB /opt/blast-2.2.13/db/ for ~/.cshrc
export BLASTDB=/opt/blast-2.2.13/db/ for ~/.bash_profile
```

For these environmental variables to take effect it is necessary to run:

```bash
#> source ~/.cshrc
#> source ~/.bash_profile
```

### 2.7. Tutorial Files

With prot4EST successfully installed, you should download the tutorial files for this chapter (from http://www.nematodes.org/bioinformatics/prot4EST). The compressed archive should be unpacked:

```bash
#> tar -zxvf HMC_example.tgz
```

You will now have the following files:

*HMC_nuc.fsa* – the EST contigs for the butterfly *H. melpomene*,  
*HMC_v_uniref100.blx* – BLASTX reports comparing the *H. melpomene* sequences with the UniRef100 division of the UniProt protein database (16)  
*protein/* – a directory containing the sequence quality files for each EST contig.

### 3. Methods

The EST clusters used to illustrate the procedures described in this chapter are from the butterfly *H. melpomene* (17). Here, we describe just one way in which prot4EST can be used to analyse a collection of EST contigs with quality files available for each contig. As described above, we have found that accurate predictions for organisms with small training sets can be achieved by using simulated transcriptomes as the training sets for the ESTScan and Decoder components.
3.1. Creating the Simulated Transcriptomes

The program `simulaTomes.pl` is used to create a simulated transcriptome. Briefly, a comparison between the EST contigs and a protein database is parsed to identify the distribution of codon usage for the species under study. This distribution is used to reverse translate a protein collection, typically from a model organism. Here, you will use protein sequences from *Drosophila melanogaster* as the template for the generation of a simulated transcriptome for *H. melpomene*.

3.1.1. Files Required

A BLASTX report using the EST contigs as the query sequences against a protein database (**Note 2**) is required. For this example, the database used was UniRef100 (16). The BLASTX report file must be in the default NCBI BLAST format. The file `HMC_v_uniref100.blx` is provided. The template file containing the protein sequences that will be reverse translated is `dmel-all-translation-r5.1.fasta` and was obtained from FlyBase (18). The genetic code file (`gc.prt`) is included in the NCBI BLAST package in a directory called `/opt/blast-2.1.13/data`.

3.1.2. Running the Program

At the command line to find the list of options required for the program type:

```bash
#> simulaTomes.pl
```

You will be presented with the help menu. For this instance, to create a simulated transcriptome for *H. melpomene* use the command below. For information on the `-taxid` option see **Note 3**.

```bash
#> simulaTomes.pl -blx HMC_v_uniref100.blx \\
 -queryF HMC_nuc.fsa -gcF$NCBI/data/gc.prt \\
 -gcN 1 -templateF dmel-all-translation-r5.1.fasta \\
 -species Heliconius_melpomene -taxid 34740
```

The program will now run to completion with progress summaries printed to screen. Three files are generated, two of which are required for later analyses. The simulated transcriptome is saved as `Heliconius_melpomene.embl`. This file will contain 19,770 coding regions in the EMBL database format and will be used for training the model used by ESTScan.

3.1.3. Codon-Usage Table

The codon-usage table, which will be used by Decoder, is saved as `out.cod`. You may change this to a more useful name:

```bash
#> mv out.cod HMC.cod
```

3.1.4. Training ESTScan

The ESTScan package comes with its own script to train the hidden Markov model (HMM) used to identify coding regions in EST contigs. This script, called `build_model`, can be run at the command line (see below), or within prot4EST if you do not have a training set ready (**Note 4**). It is important that the installation
process has been run as detailed above, with particular reference here to editESTScan.pl (version 2.3.2.5). Your control of build_model is through a configuration file, an example of which (Hm_embl.db) has been included in this chapter’s tutorial package (2.7). If you have different filenames or species, the variables $organism$ and $dbfiles$ in the configuration file should be altered to suit.

To train the ESTscan models, type:

```
#> build_model Hm_embl.db
```

Once the program has finished, a number of new directories and files would have been produced in your working directory. The matrix file that contains the probabilities for the ESTScan HMM is found in the Matrices directory. This should be copied to the present working directory and renamed:

```
#> cp Matrices/6_00030_0000001_4242.smat HMC.smat
```

### 3.2. prot4EST – Preparing Data for Coding Region Prediction

The simplest way to use EST contigs as input is to have them in a single file in the FASTA format (for help with sequence formats see [http://www.ebi.ac.uk/help/formats.html](http://www.ebi.ac.uk/help/formats.html)). This file should contain all the EST contigs that you wish to translate. If you wish to make use of the Decoder component, you will also need to set up a directory that contains both the sequence and quality scores for each consensus (Note 5). These files are provided with the packages downloaded for this tutorial.

### 3.3. prot4EST – The Configuration File

The primary point of interaction between the user and prot4EST is the configuration file. We advise that sufficient time is taken to ensure that all the options are entered correctly. The program will review all the information entered. These checks include ensuring the paths for the various files are valid and taking steps to prevent overwriting of results from previous runs. In the downloaded package is a file called config, which should be opened in your favourite text editor (e.g. emacs, vi, pico etc.). Each option is explained below. The sample entry given should be altered depending on how you have set up your system.

1. **Input File [fasta format]:**

   This is the full path to the file or directory that contains the consensuses you wish to translate. In this instance, you should use the single file provided:

   ```
   /home/your_username/est_solutions/heliconius/HMC_nuc.fsa
   ```

2. **Output Directory:**

   The name of the directory you wish the output files are stored. The only two output files not placed in this directory are the progress and error logs:
3. Organism Name (full):

The name of the organism from which the EST contigs are derived. prot4EST will only process sequences from one species at a time:
Heliconius melpomene

4. Location of genetic code file:

The full path to the `gc.prt` file. This file is included with the prot4EST download package. This file contains the various genetic codes available for BLAST analysis. Later you will be asked which genetic codes you wish to use for translation (see Section 3.4). Ideally this file should be placed with your BLAST installation:
/usr/local/ncbi/data/gc.prt

5. Ribosomal RNA database [.fsa .nhr .nin .nsq]:

The full path to a BLAST-formatted database of ribosomal RNA genes. The one provided with the prot4EST package is from the Ribosomal Database Project (19). We advise that all BLAST database files are placed in a standard BLAST database location on your computer:
/usr/local/ncbi/db/rRNAeukAll_nuc.fsa

6. Mitochondrial protein database [.fsa .pin .phr .psq]:

The full path to the BLAST database of mitochondrion-encoded proteins. The one provided with the prot4EST package was assembled from the NCBI GenBank FTP site. We advise that all BLAST database files are placed in a standard BLAST database location on your computer:
/usr/local/ncbi/db/mitoMetazoan90_pro.fsa

7. E-value for rRNA search (BLASTN):

The expectation (E) value for the BLASTN search performed in tier one of the prot4EST process (see below). You can enter the value in one of two formats: 0.001 or 1e-3. Our experience suggests that the default value given is optimal for detection of true rRNA sequences:
1e-65

8. E-value for BLASTX searches:

The expectation (E) value for the BLASTX searches performed in tiers two and three (Note 2). You can enter the value in one of two formats: 0.001 or 1e-3. From extensive empirical testing, we recommend the following default value:
1e-8
9. Location of pre-computed BLASTx report files/directory:

It is possible to run BLASTx searches within prot4EST. However, we recommend that these searches are performed outside the prot4EST program for reasons of speed and efficiency. For very large datasets (those containing more than 2,000 EST contigs), we suggest using a computer cluster (Note 2). The location of the BLAST report file is provided here:

/home/your_username/est_solutions/heliconius/HMC_v_uniref100.blx

10a. Path to sequence and quality files:

This is where you specify the location of the directory containing both sequence and quality files for your EST contigs. It is important that this directory contains only these files:

/home/your_username/est_solutions/heliconius/protein

10b. Suffix for EST files:

The file extension used to designate sequence files in the directory given above:

.seq

10c. Suffix for quality files:

The file extension used to designate quality files in the directory given above:

.qlt

10d. Path to precomputed DECODER predictions:

If decoder predictions are already available for the consensuses, then you can enter the path to the directory containing these files. For this exercise leave this blank.

11. ESTScan Matrix File [optional]:

A matrix file can be generated using prot4EST: the EMBL database is searched for full-length CDS of the species under study, and these are then processed through the build_model script automatically (Note 4). However, for this exercise you should use the matrix built in Section 3.1, as it is more representative of the sequence content for H. melpomene:

/home/your_username/est_solutions/heliconius/Hc.smat

12. Codon-usage Table (gcg format) [optional]:

Similar to the matrix file, prot4EST is able to search the codon-usage database (20) (Note 6). This is not necessary here, as simulaTomes.pl generated this to reverse translate the template proteome:

/home/your_username/est_solutions/heliconius/Hc.cod
3.4. *p4E in Action*

The following files should be in your current working directory: *config*, *protein/*, *HMC_all.fsa.*

To start prot4EST enter:

```
#> prot4EST_2.2.pl
```

After initial checks for dependencies and environmental variables (see above section), you will see the welcome screen (Fig. 10.1). You have already composed a configuration file (Section 3.3), and so should select option:

2

You are now asked to provide the location and name of the configuration file. For this example simply type:

```
config
```

At this point, you can choose to edit the configuration file using the program pico. However, you should choose to use the file by typing:

```
U
```

If the checks performed by prot4EST find an incorrectly entered file, you are returned to the welcome screen. From there you can either exit the program or use the edit (E) function.

```
fusiformis heliconius_melpomene> prot4EST_2.3.pl
starting prot4EST checks...

# Initialize prot4EST
# Version: 2.3 - Feb '07
#
# Predicting coding regions from expressed sequence tags has never been so easy.
#
# Developed at:
# The Institute of Evolutionary Biology, University of Edinburgh.

Please set up the config file:
1. Create a configuration file.
2. Use or Edit an existing configuration file.
4. Exit Program.
```

Fig. 10.1. The welcome screen of prot4EST.
You are then asked which genetic codes you wish to use for translation. The choices presented here are taken from the gc.prt file that comes with the distribution of prot4EST. These choices dictate the genetic code used to translate the consensuses. It is important to choose the correct codes.

The first choice is for the ESTs encoded by genes in the nuclear genome. For the majority of cases, including this one, you want the standard (nuclear) code:

Next you should choose the invertebrate mitochondrial genetic code:

You have now provided all the information necessary for the prot4EST run. Each stage in the pipeline is accompanied by progress reports printed to the terminal screen. For more details consult the user guide included with the installation package.

3.5. prot4EST Output

The directory that you provided in Section 3.2.1 (e.g. HSC_output) will now contain all the output generated by this run of prot4EST. There are four key files in this directory that are of interest to you.

3.5.1. translations_xtn.fsa

This FASTA-formatted file contains the polypeptide sequences predicted from the consensuses. The sequence header contains the sequence identifier (peptide accession). For PartiGene identifiers, the final letter of the three-letter code is altered to indicate that it is a peptide object. For example, the nucleotide accession id, HMC, becomes the peptide accession id, HMP. Further information in the FASTA header includes the method used to produce the peptide.

3.5.2. nt_coding.fsa

This file stores the nucleotide sequence of the protein-coding region and matches the polypeptide in translations_xtn.fsa.

3.5.3. prot_main.psql

Nearly all the information relating to each predicted translation is contained within this file. The comma-separated variable (CSV) format ensures easy transition into a database table or spreadsheet. The contents of each column are:

peptide_accession | nucleotide_accession | method of translation | gene_location | ext_start | conf_start | start_frame | conf_end | ext_end | end_frame

The first two fields have been covered above (Section 3.5.1). method of translation – the algorithm used to translate that sequence (sequence similarity, ESTScan, Decoder, longest ORF) gene_location – indicates whether the sequence is predicted to be nuclear or mitochondria encoded.

ext_start – the sequence coordinate of the coding region. This is with respect to the nucleotide EST contig.
conf_start – this coordinate indicates the start of the coding region if the extension process had not been run. This can be the same as ext_start and is with respect to the EST contig.

start_frame – the reading frame in which the start of the peptide is translated.

The other three fields (ext_end, conf_end and end_frame) are the 3\textquotesingle equivalents of the above descriptions.

3.5.4. prot_HSP.psql

This file presents the regions of similarity between a BLAST-translated consensus and a known protein. These are termed the highest scoring pairs, or HSPs. The file is in CSV format to make it easy to import into a database.

\[\begin{array}{|c|c|c|c|c|}
\hline
hsp\_reference & peptide\_acc & start & end & frame & evalue \\
\hline
\end{array}\]

hsp\_reference – a unique number included to satisfy database normalisation requirements.

peptide\_acc – this identifier matches that in prot_main.psql

start – the starting location of the translation with respect to the consensus

end – the end location of the translation with respect to the consensus

frame – the reading frame the translation is in with respect to the consensus

e-value – the e-value score associated with this HSP (the hit between the query and database entry).

3.5.5. Other Files

The following files are generated during the prot4EST process. If additional clusters become available, these files could be used by prot4EST to ensure consistency between datasets.

Cs.\textit{smat} – this is the matrix file used to construct the hidden Markov models necessary for ESTScan

decoderfiles/ – a directory that is organised to allow DECODER to run efficiently

blastreport/ – the blast reports are held in this directory

4. Notes

1. **Installing CPAN modules**

   prot4EST.pl uses coloured fonts when printing progress reports to the terminal window. This is done with the Perl module \texttt{Term::ANSIColor.pm}. The easiest way to install this module is through CPAN (Comprehensive Perl Archive Network), with the following commands:
You may need to run this command as root/superuser. You will now open up a shell into CPAN. If this is the first time you have used CPAN, you will be lead through some brief questions about the setup of your computer. If you are unsure of the answer, simply hit return.

cpan> install Term::ANSIColor
The installation should proceed without further user intervention.

cpan> exit

2. Uses of BLAST in prot4EST
prot4EST is able to run BLAST searches locally. The tutorial provided in this chapter provides a precomputed set of results from BLASTX searches. This serves two purposes: first, it is used to calculate codon usage for the species, which is then employed to generate the simulated transcriptome. Second, prot4EST is able to accept it instead of running the BLASTX searches itself. This is important, as the protein databases you are likely to use contain too many sequences for BLAST searches to be completed in a realistic time. Dedicated computer clusters are now relatively common and even the most modest will allow you to perform BLAST searches more effectively than on your personal computer. To run BLAST searches from within prot4EST, you need to leave question 9 in the configuration file blank and the BLASTDB environmental variable must be set (see Section 2.6). The available protein BLAST databases are then listed on the terminal window and you are asked to choose one. The searches then take place.

3. Using taxonomic identifiers
When creating the simulated transcriptomes with simulaTo-
tomes.pl, it is necessary to provide the taxonomic identifier (or taxid) for the species of study. The taxonomic identifier is assigned to each species that is represented in the taxonomy database available at the NCBI (http://www.ncbi.nlm.nih.-gov/sites/entrez?db=taxonomy). If your species lacks a taxo-
nomic identifier, then feel free to assign your own as it is used only as an internal check within prot4EST.

4. Building ESTScan models within prot4EST
In this tutorial, a simulated transcriptome was created for training the HMM used by ESTScan. This training set was created before running prot4EST. However, if you prefer to use only true CDSs available for the species under study, then prot4EST can perform all the necessary tasks if the following steps are carried out:
a) When writing the configuration file (Section 3.3), do not enter any information for question 11.

b) prot4EST will then connect to the EMBL database (21), search and extract all full-length CDSs available for the species.

c) If performed for *H. melpomene*, you will be informed that there are only 1837 coding nucleotides available for this species. Type ‘Y’ to continue. At this stage you will be asked if you wish to use the results from the BLAST searches to complement this number – type ‘Y’ to accept this option. This process differs from the simulated transcriptomes described previously. The BLAST report is scanned for EST contigs from the studied species that are suitable for inclusion in the training dataset (e.g. appear to be full length or include the coding region start site). The `build_model` script (see Section 3.1) is then used to build the matrix needed by ESTScan.

5. **Using Decoder in prot4EST**

While ESTScan corrects errors in the contig through modelling sequence composition, Decoder also uses quality scores associated with each base to identify regions that may contain a possible frame shift. We have designed prot4EST to take a directory containing all the sequence and quality-score files and iteratively process these through Decoder. To ensure the smooth running of this process, we have applied constraints on the naming conventions of the files. Briefly, for each EST contig there should be one file containing the FASTA-formatted nucleotide sequence, and one file containing the quality scores (see Fig. 10.2 for the format). The file name of these files must be the identifier for the sequence entry should be accompanied by a file containing the phrap quality scores for each nucleotide. The format is adapted from the FASTA format. The scores are separated with a space character and are limited to fifty per line. The grey box highlights a region with lower quality nucleotides and is one possible location for a frame shift error.
sequence, with the suffix for each file specified in the configuration file (see Section 3.3). For example, a sequence with the identifier HMC00493_1 will need two files – HMC00493_1.seq and HMC00493_1.qlt. For further examples look at the protein directory provided with the tutorial download.

6. Building your own codon-usage table for Decoder

Through generating a simulated transcriptome (Section 3.1), a codon-usage table for the species under study is estimated. This can be used by the Decoder component of prot4EST. However, if no codon-usage table is available before starting the program, prot4EST will connect to the codon-usage database (20) and search for entries for the chosen species. For *H. melpomene*, leaving question 12 in the configuration file blank results in the program searching for and finding a pre-computed codon-usage table. Unfortunately the publicly available data are meagre, and this table is based on only 249 codons. At this point you could choose another species: we advise against this and suggest that it is better to use the codon-usage table generated previously (Section 3.1).

Acknowledgments

We would like to thank our colleagues in Edinburgh and Toronto for support and user feedback, and the authors of the Decoder and ESTScan programs for permission to use their code. This work was carried out while JW was a BBSRC-funded PhD student.

References


Obtaining Accurate Translations from Expressed Sequence Tags


Chapter 11

EST Databases and Web Tools for EST Projects

Yao-Qing Shen, Emmet O’Brien, Liisa Koski, B. Franz Lang, and Gertraud Burger

Abstract

This chapter outlines key considerations for constructing and implementing an EST database. Instead of showing the technological details step by step, emphasis is put on the design of an EST database suited to the specific needs of EST projects and how to choose the most suitable tools. Using TBestDB as an example, we illustrate the essential factors to be considered for database construction and the steps for data population and annotation. This process employs technologies such as PostgreSQL, Perl, and PHP to build the database and interface, and tools such as AutoFACT for data processing and annotation. We discuss these in comparison to other available technologies and tools, and explain the reasons for our choices.

Key words: EST database design, data processing, annotation, Web interface, data submission.

1. Introduction

EST data are generated for a number of reasons, from the exploration of the coding content of poorly studied species to the investigation of differential gene expression under different physiological or environmental conditions. Irrespective of the particular scientific aim, there are practical issues relevant to all EST projects: EST sequences need to be assembled into contigs to reconstruct the sequences of the (ideally complete) transcripts; contigs must then be associated with the inferred sequence of the protein and its potential function. These data, together with information on the organism, library, tracefiles, etc., must be stored in a systematic and safe fashion, such as a database system. It is critically important that the database allows selective retrieval of sequences and associated information, so that the researcher can draw specific insights from the data.
This chapter addresses the considerations involved in the design and implementation of an integrated EST database system that performs all of the above functions. Rather than specifying explicit protocols for building a database, we focus on providing the reader with advice concerning which software is best suited to the needs of EST databases, and which design decisions are important. As a readership, we envision computer-knowledgeable biologists, bioinformaticians, or research teams with combined biology and informatics background.

Design questions that we will address touch on efficient ways to populate a database with large amounts of ESTs. In addition to managing data, an EST database is useful for guiding the data-production process. We describe how to assess sequence quality and redundancy, which will provide critical feedback to the data-generating wet lab. Since database usage vitally depends on ease of access by scientists, Web interfaces must be intuitive and simple to use. Here, we will discuss three central issues, the performance of data retrieval, the visual representation of the query interface (graphics, lay-out, tools), and data visualization that aims at the comprehension of complex biological data. Evidently, large-scale EST projects, if conducted efficiently, must include automated protein-function assignment. We show how existing tools can be integrated in the data-processing pipeline of a database. Many of the examples we use are drawn from a database initially constructed in the context of the Canadian collaborative Protist EST project with a focus on protists (1), the Taxonomically-Broad eukaryote EST Database (TBestDB) (2) (http://tbestdb.bcm.umontreal.ca/).

The overall EST data processing and functionality of an integrated EST database are illustrated in Fig. 11.1. The major steps will be described in the following sections.

2. Materials

Implementation of a Web-accessible EST database as described in the Methods Section requires the following resources.

2.1. Hardware

The minimum requirement are two servers, one that hosts the live version of the database available to the end users, and another computer, configured as similarly as possible, on which to maintain a copy of the database for development purposes. The second server is also used as a backup, should the live server fail for any reason.

2.1.1. Database Servers

Depending on the amount of EST data, it may be possible to carry out much of the processing pipeline on the same computer where database development takes place. However, for processes that
Fig. 11.1. Flowchart for EST data processing as realized in TBestDB. Clustered ESTs go through several processes in parallel: annotation with AutoFACT; an initial rapid annotation procedure to identify a conserved core of proteins; translation into hypothetical protein sequences in all six frames; and checking for chimerism and potential contamination.
take up large amounts of space or memory, such as function annotation routines involving searches of many different database resources, it is preferable to have additional computational power devoted to those tasks.

2.1.3. Data Transfer Server

It is also necessary to consider how data will initially enter the system. A database serving several laboratories at geographically distributed locations will need some organized method for receiving tracefiles. Given the size of these files (usually of the order of a megabyte each) and the numbers usually involved in an EST project, an sftp server is a reasonable approach to this requirement. As the sftp server must be open for nonlocal users to download files and be secure at the same time, it is recommended that a dedicated computer be used for the purpose.

2.2. Software

Below we list the freely available software for the implementation and maintenance of a Web-accessible EST database, which needs to be installed on the local computer.

2.2.1. Database Management System

PostgreSQL (http://www.postgresql.org/) is a freely available relational database management system. It has ample power and flexibility to address all the needs of an EST database.

2.2.2. Programming Language

The programming for data population, submission, pipelining, or format change can be done in the Perl programming language (http://www.perl.com/), which, together with the PHP Web-design language (http://www.php.net/), is freely available and thoroughly supported. Perl scripts call the various proprietary programs involved, transfer data between them, and populate the database accordingly at each step (see Fig. 11.1).

2.2.3. Large-Scale Function Annotation Tool

Developed for the LINUX/UNIX platform, AutoFACT (3) is a tool for automatic function annotation and classification of ESTs. Code and detailed instructions for installation and use are available at http://megasun.bch.umontreal.ca/Software/AutoFACT.htm.

2.2.4. Web Server

A functional Web database requires a Web server to transfer information from the database to the users. We recommend the freely available Apache Web server software (http://httpd.apache.org/), which is powerful and flexible enough to meet all the needs of an EST project.

2.3. Human Resources

With regard to human resources, several skill sets are needed. The initial installation of the appropriate software, integration of the computers in a larger network environment, and maintenance of this infrastructure require substantial support by a systems administrator. Once a database has been set up, this effort reduces to occasional troubleshooting.
The construction of an EST population pipeline requires a programmer skilled in the appropriate scripting language. The development of a Web interface demands some experience in Web design. For a mature database, both these roles can reasonably be filled by one person, but in the set-up phase these are two distinct full-time jobs for a timescale of several months.

Finally, there is the task of curating the database and maintaining the content, one which in the initial phase requires a great deal of interaction with programmers and users as issues specific to the individual installation arise. This position is ideally filled by someone familiar with both the relevant biological background and the computational tools used.

3. Methods

3.1. EST Database Design

The types of data which an EST database needs to store depend to some extent on the intended purpose of the EST project. This is reflected by the diversity of current EST databases. Databases for single-organism projects, such as the ChickEST database (http://www.chick.manchester.ac.uk), and PoplarDB (http://mycor.nancy.inra.fr/PoplarDB/index.html) can have similar structures, but in an organism-specific context. ESTIMA (4) is a tool for the management of such EST datasets and has been applied to many species, including honeybee, cattle, and zebra finch (http://titan.biotec.uiuc.edu/ESTIMA).

More general databases, such as TIGR gene indices (http://compbio.dfci.harvard.edu/tgi), Nembase (http://www.nematodes.org/nematodeESTs/nembase.html), and PartiGeneDB (http://theileria.ccb.sickkids.ca/), host ESTs from multiple species, and therefore require more complex and flexible infrastructures. With sequences from multiple species stored under the same database structure, general databases facilitate the comparison of ESTs across organisms, such as transitive annotation (see Section 3.3). Here we use TBestDB (2) as an example to illustrate what needs to be considered in database design for taxonomically broad EST databases.

The most important decision in the design of the database is whether query speed or usage of memory and disk space is the higher priority. In the case of a database accessed through the World Wide Web, particularly one that is publicly available or serves a large number of users, speed is paramount. This point is noteworthy because many textbooks on database design [for example (5, 6)] teach optimization techniques based on the
historically valid assumption that reducing disk space usage is always the highest priority. Yet optimizations slow queries down and can considerably reduce the utility of a Web interface.

**Figure 11.2** shows an overview of the structure of TBestDB (http://tbestdb.bcm.umontreal.ca/). A more detailed schema is available at http://tbestdb.bcm.umontreal.ca/docs/
TBestdb-schema.html. This database was designed with the objective of processing daily new sets of EST data in the order of a few hundred to a thousand ESTs. Certain basic data are associated with an individual EST, such as its individual EST identifier (ID), sequence data, and a link to the original tracefile. These data are then organized hierarchically, by batch (corresponding to an individual dataset submitted by a given user), by library of origin, and by organism, with appropriate additional information at each level of organization. In multiorganism databases such as TBestDB, each organism is given an identifier, usually the initial letters of the species (e.g., ‘PA’ for Paramecium aurelia), which is then used as the first two characters for all IDs related to data from that species. This makes automatic searching and sorting of the data a trivial task, and also makes the species to which any given ID belongs immediately human-recognizable.

Cluster consensus sequences and data associated with each cluster (generated as described in the next section) are stored in tables. Each row corresponds to one cluster and is given a unique cluster ID. Clusters are newly assembled each time new data are added to the database. Subsequently, the EST table is updated with information for each EST, detailing which cluster it is assigned to and the position it covers on the consensus sequence. This allows the generation of graphics and multiple alignments illustrating cluster structure (see Fig. 11.3). Moreover, statistics are calculated at different scales to show, for example, the distribution of clustered ESTs and EST lengths for a given library, providing crucial information for deciding whether to continue sequencing ESTs from that library. Additional data, mostly derived from various functional annotation procedures, are stored in other tables as shown in Fig. 11.2. For example, consensus sequences translated in all six frames and putative open-reading frames are stored in the ORF table, all indexed by cluster ID.

Fig. 11.3. Graphical representation of the composition of an Acanthamoeba castellanii cluster in TBestDB. EST, labeled with EST IDs (ACE...), are shown as grey bars at the appropriate positions on the consensus sequence (cluster ID ACL00010836). ESTs originating from different libraries are indicated using different degrees of grey shades. Arrowheads indicate the original direction of sequence readings, and ESTs that have been reverse-complemented during the process of cluster assembly are indicated as hollow bars.
The organization of the data associated with EST projects makes it particularly suited to the relational database model. We have chosen to implement TBestDB in the relational database management system PostgreSQL that uses the SQL query language (Note 4.1). For the sake of speed, EST databases should be designed so that frequently used information can be obtained by querying a single table, instead by joining tables, which is a time-consuming process. This implies that information is stored in more than one table, with the consequence of increasing the database size. Again, for the sake of speed, statistics — particularly those for the whole database — are best calculated during the population process and stored in a separate table, which can be looked up directly, in preference to calculating them “on the fly” at the time of the query.

Data are populated into these tables at various points in the EST-processing pipeline. TBestDB uses a set of scripts written in-house to handle dataflow (see Fig. 11.1), call the various programs for clustering, annotation, etc., and write the data to the database at each step of the process. The flexibility and powerful string-handling capacities of perl make it the appropriate programming language for this task, and its speed is adequate for the scale of processing required.

### 3.2. EST Data Processing

EST reads are received from automated sequencers in the form of tracefiles that contain a series of peaks representing the residues at each position along a sequence. The first stage in processing of EST data for storage and analysis is to convert tracefiles into a string of letters representing the nucleotide sequence. We use the phred/phrap suite of programs for processing tracefiles. To reconstruct the sequence of gene transcripts, ESTs are grouped into clusters based on sequence identity. We recommend the phrap program for clustering and calculation of the consensus sequence, because it uses output files from phred containing numeric quality information for each position, which allows for manual or automatic assessment of overall sequence quality (Note 4.2).

A persistent problem with clustering EST data is that of chimeric ESTs, where a single EST contains sequence originating from two genes, leading to incorrect clustering. Chimerism can be recognized by several criteria, most simply by misplaced polyA tails or sequences with a polyA “tail” at either end. It is also possible to check whether at any point along a (well populated) cluster there is only a single EST supporting the consensus sequence. Chimerism can be corroborated by dividing the consensus sequence at the point of suspected incorrect assembly, and carrying out similarity searches separately with each subsequence, which should match transcripts from different genes. Potentially chimeric sequences are labeled, and can be removed from the data at the user’s discretion.
Identifying the protein sequence corresponding to a consensus EST sequence is a challenging task, as there is no guarantee that a given cluster will contain the start codon or the polyA tail. Any method of inferring protein sequence from ESTs must also be capable of handling frameshifts (Note 4.3). A protocol for predicting the most probable protein sequence from EST sequences is provided in Chapter 10. In TBestDB, the nucleotide sequence is simply translated into protein in all six reading frames, all the putative ORFs are illustrated in a diagram, and the positions of the best BLAST matches used to annotate the cluster (see Section 3.3) are marked on the corresponding ORFs. This is usually a reliable indication of the correct protein sequence.

### 3.3. Web-Based Tools for Analysis of EST Datasets

To identify the function of the genes whose transcripts are present in the cDNA library, cluster consensus (and singleton) sequences are annotated via an automated pipeline. The fastest way to assign function to an unknown sequence is by similarity search against sequences of known function in repositories such as GenBank (Note 4.4). Such searches are usually performed by BLAST (http://www.ncbi.nih.gov/BLAST/). BLAST-based annotation is in principle straightforward, but the challenge is to provide the most informative description of sequence function. Obviously, simply transferring the annotation of the top BLAST hit is not meaningful, if for example the matched sequence is annotated as “hypothetical protein.” Several tools have been developed for the purpose of automatic sequence annotation [e.g., (4, 7, 8)], but each has limitations such as using only a single source of annotation; being part of a much larger pipeline; requiring complicated installations and setup; generating only a single output format; or requiring manual intervention. We recommend the Automatic Functional Annotation and Classification Tool AutoFACT (3), particularly for projects investigating taxa that are not closely related to well-studied model organisms. This pipeline has been developed specifically for EST sequences, and can be easily integrated into the data-processing pipeline of EST databases as it is a command-line-driven program written in Perl for LINUX/UNIX operating systems.

As shown in Fig. 11.4, AutoFACT takes nucleotide sequence as input and conducts BLAST searches against multiple databases (Note 4.4) as specified in the configuration. The corresponding results are evaluated using a hierarchical procedure. Any sequence giving a significant match against ribosomal RNA is classified as such. All BLAST results for remaining sequences are examined, down to a user-specified e-value cut-off, for the best informative hit against each protein database. Uninformative hits, as for example those labeled as “hypothetical” or “unknown,” are disregarded. Informative hits are then compared for common terms based on a user-defined order of preference, and if the same terms are found...
in more than one database, the sequence is annotated with those terms. Sequences with no informative hits, or no shared informative terms, are then compared against protein-domain databases and annotated appropriately if they contain a known domain. If not, they are labeled as “unassigned protein.” In the absence of significant hits against either full-length protein sequence or protein-domain databases, the sequence is searched against the est_ others database, and labeled as “unknown EST” if a match is found, and “unclassified” otherwise. Functionally annotated sequences are then assigned to KEGG pathways, COG functional groups, Enzyme Commission (EC) numbers, GeneOntology (GO) terms and locus names based on their descriptions. By default, AutoFACT uses the order: uniref90, kegg, cog, nr. However, the databases used, and the order of priority in which annotations are applied, can be set by the user. AutoFACT performs as well as, and in some cases better than, the alternatives, with an error rate of only 1–2%.

Annotation of short EST sequences and/or ESTs from nonmodel systems is often inefficient. To alleviate this shortcoming, AutoFACT has an option to employ “transitive
annotation,” consisting of two steps. If an EST sequence does not match a functionally annotated protein but matches an entry in dbEST (or any other EST database), the dbEST sequence homologous to the unknown EST is extracted and used as input to the **AutoFACT** pipeline. Functional classifications that are assigned to this homolog are then used transitively to annotate the initial unknown EST sequence. This process is only permitted to generate links one level deep, in order to avoid the possibility of long chains of transitive annotation extending through a series of related ESTs.

It is important that the users of an EST database can verify the evidence of an annotation assigned to a cluster sequence. E-values, alignment scores, and length of hits against each of the target databases are provided so that the user can assess the quality of the annotation and give more weight to annotations well supported by more than one data source. Additional formats (flatfile and GFF) are useful for easy incorporation of the searches into any database, and it is straightforward to implement searches of the annotation text. **Figure 11.5** shows how AutoFACT results can be integrated into a database interface for user assessment. On the technical side, **AutoFACT** is simple to install and highly configurable and has modest hardware requirements.

### Annotation Data for ACL00000152

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
<th>Percentage Identity</th>
<th>LOCUS</th>
<th>EC Number</th>
<th>Informative Hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniRef 90</td>
<td>Alcohol dehydrogenase class III related cluster</td>
<td>635</td>
<td>5e-65</td>
<td>74.8 (117/157)</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>NCBI 82</td>
<td>Alcohol dehydrogenase class III (Marinobacter aquaeolei V79)</td>
<td>635</td>
<td>5e-65</td>
<td>74.8 (117/157)</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>KEGG Pathway</td>
<td>Alcohol dehydrogenase class III related cluster</td>
<td>635</td>
<td>5e-65</td>
<td>74.8 (117/157)</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>dbEST Others</td>
<td>Alcohol dehydrogenase class III related cluster</td>
<td>635</td>
<td>5e-65</td>
<td>74.8 (117/157)</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>KEGG</td>
<td>Alcohol dehydrogenase class III (EC1.2.1.1.1.1.1)</td>
<td>635</td>
<td>5e-65</td>
<td>74.8 (117/157)</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Fig. 11.5.** Sample HTML output for **AutoFACT** annotation from TBESTDB showing the results of a query for *Acanthamoeba castellanii* EST cluster ACL00000152. Automatic annotation results are displayed at the top of the page and all the information used to infer the annotation is represented in the table, with one row per database accessed (see Section 3.3 of the text). **Percentage Identity** is the extent to which two (nucleotide or amino acid) sequences, in a High Scoring Segment Pair (HSP), are invariant. In the case of the results from dbEST Others, the reported percentage of sequence identity refers to a “translated nucleotide – translated nucleotide” comparison. The values in the **Informative Hit** column specify whether the first, second, etc., BLAST hit in the corresponding database was informative. The page also contains links to relevant database entries.
On a cautionary note, the principle of function prediction by sequence similarity has limitations, and examples of similar sequences that do not share the same or even related functions have been documented in the literature. Obviously any annotation system based on existing databases can only be as good as the source data, and there is a risk that erroneous annotations may be propagated. Automatic annotations give an excellent lead, but do not replace critical assessment by a scientist.

When the BLAST-based AutoFACT annotation returns “unassigned protein,” “domain (e.g., zinc-finger)-containing protein,” or “unknown EST,” this means that no recognizable similarity to known proteins in other organisms was found in the searched data repositories. In the case that the focus of an EST project is on hypothetical proteins, machine-learning based annotation methods can be applied to sequences without obvious homologs (Note 4.5).

In addition to functional annotation, more detailed analyses can be conducted with EST data, and several Web tools have been developed to this end. For example, SpliceNest (http://splice-nest.molgen.mpg.de/) visualizes gene structure and alternative splicing based on EST clusters. GEPIS (gene expression profiling in silico, http://www.cgl.ucsf.edu/Research/gentech/gepis/gepis.html) integrates EST and tissue-source information to compute gene expression patterns in normal and tumor samples. Finally, SimiTri (http://www.nematodes.org/bioinformatics/SimiTri/index.shtml discussed in Chapter 12) allows simultaneous display and analysis of taxonomic relationships of multiple datasets.

Note that some pipelines, such as ESTAP (http://staff.vbi.vt.edu/estap), combine the functionalities described in Sections 3.2 and 3.3. We carry out the two procedures separately to provide more flexibility.

3.4. Web Interfaces for EST Databases

The function of a database Web interface is to provide important information to the user as clearly as possible; what information is important will vary from user to user. It is therefore essential to allow for flexibility in how the user can query the database. While details of interface design will depend on the requirements of a given project, there are some basics inherent to the structure of clustered EST data. An interface to an EST database needs to represent data associated with each EST, and also to provide views of this data at higher levels of organization, e.g., by library or by organism.

It is vital to represent the arrangement of ESTs within a cluster in an accurate and informative fashion. This can be illustrated in two complementary ways. A multiple sequence alignment shows
the precise details of the position of every EST in the consensus sequence and illustrates the choices made by the clustering algorithm in cases of ambiguity. However, this alignment is usually too large to be easily studied as a whole. Therefore, another image summarizes the make-up of the cluster, representing sequences as appropriately positioned bars. For visual comprehension, bars corresponding to ESTs from different libraries can be filled by different colors, ESTs that have been reverse-complemented in the course of the assembly procedure can be shown as hollow bars, and arrowheads may indicate the direction of transcription (Fig. 11.3). Several different relevant pieces of information are thus shown on a simple image, which is generated in real-time by the PHP interface, with an explanation immediately to hand. The image provides a visual representation of the degree of coverage of any given consensus sequence; coverage at each position can be measured and averaged in order to measure overall redundancy within a dataset.

We recommend implementing the interface in the widely used PHP language. This freely available Web-design language allows user-specified search criteria to be composed into sophisticated SQL queries, and to generate Web pages containing the query results. With queries returning large numbers of results, such as retrieving all annotated clusters from an organism, the data are best divided across pages showing 100 results each. Otherwise, these pages become too unwieldy and slow to display. Facilities need to be provided for downloading sequences or traces, individually or in bulk, for users who wish to analyze sequences on their own computer. It is important that the code and query structure is made as straightforward as possible for the benefits of speed and ease of maintenance. In TBestDB, for example, the underlying structure of a table matches almost exactly the structure in which the information is displayed on the screen, as this approaches the ideal for ease of query formulation and speed of result retrieval. PHP also allows editing the underlying data via the Web interface, for example to add comments or manual annotations.

The interface to an EST database can provide password-controlled user accounts, so that data may be kept confidential. Access permissions are best controlled at an organism-by-organism level, whereby each user obtains access to a defined set of organisms within the database. This level of access control can easily be overlaid on an EST database, though it requires secure storage of usernames and password information in a separate external database. Different levels of access permission can also be defined, so that, for example, some users may only be able to read certain data, while others may also have the facility to edit annotations.
4. Notes

1. EST data can be stored in various Database Management Systems (Oracle, Sybase, MS Access/SQL, MySQL, PostgreSQL, etc.) and their features are well documented in the corresponding manuals. There is no single optimal choice of DBMS for an EST project. ESTIMA, for example, uses Oracle, and PartiGeneDB uses PostgreSQL. MySQL is employed for PoplarDB and ChickEST database. We chose PostgreSQL for TBestDB, because it possesses many of the powerful and secure features of high-end commercial RDBMSs such as Oracle, while being freely available.

2. Several other clustering algorithms have been developed: Chapter 9 describes CLOBB, which allows the incremental updating of clusters; TGICL is a pipeline for analysis of large ESTs and mRNA databases; d2-cluster, included in stack-PACK (http://genomics.msu.edu/stackpack/index.html), allows for clustering of EST sequences in situations where quality files are not available for ESTs, with the drawback of providing less auxiliary information as to the decisions taken in the clustering process.

3. Available tools that are developed for correction of frameshifts caused by sequencing errors (e.g., insertion and deletion) include ESTScan (http://www.ch.embnet.org/software/ESTScan.html) (9), and DIANA-EST (10).

4. Standard set of databases used:
   - European Ribosomal Database: Large subunit (LSU) ribosomal RNAs, Small subunit (SSU) ribosomal RNAs (11).
   - Uniprot’s UniRef 90, Uniprot’s UniRef100: GeneOntology terms Enzyme Commission numbers Locus names (12, 13).
   - Clusters of Orthologous Groups (COG): Functional categories (14, 15).
   - Protein Families Database (Pfam): Protein domains (17).
   - Smart: Signaling domains, domain architectures (18).
5. Machine-learning based function prediction allows capturing complex and hidden features of proteins. A promising, but still investigational, approach is to represent proteins by their physico-chemical properties; ungapped or gapped dipeptides; and predicted secondary structure or protein fold; then train a predictor with known proteins, and finally use it on unknown proteins (19–28). For research in this area, we recommend the public-domain machine-learning package Weka (http://www.cs.waikato.ac.nz/ml/weka/), which includes modules such as decision tree, support vector machine, and Bayes network algorithms.

Knowing the subcellular localization of an unknown protein can also provide important clues about its function. A dozen or so predictors are available today, using different sequence features and employing various machine-learning approaches. Available tools that can be applied to EST-derived proteins include SubLoc (29), Loctree (30), LOCSVMPSI (30, 31), iPSORT (32), MITOPRED (33), and ESLpred (34). However, the performance of these predictors is not satisfying with those data. The major challenge of machine-learning based prediction of protein function and subcellular localization of EST-derived proteins is that these are often incomplete, typically lacking the N-terminus when proteins are long. We are currently developing a method for subcellular-localization prediction specifically for ESTs, which will be tested in the context of TBestDB.

References

Phylogenomic Analysis of EST Datasets

José M. Peregrín-Alvarez and John Parkinson

Abstract

To date the genomes of over 600 organisms have been generated of which 100 are from eukaryotes. Together with partial genome data for an additional 700 eukaryotic organisms, these exceptional sequence resources offer new opportunities to explore phylogenetic relationships and species diversity. The identification of highly diverse sequences specific to an EST-based sequence dataset offers insights into the extent of genetic novelty within that dataset. Sequences that are only shared with other related species from the same taxon might represent genes associated with taxon-specific innovations. On the other hand, sequences that are highly conserved across many other species offer valuable resources for performing more in-depth phylogenetic analyses. In the following chapter, we guide the reader through the process of examining their sequence datasets in the context of phylogenetic relationships. Performed across large-scale datasets, such analyses are termed Phylogenomics. Two complementary approaches are described, both based on the use of BLAST similarity metrics. The first uses an established Java tool – SimiTri – to visualize sequence similarity relationships between the EST dataset and three user-defined datasets. The second focuses on the use of phylogenetic profiles to identify groups of taxonomically related sequences.

Key words: Expressed sequence tags, ESTs, phylogenetic profiles, comparative genomics, bioinformatics, clustering.

1. Introduction

For many organisms, the generation of expressed sequence tag (EST) datasets provides a first glimpse into their genetic makeup. As such, the focus of many analyses center on the functional annotation of the dataset. This is typically achieved through BLAST searches of either the raw ESTs or the derived partial genomes (see Chapter 9) against well-annotated protein databases such as UniProt (1). Provided peptide translations can be obtained (see Chapter 10); additional analyses include the application of
hidden Markov models to identify protein domains or more specialized tools such as SignalP (2), TMHMM (3), or PSORT (4) to reveal additional sequence features. Functional analyses have proven extremely useful in identifying genes that underpin an organism’s biology. For example, EST projects have unearthed a wealth of new data on allergens, proteases, and immunomodulators for a variety of parasites including mites (5), nematodes (6, 7) and the Apicomplexa (8). However, aside from revealing important functional insights into their host organism, EST-based datasets can also be exploited to reveal evolutionary insights.

The extension of small-scale evolutionary analyses based on a single or small set of genes (phylogenetics) to larger scale analyses focused on genomic datasets (phylogenomics) allows one to explore the diversity of the genetic landscape of an organism. Due to the scale of the analyses, phylogenomics is typically reliant on the use of automated sequence comparison tools such as BLAST. As such, analyses are considered in the context of comparisons of sequence similarity as opposed to precise evolutionary relationships. Although sequence similarity measures are inherently flawed for accurate phylogenetic reconstructions (9), it should be noted that the incomplete nature of the EST datasets renders the identification of orthologous relationships highly problematic in any event. Consequently, for ESTs, sequence-based approaches offer the most practical solution for examining evolutionary relationships with other species as well as allowing the identification of sequences potentially responsible for taxon-specific innovations. Furthermore, phylogenomic analyses serve as a useful filter for the identification of highly conserved sequences, which may be exploited for in-depth phylogenetic analyses. Several large-scale phylogenetic studies have now been published, which demonstrate the power of combining EST and genomic sequences to compile large multigene datasets for resolving deep taxonomic relationships in eukaryotes (10, 11). For example, Philippe and coworkers used a large dataset of 146 conserved genes derived from genome and EST-based datasets for 35 species to resolve evolutionary relationships between the major metazoan phyla (11).

Here we concern ourselves with two standard phylogenomics approaches that may be readily applied as a first-pass investigation into the diversity of an EST-based dataset. SimiTri is a Java-based tool that visualizes sequence similarity relationships between a query dataset and three reference datasets (12) (Fig. 12.1). In addition to showing general trends in such relationships, it also allows the user to identify outlier sequences that might represent genes undergoing different selective forces. The second approach is based on the generation of phylogenetic profiles (13). Once created, these profiles may be readily manipulated to provide insights into sequence diversity and their potential evolutionary origins (14). In the following, we illustrate the methods through the use of a set of 2,800
sequences representing the consensus sequences of ~10,000 ESTs derived from the parasitic cestode, *Echinococcus granulosus*, and processed via the PartiGene process (Chapter 9) (Parkinson, Mai-zels and Fernandez, unpublished data).

2. Materials

This chapter is aimed at users who possess basic computing skills including experience in using basic Unix/Linux commands, writing simple Perl scripts, and installing publicly available software.

2.1. Equipment

The minimal computational infrastructure required is a standard Intel-based workstation with an Internet connection. A suitable model can typically be purchased nowadays for just a few hundred
dollars, e.g., a dual-core processor (2.00 GHz) with 2 GB RAM and 320 GB hard drive. For more extensive analyses involving comparisons of tens or hundreds of thousands of sequences to several hundred species datasets, the reader may wish to consider using a larger number of processors such as a computing cluster (Note 1).

2.2. Operating System and Software

The analyses described here could be readily implemented in most flavors of the Linux operating system; however, we suggest the user considers the use of Ubuntu (http://www.ubuntu.com), which is well supported and offers many useful features. Ubuntu is freely downloadable or available as a free CD. Further details of how to install Ubuntu are available through the Web site. In addition to the Linux operating system, we recommend the following freely available packages:

2.2.1. Perl, an extremely useful programming language for bioinformatics (http://www.perl.org/) – see Chapter 9, Section 2.2.1 for further details.

2.2.2. PostgreSQL, a powerful database (http://www.postgresql.org/) – see Chapter 9, Section 2.2.5 for further details.

2.2.3. BLAST (15), the standard sequence comparison tool (http://ftp.ncbi.nih.gov/blast/) – see Chapter 9, Section 2.2.3 for further details.

2.2.4. Cluster3.0 (16), software featuring a number of commonly employed clustering methods (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm).

2.2.5. Treeview (17), software for visualization of clusters produced by the Cluster3.0 software (http://jtreeview.sourceforge.net/).


2.2.7. Java, a powerful programming language required by Cluster3.0, Treeview, and SimiTri (http://www.java.com/).

2.2.8. BioPerl, a Perl module that provides many useful bioinformatics functions (http://www.bioperl.org/) – see Chapter 9, Section 2.2.1 for further details.

2.2.9. DBI Perl module, typically part of a Perl installation, which may otherwise be obtained from the CPAN Perl repository (http://www.cpan.org) – see Chapter 9, Section 2.2.1 for further details of installing Perl modules.

2.3. Sequence Datasets

In addition to your species (or “query”) dataset of EST-derived sequences, you will need to obtain a series of comparator-sequence (“reference”) datasets. Depending upon the size of available
datasets, the user can consider using nucleotide- or protein-sequence data. We suggest the reader consider the following three sources for their reference datasets (Note 2):

2.3.1. **Sequences from Species Derived from Specific Taxonomic Groups.** In the SimiTri analyses presented here, we were interested in examining the relationships of \(~2,800\) sequences derived from the parasitic cestode, *Echinococcus granulosus*, with a related species, *Echinococcus multilocularis*, and two sister platyhelminth taxa: *Trematoda* and *Turbellaria*. Such datasets are readily retrievable through the NCBI taxonomy Web browser (http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy). For example, in the search box enter “trematoda” and click the “Go” button. Click the Trematoda link to go to the tree view. Then select the “Nucleotide” and “Protein” check boxes and click the “Display” button. You will see the number of sequences (nucleotide and protein) available for each node of the tree. Clicking on the numbers will take you to the sequence download page. Choose the “FASTA” option under display and “File” under send to, to download the sequences. Note that you may need to consider that a group of sequences that you are downloading may include your query sequences (Note 3).

2.3.2. **Sequences Associated with Complete Genomes.** We suggest users obtain nonredundant sets of ORFs for each genome sequence they wish to compare against. The Genomes Online Database (GOLD – http://www.genomesonline.org/) (18) provides an excellent resource detailing information on available genomes and their associated Web sites (Notes 4 and 5). We recommend that for the generation of the most informative phylogenetic profiles (Section 3.2.2) that as many eukaryotic genomes are included as are available. However, a representative set of 20–50 bacterial and archaeal genomes is probably sufficient to compare across prokaryotes (Note 6).

2.3.3. **Sequences Derived from Partial Genomes.** Previously assembled datasets of ESTs offer an extremely useful source of sequence data to supplement the relatively poor taxonomic coverage afforded by the complete genome datasets. Download and save in different files the putative genes predicted for the partial genomes present in the PartiGene Database (http://www.compsysbio.org/partigene) (19) (Note 7).

Sequence datasets should be downloaded as individual FASTA-formatted files (see Note 6 in Chapter 9) and given appropriate names reflecting their content. To build a BLAST database for each dataset run the command:

```
formatdb -i <name_of_FASTA_file> -p T
```
If the sequence dataset is nucleotide, remember to add the ‘-p F’ option to the above command. Once constructed, we recommend that the reference datasets be placed in a single directory, e.g., `/home/my_user_name/blastdbs`.

3. Methods

3.1. Generation of Phylogenetic Profiles

The first step in our comparative analyses is to build so-called phylogenetic profiles. In its simplest form, a phylogenetic profile represents the presence or absence of a sequence across a number of species. Once generated, these profiles can be readily manipulated to examine the phylogenetic diversity of the sequence. To construct the phylogenetic profile for each sequence in your query dataset, a sequence similarity search needs to be performed against each of the downloaded reference datasets. Given the large number of sequence comparisons, such analyses require a high-throughput sequence similarity search tool. One of the most widely adopted is BLAST (15). Although it only analyses local sequence alignments (Note 8), BLAST nonetheless provides a fast solution ideal for large-scale phylogenomic comparisons such as those outlined here. Before proceeding with the analyses, we suggest that you create a project directory with the following subdirectories: `blastoutput` and `parsedblasts`. Move the FASTA file containing your target dataset into your project directory.

3.1.1. Sequence-Similarity Searches

For each query dataset that you wish to compare against, you will need to perform a BLAST search. For example, if our query dataset is a set of nucleotide sequences from *Echinococcus granulosus* contained in a file called `echino.fsa` and there are three nucleotide databases you wish to compare against `turbellaria.fsa`, `trematoda.fsa`, and `e_multilocularis.fsa` you would run the following (Note 9):

```
blastall –p blastn –i echino.fsa –d turbellaria.fsa –e 0.001 –o echino_turbellaria.out
blastall –p blastn –i echino.fsa –d trematoda.fsa –e 0.001 -o echino_trematoda.out
blastall –p blastn –i echino.fsa –d emultilocularis.fsa –e 0.001 –o echino_emulti.out
```

Depending upon the relative numbers of sequences you are comparing against, each search may take from hours to days (Note 1). If you are contemplating BLAST searches against many reference datasets, you may wish to consider writing a simple Perl script to perform them automatically. Once completed, move each BLAST output file into the `blastoutput` subdirectory of your project directory (Section 3.1).
3.1.2. Parsing BLAST Reports

From each BLAST output file (report), we need to obtain the highest BLAST similarity scores for each query sequence that matched a reference sequence. We use the bit score instead of the expectation value (E-value) to avoid the need to normalize comparisons involving datasets of different sizes. BLAST reports can be readily parsed using a simple Perl script and invoking functions from Bioperl (Section 2.2.8). An example of such a script is given below and may be downloaded from (http://www.compsysbio.org/download/papers/methods2008/extract_blasts.pl).

```perl
#!/usr/bin/perl -w
# extract_blasts.pl - A simple script to extract Blast scores
# usage: perl extract_blasts.pl <name of blast output file>
use strict;
use Bio::SearchIO;
my $blastfile = $ARGV[0];
open(FILEOUT,">$blastfile.txt") ;
my $blastin = new Bio::SearchIO(-format => 'blast', -file => "$blastfile");
while( my $result = $blastin->next_result )
{
    my $score = 0;
    my $name = $result->query_name;
    while( my $hit = $result->next_hit )
    {
        my $blastscore = $hit->raw_score;
        if($blastscore > $score) {$score=$blastscore; }
    }
    print FILEOUT "$name $score\n";
}
close(FILEOUT);
```

Copy the script into a file (e.g., extract_blasts.pl) and run it using the command:

```
perl extract_blasts.pl echino_turbellaria.out
```

Substitute echino_turbellaria.out with the name of any BLAST output file you wish to analyze. The output will be a file of the same name with a `.txt` appended to it and contain a list like:

```
EGC00001.Contig1 0
EGC00002.Contig1 0
EGC00004.Contig1 560
EGC00005.Contig1 185
EGC00006.Contig1 506
```
which indicates that the \textit{E. granulosus} query sequence EGC00001.Contig1 had no similarity to any sequence from the turbellaria reference dataset, while the sequence EGC00004.-Contig1 had a maximum bit score of 560 and so forth (Note 10).

3.1.3. Parsing the Data into a PostgreSQL Database

Although it is not strictly necessary to build a database to hold your data (see Note 11), the use of a database greatly simplifies the ability to organize and manipulate your data. Once you have constructed a database (this could either be a new database or a database previously constructed as part of the PartiGene process described in Chapter 9), create a new table in which the first field (column) is used to store the identifier of each query sequence for which BLAST searches have been performed. Additional fields then represent the best BLAST bit scores of query sequence matches to the various reference datasets (See Table below). The following shows an example of how to set up a table for storing the results of BLAST searches of our \textit{E. granulosus} sequences performed against the three example reference datasets. First create a new database if one does not already exist:

\textit{createdb echinodb}

Enter the database programming environment:

\textit{psql echinodb}

Then create a new table to store the BLAST results:

\textit{create table blastresults (seqid varchar(20) null, turbellaria int null, trematoda int null, e_multilocularis int null);}

Extra fields can then be added to store additional BLAST data using:

\textit{alter table blastresults add column cestoda int null;}

To view the table, use the command \textit{'\d blastresults'} which will show:

\textbf{Table "public.blastresults"}

\textbf{| Column | Type | Modifiers |}

<table>
<thead>
<tr>
<th>Seqid</th>
<th>character varying (20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>turbellaria</td>
<td>integer</td>
</tr>
<tr>
<td>trematoda</td>
<td>integer</td>
</tr>
<tr>
<td>e_multilocularis</td>
<td>integer</td>
</tr>
<tr>
<td>cestoda</td>
<td>integer</td>
</tr>
</tbody>
</table>

Finally data from the flat files (Section 3.1.2) need to be imported into the database. Again a simple Perl script can be created invoking functions from the DBI Perl module. The one
#!/usr/bin/perl -w
# insert_db.pl - a simple script to insert
   parsed BLAST output
# into a specified database (here called echinodb)
# usage: perl insert_db.pl <filename> <taxon name in db>

use strict;
use DBI;
use DBD::Pg;

if(!($ARGV[0] ||$ARGV[1])){ exit(); }

my$filename=$ARGV[0];
my$taxaname=$ARGV[1];
my$database="echinodb";

my$conn=DBI->connect("dbi:Pg:dbname=$database", "","", {PrintError=> 0});

open (FILE,"$filename");
while(my$line =<FILE>)
{
   $line=~/(.+)\s(.+)/;
   my$sequence=$1;
   my$score=$2;
   if(!$score) { $score=0; }

   my$result=$conn->prepare("SELECT seqid,$taxaname
   from blastresults where seqid=' $sequence';","", {PrintError=> 0});
   $result->execute();
   my @array=$result->fetchrow_array;
   my$sequence_reference=$array[0];
   if($sequence_reference)
   {
      my$oldscore=$array[1];
      if(!$oldscore) { $oldscore=0; }
      if($score >=$oldscore)
      {
         my$update=$conn->do("update blastresults set
            $taxaname='$score' where seqid=' $sequence';","", {PrintError=> 0});
      }
print "$sequence_reference updating blastresults\n$taxaname set to $score for $sequence\n";
}
}
else
{
    my $insert = $conn->do("insert into blastresults (seqid, $taxaname) values ('$sequence', '$score');
    print "inserting $sequence into blastresults and setting $taxaname $score\n";
}
}
$conn->disconnect;

Copy the script into a file (e.g. insert_db.pl) and run it using the command:

perl insert_db.pl echino_turbellaria.out.txt turbellaria

To view contents in the table use the command ‘select * from blastresults;’ which in our example shows:

<table>
<thead>
<tr>
<th>seqid</th>
<th>turbellaria</th>
<th>trematoda</th>
<th>e_multi-localis</th>
<th>cestoda</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGC00189.Contig1</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGC00001.Contig1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGC00002.Contig1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGC00004.Contig1</td>
<td>560</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGC00005.Contig1</td>
<td>185</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGC00006.Contig1</td>
<td>506</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGC00006.Contig2</td>
<td>729</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC00007.Contig1</td>
<td>710</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other blast results can be readily inserted into the table by substituting: echino_turbellaria.out.txt with the name of any BLAST output file you wish to analyze, and turbellaria with the name of the database column that you wish to populate. Advanced users may wish to automate these steps to populate large databases containing many columns to store the results from tens or even hundreds of blast reports (Note 12).

**3.2. Analysis and Visualization of Phylogenetic Relationships**

Having constructed the basic resource, we next present two methods to readily analyze your sequence datasets in the context of phylogenetic relationships. The first method is relatively straightforward and relies on the SimiTri Java tool (12), which facilitates the visualization of sequence-similarity relationships between four
sequence datasets (your query dataset and three reference datasets – Fig. 12.1). The second method exploits phylogenetic profiles constructed from all available datasets to identify sequences with interesting evolutionary patterns.

3.2.1. Using the SimiTri Tool

The SimiTri package obtained from 2.2.6 includes a script (xfromDB.pl) to extract BLAST similarity data from a selected database and format it for use with the SimiTri tool. To run the tool to extract data from our given example, use the command:

```
perl xfromDB.pl trematoda turbellaria e_multilocularis
```

This will extract the entries from the columns trematoda, turbellaria, and e_multilocularis from the database that we set up previously. These entries are then used to calculate the relative positions of each of our query sequences according to their relative similarities to the three reference datasets. To visualize these relationships run:

```
appletviewer simitri.html
```

You will see a Java applet appear, providing an interactive graphic similar to that in Fig. 12.1. Within this graphic, each colored tile represents an individual query sequence; clicking on a tile highlights its sequence ID. In addition to revealing general trends between datasets, the SimiTri tool is especially useful for identifying sequences with specific phylogenetic properties. For example, those sequences at the center of the triangle-colored red are likely to be associated with highly conserved housekeeping genes. However, sequences closer to one axis representing a sister taxon and colored purple are likely to be recently diverged genes.

3.2.2. Generating and Analyzing Phylogenetic Profiles

While the SimiTri tool is useful for visualization, more quantitative analyses of sequence-similarity relationships require a different approach. Here we introduce the use of phylogenetic profiles to identify groups of sequences with interesting taxonomic properties. Due to the more complex nature of these analyses, we assume in the following that you have performed BLAST searches against a significant number of genomes and placed the results in an appropriate database table. An important consideration is that the reader is aware of the taxonomic relationships between each column within your database table.

3.2.2.1. Generating the Profiles

In our example we have created a database in which we have stored the results for BLAST searches of our query dataset performed against 52 genomes. We use a script – `gen_phyloprofiles.pl` (downloadable from http://www.compsysbio.org/download/papers/methods2008/gen_phyloprofiles.pl) – to query the database and extract a binary phylogenetic profile for each sequence. In this profile, a “1” represents those genomes (or partial genomes or other reference datasets) for which the query sequence has a
BLAST bit score above our defined cutoff (we use a bitscore cutoff of 50, but other users depending upon their application may prefer a more stringent cutoff of 80 or 100) and “0” represents those genomes in which the query sequence does not (potentially representing its absence from that genome/dataset). The results are stored in a file phylo_profile.dat. In our particular example with queries against 52 reference datasets, we obtain entries that look like this:

Sequence1 (37) 1101001010110101101100011111111
111110111111011011111
Sequence2 (10) 11010001000000000000010000000110
0011000000000000000001
Sequence3 (11) 1100001010000000000000000100010
001110000011000000000001
Sequence4 (19) 001010000011000000101100110
00110011011000100001
Sequence5 (9) 000000000000000000000001100011
00000010100000000011000
Sequence6 (34) 111111111101110110001110110111
01110110011110011001111
...

From this we may conclude that Sequence1 is potentially present in 37 genomes: genome 1, 2, 4, 7, 9, 11, 12, 14, and so forth. At the same time, the gen_phyloprofiles.pl allows the construction of abbreviated phylogenetic profiles which group genomes/datasets into user-defined taxonomic groups. For example, we may define genomes 1, 2, and 4 as deriving from archaeal genomes. The use of these abbreviated profiles provides a more systematic measure of sequence conservation. Typically we define sequences into the following categories: Archaea, Bacteria, Fungi, Metazoa, Protista, and Viridiplantae (see Note 13) (20). However it is up to the user to define groups appropriate to their own studies. This can be achieved through careful editing of the gen_phyloprofiles.pl script (see script for more details). The resultant file, abrv_phylo_profile.dat, contains output that for our example looks like this:

Sequence1 111001
Sequence2 111010
Sequence3 111110
Sequence4 011101
Sequence5 011100
Sequence6 111111
...

where the first entry indicates that Sequence1 is potentially present in at least one archaeal, bacterial, fungal, and plant genomes (Note the script orders the defined taxonomic groups alphabetically).
Two useful measures for analyzing the conservation of a sequence is the number of genomes in which it is found and their phylogenetic extent. These metrics are readily derived from Section 3.2.2.1 and may be used to define groups of sequences with interesting conservation properties. Three different groups of sequences are of particular interest:

3.2.2.2.1. Species-specific Genes  Sequences that are determined to be in zero genomes represent potential species-specific sequences since they appear unique to the query dataset (i.e., they do not have any homologs in any of the other sequence dataset). Typically we find that from 30 to 60% of a newly sequenced organism’s sequences fall into this category (14). These sequences represent genes that are under reduced selective constraints, provide newly acquired functionality, or are simply redundant and in the process of being lost. Many genes in these categories typically have no functional annotation.

3.2.2.2.2. Taxon-Specific Genes  Sequences with homologs in one or more reference genomes, which are all restricted to the same taxonomic group as the query sequence, indicate taxon-specific sequences. Taking as an example our Echinococcus sequences, we could define our taxonomic group at the level of cestodes, platyhelminths, or even metazoans depending upon the groups defined in Section 3.2.2.1. To identify metazoan-specific genes, we simply extract the sequences in the file abr_phylo_profile.dat with the profile “000100.” The number of genomes associated with these sequences (obtained from the file phylo_profile.dat) provides an indication of how conserved these sequences may be within the selected taxon. Previously we have found that ~8% of metazoan genes are broadly conserved across the phylum (14). In addition to ancestral genes that may have been lost in other lineages, these sequences represent relatively recently evolved genes that underlie taxon-specific innovations (for example metazoan-specific sequences might be associated with multicellular functions such as development and cell-cell communication).

3.2.2.2.3. Highly Conserved Genes  Sequences derived from highly conserved genes tend to be very well annotated and represent a useful source for performing more in-depth phylogenetic analyses (10, 11) as well as provide useful benchmarks for analyses of diversity. Two complementary approaches may be used. The first is to identify sequences with homologs in a large number of genomes (e.g., 80% of all genomes/partial genomes searched – Note 14). The second is to identify those present in a majority of the defined taxonomic groups (e.g., present in at least five of the six taxa defined in Section 3.2.2.1, i.e.,
Sequence3 111110 and Sequence6 111111). Alternatively, we might consider as highly conserved those sequences present in at least three of the six taxa provided that two of them are also Archaea and Bacteria (i.e., Sequence1 111001 and Sequence2 111010 in addition to Sequence 3 and 6). These sequences are also potentially highly conserved sequences, through the identification of homologs in all three domains of life. Previously we have found that ~20% of eukaryotic sequences may be defined as highly conserved based on this latter criteria and likely represent ancestral eukaryotic genes under significant selective constraints.

3.2.3. Profile Clustering

A more sophisticated analysis of the derived phylogenetic profiles involves the use of clustering software. Clustering profiles has proven to be extremely useful for inferring functional annotations and protein interactions (8–10). Here we describe the use of the software tools Cluster3.0 (16) and Treeview (17) to hierarchically cluster and visualize a set of phylogenetic profiles.

3.2.3.1. Clustering Phylogenetic Profiles Using Cluster3.0

First you will need to reformat the file containing the phylogenetic profiles such that each binary character that represents presence/absence from a genome is separated by a tab. In addition the first line of the file needs a tab-delimited list of each entry. For example

```
Seq_ID  1 2 3 4 5 6 7 8 9 10
Sequence1 1 1 0 1 0 0 1 0 1 0
Sequence2 1 1 0 1 0 0 0 0 1 0
Sequence3 1 1 0 0 0 1 0 1 0 0
Sequence4 0 0 1 0 1 0 0 0 0 0
Sequence5 0 0 0 0 0 0 0 0 0 0
Sequence6 1 1 1 1 1 1 1 1 1 1
```

A file (phylo_profile_4cluster.txt) with this correct format is automatically generated by gen_phyloprofiles.pl (Section 3.2.2.1). Cluster3.0 is available for Microsoft Windows, Linux, and MacOS (see Cluster3.0 documentation for more information on installation and usage.) Launch the Cluster3.0 software using the appropriate command and load in your tab-delimited file (e.g., phylo_profile_4cluster.txt) by using the option “File -> Open data file” and selecting the appropriate file. Next click on the “Hierarchical” tab in the Cluster3.0 window. Check the “Cluster” box and select “Correlation (centered)” from the drop down menu under the section “Genes”. Click the “Single linkage” box under the section
“Clustering Method” (Note 15). The new file generated after clustering will be automatically created with the extension “.cdt” by the Cluster3.0 software.

3.2.3.2. Visualization of the Clusters

TreeView is a package that facilitates the visualization of cluster data generated by the Cluster3.0 software and is also available for Microsoft Windows, Linux, and MacOS. Again refer to the TreeView documentation for installation and usage. Launch the TreeView software and use the menu option “File -> Open” to select and import the .cdt file produced by Cluster3.0. Once the clustered profiles have been imported, you may wish to alter the colors by opening the “Settings” -> “Pixel Settings” menu. Use the “Contrast” bar to provide a clearer contrast between presence and absence of a sequence in a genome. See Fig. 12.2 for an example of hierarchical clustering of phylogenetic profiles. From these views, clusters of sequences sharing similar

Fig. 12.2. Hierarchical clustering of phylogenetic profiles. Hierarchical clustering was performed using cluster3.0 (16) and visualized using Treeview (17). Each row in the image represents the phylogenetic profile for an individual sequence of the partial genome *Zeldia punctata*. Each column represents the pattern of presence/absence to a battery of 192 other partial genomes. We use partial genomes in this example to illustrate their utility – columns could also readily represent complete genomes or other sequence datasets (e.g., groups of sequences from a defined taxon). Columns were sorted by phylogenetic relationships (not shown for simplicity). The top (white background), middle (dark background), and bottom (scattered background) areas of the graphic indicate regions of species-specific, highly, and intermediately conserved sequences, respectively.
profiles may be readily identified. Annotation associated with sequences within these clusters may reflect common functionality.

4. Notes

1. Access to adequate computational infrastructure can be a limiting factor in large-scale phylogenomic analyses. It is possible to perform BLAST comparisons of a modest dataset of ~3,000 sequences against 20 datasets of similar size on a desktop workstation in approximately a day. If the reader wishes to search many more and/or larger reference datasets, they may wish to locate a computing cluster on which several large-scale BLAST searches may be performed in parallel. For our purposes we use a local computing cluster located at the Centre for Computational Biology (Hospital for Sick Children, Toronto, Ontario, Canada).

2. The precise choice of sequence datasets is dependent upon the question being asked. If one wishes to use the SimiTri tool to examine sequence relationships of your target sequences to three reference datasets, then you may wish to download three sequence datasets from species closely related to your own (see Section 2.3.1). Alternatively, if you wish to derive detailed phylogenetic profiles, then you should consider most of the 100 or so complete eukaryotic genome datasets currently available in addition to a handful of representative prokaryotic datasets (see Note 6).

3. To remove sequences belonging to a specific species from a larger group, you can use the NCBI Entrez tool. For example, to compare our echinococcus dataset against other cestodes, we could exclude the echinococcus parent group – Taeniidae – from our set of cestode sequences. To do this we first find the taxonomy IDs of the respective groups (Taeniidae and Cestoda) by typing the name of each group into the NCBI taxonomy browser and click “Go.” Next click on the taxonomic group name. In our example, we find Taeniidae has a taxonomy ID of 6208, while Cestoda has a taxonomy ID of 6199. To download cestode sequences while excluding Taeniidae sequences, you can then enter the following into the Entrez search panel: \texttt{txid6199[Organism:exp]} \texttt{NOT txid6208[Organism:exp]}. 
4. In this study, we used a filtered subset of the 335 complete genomes available in the GOLD database (see Section 3.1) on January 17, 2006, that is, only one strain was used per species. This was done in order to avoid taxonomic redundancy. This resulted in the download of 198 complete genomes (19 Archaea, 129 Bacteria, and 34 Eukarya). The user may want to download an updated and/or smaller number of complete genomes to reduce the number of BLAST searches that need to be performed. This reduction should be undertaken with the goal of maximizing taxonomic coverage. Species can be assigned to specific taxonomic groups on the basis of the information derived from the National Center for Biotechnology Information’s (NCBI) TaxBrowser resource (21).

5. Alternatively, the user may want to use other biological databases to retrieve these sequence data, such as COGENT (http://cgg.ebi.ac.uk/services/cogent/) (22) or NCBI (http://ftp.ncbi.nih.gov/genomes/) (21).

6. During the generation of phylogenetic profiles, the evolutionary relationships of the organisms compared against must be taken into account. To date there are over 600 genomes for prokaryotes. In some cases, 3 or 4 genomes may have been sampled from the same species. In other cases, there may have been dense genomic sampling within a relatively narrow taxon. Since inclusion of many closely related datasets can greatly influence profile clustering (23), we recommend the user includes sample genomes from as diverse a set of prokaryotic taxa as possible.

7. For the purpose of the clustering analysis, we used a subset of 192 partial genomes obtained from the PartiGene database (http://www.compsysbio.org/partigene) on January 17, 2006. The user may want to download an updated number of partial genomes in order to increase taxonomic coverage in Eukarya. The PartiGene database provides an interactive page in which users can select and download individual species or larger taxonomic groups of partial genomes in FASTA format, which may be readily transformed into BLAST databases.

8. BLAST performs searches based on local alignments as opposed to global alignments. Thus, while it can rapidly identify sequence matches, such matches may reflect only local regions of sequence similarity such as protein domains that may not be indicative of true homology relationships. For a more accurate search tool, the advanced user could consider using BLAST as a prefiltering step to identify a subset of
candidate sequences, which can be more accurately analyzed using a global alignment algorithm such as Needleman-Wunsch (24).

9. The use of the BLAST option “–e 0.001” specifies a BLAST E-value threshold of $10^{-3}$. We choose this threshold over a less stringent threshold to reduce the potential size of the BLAST output files. This initial E-value threshold may be complemented in later steps by applying an additional BLAST bit score threshold of 50, which was found to be appropriate for the analysis of sequence similarity across genomes.

10. Rather than using BLAST scores, the user could also consider using metrics such as expectation values or % sequence identity, which may be readily obtained by modification of the extract_blasts.pl script.

11. Instead of a PostgreSQL database, the user may want to create an MySQL database (http://www.mysql.com/) or even a flat file. We advise the use of an SQL database because it allows the storage of data in a flexible and readily accessible format that facilitates future detailed analyses.

12. Populating a table with many hundreds of results could be achieved through modification of the insert_db.pl program through the specification of a directory holding all the BLAST output. Columns can be dynamically added using the alter table command in Section 3.1.3 and referring to the name of the BLAST output file for the column title.

13. The taxonomic categories described correspond to the three domains of life: Archaea, Bacteria, and Eukarya, with the eukaryotic domain divided into Protista, Fungi, Metazoa, and Viridiplantae. Depending upon the question being addressed, different taxonomic classifications might be more appropriate. For example, if one wishes to compare a dataset of insect ESTs, one could consider taxonomic groups such as other insects, Crustacea, Myriapoda, Chelicerata, and Panarthropoda in addition to the categories mentioned previously, to provide more information on deeper evolutionary relationships.

14. Note that the criteria are flexible but should take into account the presence of potential homologs in prokaryotic as well as eukaryotic datasets.

15. Alternatively, the user may want to try other clustering methods such as “Centroid linkage” in order to identify the most clearly delineated groups of sequences.
Acknowledgments

This work was supported by the Hospital for Sick Children (Toronto, Ontario, Canada) Research Training Center. J.P is supported by a Canadian Institute for Health Research New Investigators Award and an Early Researchers Award from the Ontario Ministry of Research and Innovation.

References


Statistical Analysis of Expressed Sequence Tags

Edward Susko and Andrew J. Roger

Abstract

Expressed sequence tag (EST) surveys are an efficient way to characterize large numbers of genes from an organism. The rate of gene discovery in an EST survey depends on the degree of redundancy of the cDNA libraries from which sequences are obtained. We consider statistics for the comparison of EST libraries based upon the frequencies with which genes occur in subsamples of reads. These measures are useful in determining which of the libraries, having a large proportion of genes in common, is more likely to yield new genes in future reads. We also present tests, with multiple corrections adjustments, for whether genes are equally represented or expressed in a pair of libraries.

Key words: Expressed sequence tags, coverage, gene discovery, redundancy measures, cDNA library comparison, tests for expression.

1. Introduction

In many cases, the redundancy of highly expressed transcripts makes it necessary to perform expensive normalization protocols on cDNA libraries before large numbers of ESTs are gathered from an organism. As sampling of a library continues, the probability of sampling previously identified genes increases. Redundancy checks throughout an EST project can be used to determine when it is no longer economically profitable to keep sequencing from the same library. The measures of redundancy considered here, coverage and the expected number of reads to discover a new gene, are extensions of measures used in the problem of quantifying species frequencies.

EST surveys are not only used for gene discovery, but are often conducted to evaluate differences in gene expression in
different tissues or cells exposed to different conditions. For example, one might be interested in the expression of a particular gene in a cancerous cell relative to a normal one. Microarray experiments provide an effective way of addressing this question, but EST surveys allow for comparison as well. Where microarray experiments quantify expression through continuous fluorescence measures, EST surveys yield counts of the number of reads corresponding to a gene. If a particular gene is expressed twice as frequently in cancerous cells, a comparable ratio is expected of the numbers of reads corresponding to that gene from libraries of both cancerous and normal cells. For a given gene, conditioning upon the total number of reads in all libraries, we present appropriate binomial test statistics for detecting differences in expression. Tests and methods for examining redundancy are illustrated through application to normalized and nonnormalized cDNA libraries from the amitochondriate protist *Mastigamoeba balamuthi*. Additional information is available in Ref. (1).

2. Methods

In the following, we describe a number of redundancy measures. Single sample coverage (2, 3, 4) and expected reads for gene discovery are relevant redundancy measures. Standard errors (5) and confidence intervals for these measures can be useful in comparing libraries. Multiple library coverage is a measure of redundancy in a given library from the reads taken from any of the libraries. It provides a redundancy measure taking into account appearance of genes in other libraries.

A good statistical test should satisfy that the probability of rejecting a hypothesis is always larger when that hypothesis is false than when it is true; tests that satisfy such a property are referred to as unbiased. If, in addition, a test has greater probability of rejecting a false hypothesis than any other unbiased test, it is referred to as uniformly most powerful and unbiased. Such tests are broadly considered to be optimal (6). For a given gene, uniformly most powerful and unbiased statistical tests can be used to test for differences in the expression or proportional representation of a gene. When a large number of genes are tested, however, the overall probability that a false positive will be found becomes large unless a multiple comparisons adjustment (7) is made.
Coverage is defined as the proportion of the library that appears in the sample of the reads. For instance, if two genes arise with one constituting 10% of the library and the other 1% of the library, the coverage would be 11%. If coverage is large it suggests that new gene discovery is unlikely. An equivalent but perhaps more easily interpretable measure of redundancy is the expected number of reads required to discover a new gene.

To obtain the coverage, first obtain the number, \( n_1 \), of EST clusters consisting of a single sequence and the total number, \( n \), of sequences. For the nonnormalized *Mastigamoeba* data in Table 13.1, the number of clusters consisting of a single sequence was 378 and the total number of sequences was 715.

The (estimated) coverage is calculated as 1 minus the proportion of genes that appeared in a cluster consisting of a single sequence:

\[
\hat{C} = 1 - \frac{n_1}{n}.
\]

For the nonnormalized *Mastigamoeba* data, we obtain a coverage of \( 1 - \frac{378}{715} = 0.471 \).

### Table 13.1
The numbers of clusters \( (n_k) \) of sequences that were read \( k \) times from nonnormalized and normalized libraries for the amitochondriate protist *Mastigamoeba*

<table>
<thead>
<tr>
<th>( k )</th>
<th>Nonnormalized ( n_k )</th>
<th>Normalized ( n_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>378</td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
The expected number of reads to discover a new gene is estimated as:

\[ \hat{E} = \frac{1}{1 - \hat{C}} \]

Where \( \hat{C} \) is the coverage. For the nonnormalized *Mastigamoeba* data, we therefore estimate that \( \frac{1}{1 - 0.471} = 1.89 \) reads will be required to discover a new gene.

### 2.2. Standard Errors and Confidence Intervals

Since the actual coverage of a library is not directly observable (without an exhaustive survey of the library, we do not know, for example, if one particular gene accounts for 10% of the entire library), the estimate of coverage is subject to sampling error. Comparison of two libraries thus requires that standard errors of coverage be taken into account.

The true coverage of the library is the sum of the proportional representations of the genes that appeared in the sample. Since the proportional representations are unknown, \( \hat{C} \) is an estimate. Standard errors for coverage can be calculated through the formula derived in (5):

\[
se(\hat{C}) = n^{-1/2}\left[\left(\frac{n_1}{n}\right) + \left(\frac{2n_2}{n}\right) - \left(\frac{n_1}{n}\right)^2\right]^{1/2}
\]

where \( n \) is the total number of sequences, \( n_1 \) is the number of clusters consisting of a single sequence, and \( n_2 \) is the number of clusters consisting of 2 sequences. For the nonnormalized *Mastigamoeba* data, we obtain:

\[
se(\hat{C}) = (715)^{-1/2}\left[\left(\frac{378}{715}\right) + \left(\frac{2.33}{715}\right) - \left(\frac{378}{715}\right)^2\right]^{1/2} = 0.022
\]

The standard error for coverage can be converted to a standard error for the expected number of reads required to discover a new gene through the transformation:

\[
se(\hat{E}) = se(\hat{C})/(1 - \hat{C}).
\]

For the nonnormalized *Mastigamoeba* data, the expected number of reads to discover a new gene was 1.89, the coverage was 0.471, and the standard error for the coverage was 0.022. Thus the standard error for the expected number of reads to discover a new gene is \( 0.022 / (1 - 0.471) = 0.0416 \).

Approximate 95% confidence intervals can be calculated using the large-sample normal results as the estimate, plus and minus 1.96 times the standard error:

\[
\hat{C} \pm 1.96 \ se(\hat{C})
\]

\[
\hat{E} \pm 1.96 \ se(\hat{E})
\]
For the nonnormalized *Mastigamoeba* data, the coverage was 0.471 with standard error 0.022, so a 95% confidence interval for the coverage is $0.471 \pm 1.96 \cdot 0.022 = [0.428, 0.514]$. The estimate of the expected number of reads to discover a new gene was 1.89 with standard error 0.0416 so that a 95% confidence interval is given by $1.89 \pm 1.96 \cdot 0.0416 = [1.81, 1.97]$. Similar calculations for the normalized library gave an estimate of 1.82 and confidence interval (1.70, 1.93). While the nonnormalized library seems more likely to lead to new gene discovery, the overlap of the confidence intervals suggests that this can be explained by sampling variation.

2.3. Multiple Sample Coverage and Expected Reads for Gene Discovery

Because of the overlap of libraries, the expected reads for gene discovery from a given library may be small even though it is unlikely those new genes would not have been discovered in at least one of the other libraries. Multiple sample coverage is the coverage of a given library by any of the libraries considered.

To calculate the coverage of a given library, say library 1, from any of the libraries sampled, determine the number of genes that were read as a cluster of a single sequence in library 1, but did not occur at all in any of the other libraries. Divide this by the total number of genes in library 1 and subtract the result from 1:

$$\hat{C}_1 = 1 - \frac{n_{10...0}}{n_1}$$

where $n_{10...0}$ denotes the number of genes that appeared one time in library 1, but not at all in any of the other libraries and $n_1$ denotes the total number of genes in library 1. For the *Mastigamoeba* data, ESTs were available for a normalized library in addition to the nonnormalized library. Among the 378 genes that appeared in clusters of a single sequence in the nonnormalized library, 375 did not appear in the normalized library. Thus the two-sample coverage of the nonnormalized library was $1 - \frac{375}{715} = 0.476$. The coverage based on the nonnormalized library alone was 0.471. Here we see that the coverage of the nonnormalized library is only slightly higher when data from both normalized and nonnormalized libraries are considered. Thus there does not seem to be a substantial effect from normalization.

To obtain the estimate, $\hat{E}$, of the expected number of reads from library 1 required to discover a new gene, given the samples from all of the libraries, calculate:

$$\hat{E} = \frac{1}{(1 - \hat{C})}$$

as in the single sample case but where now $\hat{C}$ is the two-sample coverage for that library. For the *Mastigamoeba* data, taking into account both libraries, the coverage of the nonnormalized library
was 0.476. Hence the number of reads that need to be sampled from the nonnormalized library in order to identify a gene previously unidentified in either library is $1/(1 - 0.476) = 1.91$.  

### 2.4. Multiple-Sample Coverage: Standard Errors and Confidence Intervals

As with single-sample coverage, multiple-sample coverage is subject to sampling error. Comparison of two libraries thus requires that standard errors of coverage be taken into account.

In addition to the number of genes, $n_{10\ldots0}$, that appeared in clusters of a single sequence in library 1 and not at all in any of the other libraries, obtain the number of genes, $n_{20\ldots0}$, that appeared in clusters of two sequences and not at all in any of the other libraries. For the *Mastigamoeba* data, $n_{10\ldots0} = 375$, and out of the 33 genes that appeared in clusters of two sequences in the nonnormalized library, $n_{20\ldots0} = 21$ did not appear in the normalized library.

Calculate the standard error through the formula:

$$se(\hat{C}) = n^{-1/2}[(n_{10\ldots0}/n_1) + (2n_{20\ldots0}/n) - (n_{10\ldots0}/n)^2]^{1/2}$$

where $n$ is the total number of sequences in library 1. For the *Mastigamoeba* data, we obtain:

$$(715)^{-1/2}[(375/715) + (2\cdot21/715) - (375/715)^2]^{1/2} = 0.0208$$

As in the single-library case, the standard error for coverage can be converted to a standard error for the expected number of reads required to discover a new gene through the transformation:

$$se(\hat{E}) = se(\hat{C})/(1 - \hat{C}).$$

For the *Mastigamoeba* data, the coverage is 0.476 and the standard error of the coverage is 0.0208, so the standard error for the expected number of reads from library 1 required to discover a new gene is $0.0208/(1 - 0.476) = 0.0397$.

Approximate 95% confidence intervals can be calculated using large-sample normal results as the estimate, plus and minus 1.96 times the standard error:

$$\hat{C} \pm 1.96 \, se(\hat{C})$$

$$\hat{E} \pm 1.96 \, se(\hat{E})$$

For the *Mastigamoeba* data, the confidence interval is $0.476 \pm 1.96 \cdot 0.0208 = [0.435, 0.517]$. The confidence interval for the expected number of reads from library 1 required to discover a new gene is $1.91 \pm 1.96 \cdot 0.0397 = [1.83, 1.98]$.

### 2.5. Tests for Differences in Expression for a Gene

In many EST surveys, a primary goal is to detect differences in gene expression. Given a gene of interest, the data available is the numbers of times, $x_i$, that the gene was encountered in reads
from library $i, i = 1, 2$. A test with desirable statistical properties – uniformly most powerful unbiased (6) – rejects the hypothesis of equal expression when $x_1$ is larger than or smaller than expected under that hypothesis.

For a given gene of interest, obtain the sizes of the EST clusters that the gene was present in, for the two libraries. For the *Mastigamoeba* data, one of the genes appeared in a cluster of size $x_1 = 15$ in the nonnormalized library and $x_2 = 14$ times in the normalized library.

Determine $p$, the number of reads from the first library divided by the total number of reads from both libraries. For the *Mastigamoeba* data, 715 reads were obtained from the nonnormalized library and 363 from the normalized library. Thus $p = 715 / (363 + 715) = 0.663$.

Determine the probability, $p_L$, that a binomial random variable (see Note 1) with $n = x_1 + x_2$ and $p$ the proportion of reads from library 1, is less than or equal to $x_1$:

$$p_L = P(X \leq x_1; n, p) = \sum_{x=0}^{x_2} \binom{n}{x} p^x (1 - p)^{n-x}$$

Standard statistical packages and tables can be used to calculate this probability, and software is available; see Note 2. For the *Mastigamoeba* example comparison, the probability that a binomial random variable with $n = 14 + 15 = 29$ and $p = 0.663$ is less than or equal to 15 is 0.0736.

Determine $p_G$, the probability that a binomial random variable with $n = x_1 + x_2$ and $p$ the proportion of reads from library 1, is greater than or equal to $x_1 - 1$:

$$p_G = P(X \geq x_1 - 1; n, p)$$

For the *Mastigamoeba* example comparison, the probability that a binomial random variable with $n = 14 + 15 = 29$ and $p = 0.663$ is greater than or equal to 15 - 1 = 14 is 0.966.

Calculate the $P$-value for the null hypothesis that proportional representation of the given gene is the same in the two libraries as (see Note 3) twice the smaller of the probabilities from the previous two steps:

$$P\text{-value} = 2 \min(p_L, p_G)$$

In the *Mastigamoeba* example comparison, the probabilities were 0.0736 and 0.966 so that the $P$-value is $2 \times 0.0736 = 0.147$.

### 2.6. Adjustments for Multiple Comparisons

The probability of a false positive for comparison of expression for a single gene is $\alpha$. When large numbers of comparisons are considered however, the probability of at least one false positive is much larger. Adjustments to the thresholds for rejection are
required in order to control this experiment-wise probability of a false positive at $\alpha$.

Obtain the $P$-values for equality of expression or, equivalently, equality of proportional representation – that the frequency with which the gene occurs in library 1 is the same as the frequency with which the gene occurs in library 2. Order the distinct $P$-values from smallest to largest: $p(1) < p(2) \ldots$. Because of the discrete nature of the data, it is possible that more than one gene will give the same $P$-value. Let $N_j$ denote the number of genes that gave $P$-value $p(j)$. For the *Mastigamoeba* data, the seven smallest distinct $P$-values are given in Table 13.2.

The equality of proportional representation for the gene with the $i$th ranked $P$-value $p(i)$ is rejected with experiment-wise false positive rate less than $\alpha$ if:

$$p(i) < \alpha \sum_{j=1}^{i} \frac{N_j}{N}$$

where $N$ is the total number of $P$-values that have been calculated.

For the *Mastigamoeba* data from Table 13.2, we see that the smallest $P$-value is 0.002. A total of 663 distinct genes had tests of proportional representation for this data (Table 13.2 indicates only a few of the smallest $P$-values). Since $0.002 > 7.54 \times 10^{-5} = 0.05 \cdot 1/663$, taking into account the multiple comparisons, the

<table>
<thead>
<tr>
<th>$N$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>6</td>
<td>0.002</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>9</td>
<td>0.008</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>7</td>
<td>0.017</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0.025</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>3</td>
<td>0.076</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>0</td>
<td>0.113</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>14</td>
<td>0.147</td>
</tr>
</tbody>
</table>
hypothesis of equality of proportional representation cannot be rejected for this gene. Considering now the gene that had occurred in a cluster of 15 sequences in the nonnormalized library and a cluster of 14 in the normalized library (last row in Table 13.2) we see that:

\[
\sum_{j=1}^{i} N_j = 1 + 1 + 1 + 2 + 8 + 1 + 1 = 15
\]

The \( P \)-value 0.147 is larger than \( 0.05 \cdot 15/663 = 0.001 \) and taking into account the multiple comparisons, the hypothesis of equality of proportional representation cannot be rejected for this gene either. Indeed, because of the large number of comparisons considered, there were no rejections of the hypothesis of proportional representations for the \textit{Mastigamoeba} data. Adjusting for the large number of comparisons, there is no significant evidence of differences in expression.

3. Notes

1. A binomial random variable, \( X \), with parameters \( n \) and \( p \) satisfies that: \( P(X = x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x} \). A common example where this distribution arises is when \( X \) is the number of heads in \( n \) independent coin flips, each of which has probability \( p \) of yielding a head.

2. Software for the methods discussed is available at \texttt{http://www.mathstat.dal.ca/~tsusko}

3. The test presented here rejects the hypothesis of equal expression at the \( \alpha \)-level of significance when \( x_1 \) is smaller than the \( \alpha/2 \)th quantile or larger than the \( (1 - \alpha/2) \)th quantile of a binomial distribution with parameters \( n \) and \( p \) \((1)\). Generally, the \( P \)-value for a test can be defined as the smallest \( \alpha \)-level for which the test rejects. This turns out to be \( 2 \min(p_L, p_G) \) for this test.

4. The \textit{Mastigamoeba balamuthi} (GenBank taxonomy ID: 108607) data were derived from the Genome Canada Protist EST Program \texttt{http://www.bch.umontreal.ca/pepdb/pep_main.html}

Much more data than what is summarized in Table 13.1 are now available. As of the current writing, there are 20,111 reads and they have all been deposited in the GenBank EST database.
5. An alternative measure of multiple-library coverage, with a different and potentially useful interpretation, is the coverage of library 1 by the sample from library 2 (8, 9). It is estimated as 1 minus the proportion of genes in the sample from library 1 that were not found in library 2.

6. An additional statistical measure that can be useful in determining follow-up sampling effort is the expected number of new genes that will be discovered in new samples of varying size (1, 3, 10, 11, 12). These methods are can be divided into classes as parametric and nonparametric. Parametric methods (1, 10) give stable estimation but make assumptions about the manner in which the proportional representations of genes are distributed in the library. Nonparametric methods (1, 3, 11, 12) make no such assumptions but suffer from instability problems when numbers are required for new samples of size exceeding the original sample size.

7. Here and in our earlier work (1), we have treated the EST clusters as known and have implicitly assumed that errors arising from EST-clustering procedures are negligible. If uniformly full-length cDNA libraries are available and long, high-quality sequence reads from the 5’ end of the clones are obtained, then this is a reasonable assumption. However, in reality, truncated cDNAs and short or poor-quality reads can lead to increased clustering errors where ESTs corresponding to the same gene fail to cluster together and the numbers of smaller clusters and therefore the number of unique genes are inflated. Alternatively, ESTs from extremely similar duplicate genes (paralogs) may be artificially clustered together, causing inflation of cluster sizes and a decrease in the number of unique genes. Two recent studies (12, 13) discuss these issues in detail and propose a statistical approach to correcting these errors.

8. Alternative tests for comparing expression (14) make assumptions about the manner in which the proportional representations of genes are distributed in the libraries. Our experience, however, has been that the tests give similar results to the ones given here.

References


Chapter 14

Application of ESTs in Microarray Analysis

Weizhong Li, Lisa Olohan, Daryl Williams, Margaret Hughes, Andrew Gracey, and Andrew Cossins

Abstract

Microarray analyses provide information on the relative expression levels of large numbers of gene products (transcripts). As such they have been widely used to examine differences in gene expression across a variety of samples such as tissues and life-cycle stages. Due to a previous lack of sequence data, microarray analyses have typically centred on the study of well-characterised model organisms. However, the recent availability of large sets of expressed sequence tags (ESTs) generated for the purpose of gene discovery offers the opportunity to consider designing and applying microarray technology to a larger and more diverse set of species. Here we outline the array-design process involving the generation of an optimised set of oligoprobes from a minimally redundant but maximally representative list of sequences from raw EST data. We illustrate these principles by showing how we designed and fabricated a high-density oligoarray for the rainbow trout, a non-model species for which large numbers of ESTs, and a non-redundant assembly is available. This approach brings array technology within the reach of all investigators, even those with limited budgets.

Key words: Microarray, oligoprobe, expression profiling.

1. Introduction

Microarray screening of transcripts has radically altered the research landscape by enabling the simultaneous determination of potentially all primary gene products (transcripts) within a highly complex mixture. The critical and active element of the microarray is the DNA probe (see Note 1 on terminology), which in contrast to the liquid-phase probe used in Northern analysis is immobilised directly onto the solid phase in a 2D array format – one probe for each position in an array that may extend to millions of separate probes. Modern standards allow the fabrication of arrays containing many
tens or even hundreds of thousands of probes on a glass microscope slide, and the Affymetrix “Genechip” platform allows several millions of probes on a small silicon wafer.

While the earliest microarrays were typically based on amplitcon probes generated by PCR of cDNA inserts or by the targeted amplification of genomic fragments, the most dense and elegant arrays contain short (25–30 mer) or long (65–70 mer, see Note 2) oligonucleotide probes designed from the available cDNA sequence. Oligoarrays are increasingly preferred since they offer greater hybridisation specificity than cDNA or PCR products, can distinguish single-nucleotide polymorphisms and can be directed at sub-regions of a gene, such as for exons, exon boundaries or variant-specific regions (1). Both oligo and cDNA arrays ultimately depend upon the increasing availability of expressed sequence tag (EST) data. The genomic model species, such as C. elegans, Drosophila, mouse and humans have very extensive EST resources. However, there are also rapid increases in the EST coverage of an increasing number of other species, which are not genomically sequenced so designing oligoarrays even for these so-called “non-model” species is fast becoming the preferred route to array application.

A key requirement in microarray design is the need to maximise gene representation. The principal limitations are (i) the availability of appropriate resources for providing probe DNA (i.e. cDNA clones or sequence data) and (ii) the number of individual features that can be included on the surface of the array. Sequence lists of genes being targeted for analysis (i.e. targets) can be compiled by analysis of genomic sequence data, and verified from the detailed analysis of ESTs or full-length cDNA sequences. When sequence data are unavailable, there is no alternative to generating, de novo, a collection of cDNA clones together with the determination of the corresponding DNA sequence. This is now an achievable aim even for small labs but does require some skill in its execution; see Refs. (2, 3) for good examples. The protocols for normalisation, suppression subtractive hybridisation and serial subtraction (4–7) and for generating full-length cDNA clones are described in detail elsewhere in this book.

Regarding the number of probes, early array platforms used in academic labs could print 10,000 to 20,000 features per slide, more rarely up to 40,000. These remain a cost-effective route to small-scale array production particularly for arrays containing a small number of probes, or for multiple small arrays printed on the same slide. However, commercially sourced platforms now provide a much greater capacity for probes and this capacity is rapidly increasing as production methods evolve and mature (see Note 3). In practical terms, this allows all target sequences of interest to be represented on the array, but it also allows multiple large arrays (i.e. 44 K features) to be printed on each slide, each of
which can be separately hybridised to a different target. In the latter case and also for designs employing multiple probes per target sequence or large numbers of controls, the scale of the available sequence data generally outstrips the capacity of each array, and it is therefore necessary to be selective of the sequences represented there.

Finally, modern array production methods using on-chip synthesis are increasingly rapid and flexible. Several manufacturers now undertake small print runs of customised designs at modest cost and over a 2–4 week production cycle. This together with readily available sequence data offers a very flexible and responsive approach to array design, and this applies equally to non-model species for which only modest EST data exist (8). We describe here our in-house protocols used to design an oligoarray for the rainbow trout for which 227 K ESTs were publicly available. These sequences were informatically filtered down to 21.5 K selected probes using a series of straightforward protocols, this matching the then capacity of the Agilent platform. We describe procedures to (i) generate a minimally redundant but maximally representative list of sequences from the raw EST data, (ii) reduce the list of sequences to match the capacity of the array platform being used, (iii) design an optimised set of oligoprobes and finally (iv) perform a hybridisation experiment using an oligoarray. Whilst oligoprobe design can be much more sophisticated, and gene filtering and probe validation routines can be much more extensive (8–11), the protocols described here are suitable for relatively non-specialised labs to generate excellent results.

2. Materials

2.1. Hardware and Software Requirements

2.1.1. Database Servers

The minimum requirements are two servers, one that hosts the live version of databases available to the end users, and another computer configured as similarly as possible, on which to maintain a copy of the database for development purposes. The second server is also used as a backup should the live server fail for any reason.

2.1.2. Database Management System

PostgreSQL and MySQL are relational data-management systems, which are freely available at http://www.postgresql.org/ and http://www.mysql.com/, respectively. Both have ample power and flexibility to address all needs of a suitable EST database system.

2.1.3. Basecalling and Assembly

The phred-crossmatch-phrap package for EST electropherogram processing/assembly is freely distributed (http://www.phrap.org/).
2.1.4. Assembly and Reporting of ESTs

We used EST-Ferret, a suite of open-source routines that has been integrated by a series of PERL scripts (http://legr.liv.ac.uk). It automates all of the processes involved in assembly, BLAST homology searching, generation of gene ontology and other kinds of functional annotation. A series of files is generated that can be used in the construction of websites, relational databases and other reports.

2.1.5. Installing a Capability for BLAST Homology Searching

Download the BLAST programme and databases from the NCBI BLAST website (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml). Installation is described in the accompanying instructions (but also see Note 4). BLAST analysis of large-scale sequence datasets can be very time consuming, and Note 5 describes the advantages of using a PC cluster to help minimise this. RefSeq, and the non-redundant (nr) databases should be downloaded from the NCBI FTP site: http://ftp.ncbi.nlm.nih.gov/blast/db/. SwissProt should be downloaded from EBI http://ftp.ebi.ac.uk/pub/databases/swissprot/. Additional sequence databases are available on the NCBI FTP site.

2.2. Source of Sequence Data

2.2.1. Pre-assembled EST Collections

Since they are randomly selected, EST collections contain many repeated sequences. To minimise this redundancy, we suggest that the user attempts to locate a source of pre-assembled sequences (see Note 6), where the redundancy has been completely removed. The Gene Index Project maintains a large number of collections from a diverse range of species, including the rainbow trout.

First, using rainbow trout as an example, go to the Gene Index Project site at http://compbio.dfci.harvard.edu/tgi/tgipage.html. Select the rainbow trout logo and then click on the “DownloadRtGi” button. This downloads 57 K sequences (as of May 2006, but 83,863 in August 2007) in FASTA format to your local PC; this comprises the list of non-redundant contigs and singletons that were clustered from 227 K ESTs.

2.2.2. Assembly of ESTs

Where EST data exist only in the form of raw, unassembled ESTs, as from dbEST (http://www.ncbi.nlm.nih.gov/dbEST), or when the user has produced their own EST collection, then we suggest that sequence redundancy is minimised by adopting assembly and clustering procedures (see Note 7) such as those described in Chapter 9 of this volume. Where the sequence list needs to be trimmed back to match array capacity, then we recommend that the assembled contigs and singletons should be subjected to BLAST searches (see Section 3.1.1). In addition to identifying the sense direction of each gene, such analyses also indicate which sequences can be linked to a biological process or molecular function. Sequences lacking this annotation tend to be less useful in the functional assessment of transcript-profiling experiments and thus are prime candidates for discarding. This
carries the risk that potentially important “novel” sequences lacking either identity or function are removed from the array. However, establishing which of these sequences within a large collection of otherwise unidentified sequences are worthy of analysis can generally only be achieved after the results of the array experiment become known.

To identify meaningful sequences, we suggest that you perform BLAST searches against the following databases: (1) mRNA sequences that are already available for your species of interest; (2) a collection of full-length cDNA sequences from the closest model organism to your selected species (for rainbow trout, we use the zebrafish collection); (3) RefSeq, a collection of well-annotated genes; (4) Swiss-Prot, a collection of well-annotated proteins and (5) the non-redundant protein database, UniProt. **Section 2.1.5** provides details of how these databases can be downloaded.

### 2.2.3. mRNA Collections

As mRNAs from the GenBank are not included in dbEST or the Gene Index Project, it is necessary to retrieve them from GenBank separately. Thus, to retrieve rainbow trout mRNA sequences, go to Genbank (http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi) and select the Entrez tool (12, 13). Type the keywords “rainbow trout, mRNA, not mitochondrial, not EST” in the text box of “Search across databases”, then click “GO” to search sequences. Entrez then outputs a summary for lists of the resulting sequences. Click on “Nucleotide: sequence database,” which leads to a listing of the sequence records. Click the hyperlinked number adjacent to “CoreNucleotide records” to be taken to the list of sequence records. Then from the pull-down menus, select “Display” as FASTA and “Send to” as File. This results in the FASTA-formatted sequences being downloaded directly to your local computer.

### 2.2.4. Full-Length cDNA Collections for Mouse and Zebrafish

Download the full-length cDNA collections for one or more of zebrafish (14), mouse or *Xenopus*, all of which are available at http://mgc.nci.nih.gov/. The resulting files are provided in FASTA format. If full-length cDNA sequence data are not available, EST assemblies from other closely related species may also serve as alternative scaffolds for the co-alignment of ESTs to be described in **Section 3.2**.

### 2.3. Hybridisation Experiments

1. **Synthesis of aminoallyl-labelled first-strand cDNA**: Molecular biology grade water, oligo(dT) primer (5’-T₂₀VN-3’; 100 μM; custom order, Sigma Genosys), 20 × aminoallyl-dUTP/dNTP mixture [10 mM each of dATP, dGTP and dCTP (Bioline), and 5 mM each of dTTP (Bioline) and aminoallyl-dUTP (Sigma)], 5 × first-strand buffer, 0.1 M DTT, SuperScript III (200 units/μl; Invitrogen), stop
solution (0.45 M EDTA and 1 N NaOH; prepared in-house),
normalisation solution (0.75 M HEPES, pH 7.4 and 0.75 M
CH₃COONa, pH 5.2; prepared in-house).

2. **Purification of first-strand cDNA**: NucleoSpin Extract II Kit
   (Abgene), molecular biology grade water (pH 7.0–8.5), 80%
   (v/v) ethanol prepared using 200 proof ethanol (Sigma).

3. **Coupling of aminoallyl-labelled first-strand cDNA with Cy
dye**: 0.1 M sodium bicarbonate (for molecular biology,
Sigma), Cy3 and Cy5 mono-reactive dyes (GE Healthcare),
DMSO (ACS reagent, Sigma), 4 M hydroxylamine (Fisher
Scientific).

4. **Purification of dye-labelled first-strand cDNA**: NucleoSpin
   Extract II Kit (Abgene), molecular biology grade water.

5. **Hybridisation of target to rainbow trout oligonucleotide micro-
   array**: Molecular biology grade water, 2 × hybridisation
   buffer (Agilent), gasket slides (Agilent).

6. **Post-hybridisation array washing**: 20 × SSPE (Cambrex),
   20% N-Lauroylsarcosine (Sigma), PEG 200 (Sigma).

---

### 3. Methods

**3.1. Generation of Non-redundant List of Target Sequences**

Matching the list of target sequences to the capacity of the chosen
array platform while maximising gene representation and minimising
gene redundancy is undertaken by a series of steps, as illustrated in
[Fig. 14.1](#) for our project on the rainbow trout oligoarray. These are:

(i) Obtain sequences from databases or by production of cDNA
clones.

(ii) Assemble the ESTs to generate a minimally redundant collection
of contigs.

(iii) BLAST the sequences against appropriate sequence databases.

(iv) Use the BLAST identities to identify the most informative
sequences, first by seeking a suitable hit from a priority list of
databases, then by exploring full-length annotated cDNAs.

(v) Ensure that the “informative sequences” are non-redundant,
that is, non-overlapping sequences possessing the same gene
identity are pruned so that only one sequence remains.

**3.1.1. Identification and Filtration of Sequences**

Sequence collections generally include low-quality reads, reads
from non-RNA sources as well as sequences that read in the anti-
sense direction. A useful means of filtering these is to use a
BLAST search, which focuses attention upon sequences with
“informative” annotation, and furthermore identifies the “sense” direction which is important for specifying oligoprobes (see Note 8). It also identifies potentially redundant sequences that are non-overlapping, but still probably arise from the same gene. Figure 14.1 shows how this step reduced a list of 57,000 contigs
and singletons assembled by the Gene Index Project down to 25,200 BLAST-identified sequences all of which were placed in the sense direction.

This approach has three caveats; first, the search is limited to the sequences in the reference database(s) that are interrogated – optimally these sequences should include full-length cDNA sequences; second, BLAST “hits” of segments arising from the 3' untranslated region are unlikely to establish identity due to poor conservation of sequences between species and third, due to existence of large gene families, the co-location of ESTs onto a single full-length cDNA is not necessarily an absolute indication of their common ancestry.

3.1.1.1. Selecting the Most Appropriate Sequence Databases for Alignment Searching

Databases used for BLAST alignment should be selected on the basis of their phylogenetic relationship to the species under investigation. Thus, for rainbow trout we selected three protein-sequence databases for searching, namely Swiss-Prot (15) → RefSeq → the non-redundant (nr) database.

3.1.1.2. Conducting the BLASTX Search

After downloading each set of sequences, they need to be formatted to make them searchable by BLAST. This is readily achievable using the “formatdb” command, which is part of the BLAST package described in Section 2.1.5; formatdb -i <name of fasta file of sequences>. If the FASTA file contains nucleotide sequences, then you will need to add the flag “-p F.” To search against these databases, you will need to use the appropriate BLAST command: blastall -p <command> -i <your file of assembled ESTs> -d <your file of reference sequences> -e 0.001 -o <name of output file> -T where <command> is blastx if searching against a set of protein sequences (swissprot, uniprot); or blastn if searching against a set of nucleotide sequences (e.g. collections of mRNAs or cDNAs). The -e 0.001 option tells the program to return significant hits only, while the -T option ensures that the output is returned in the HTML format required for its subsequent parsing (see Section 3.1.1.3).

3.1.1.3. BLAST Searches

After subjecting the sequences collated in Section 2.2 (e.g. assembled contigs and singletons) to BLAST searches against your reference databases, PERL scripts can be used to extract the key information from the resulting output files. We provide four PERL scripts (“readBlastSP.pl,” “readBlastRefSeq.pl,” “readBlastNR.pl,” “readBlastMGC.pl,” available from http://legr.liv.ac.uk), which allow the extraction of BLAST matches from searches performed against Swiss-Prot, RefSeq, the non-redundant (nr) protein database and the full-length cDNA database, respectively. They provide the query sequence identity and length, the top BLAST hit information (description, identity, db accession number) and the alignment information (E-value, bit score,
alignment region, matched frame), all of which should be stored in a tab-delimited table in Microsoft Excel. This table indicates sequences lacking a significant hit (i.e. E-value \( \leq e^{-15} \)), with antisense hit(s) (alignment frame -1, -2 and -3) or with hits against mitochondrial sequences. The tab-delimited tables are useful for viewing and sorting results, and filtering the sequences. Thus, sequences lacking a suitable identity can be manually discarded from the Excel file. Similar tab-delimited tables were generated and used in the next steps of the filter by PERL scripts.

3.1.4. Filtering of Sequences with Repeated BLAST Identities

Sort the Excel spreadsheet on the column containing the top BLAST hit. Then manually remove redundant sequences leaving the longest read in place.

3.1.5. Removal of Non-meaningful BLAST Hits

Another filter is to exclude sequences that align against existing sequence databases, but which lack a “meaningful” BLAST identity. Less-than-meaningful terms frequently used as BLAST descriptors include “hypothetical protein . . .”, “predicted protein . . .”, “similar to . . .”, “protein . . .”, “mRNA . . .”, “cDNA . . .”, “unknown protein . . .”, “unnamed protein . . .”, “zgc: . . .”. A high proportion of such sequences are cloning artefacts, random pieces of genomic DNA, etc. and where array capacity is limiting then losing these is the least worst option. Using the Excel file, sort your non-redundant sequences on the basis of their BLAST descriptions to rapidly group and remove sequences with similar non-meaningful terms.

3.1.6. Recoveries of Non-meaningful Hits

Finally, as a conservative step to recover sequences discarded in Section 3.1.1 that might have a meaningful description in the less well-curated Genbank sequences, perform a BLAST search of the discarded sequences from Section 3.1.1.5 against the Genbank nr database. Previously removed sequences that exhibit “meaningful” description in this search should be added back to those retained in Section 3.1.1.5. Figure 14.1d shows how these methods reduced 22,500 sequences down to ~11,000 sequences for submission to the Nimblegen platform.

3.2. Aligning the Sequence Collection on the Zebrafish ZGC and the Mouse MGC

Unwanted redundancy in EST collections may arise when contigs or singletons from the same gene do not overlap, such that they will not be assembled together. Sometimes they will have identical BLAST identities, but in other cases one of the ESTs may have an unknown identity perhaps due to its 3’ location. Reduce this redundancy by seeking co-alignment of singletons or contigs onto a collection of full-length cDNA sequences, followed by removal of the least useful clone from the collection. Due to their “model” status there are large collections of full-length, high-quality transcript sequences for both mouse and zebrafish,
which are suitable for this purpose, at least for vertebrate animals. *Drosophila* and *C. elegans* cDNA collections are probably most useful for invertebrate animal species.

3.2.1. Co-aligning Contigs onto the Full-Length cDNAs

First, align the contigs of interest onto the full-length cDNA collections of the zebrafish gene collection (ZGC) using BLAST [with the blastn command with a conservative alignment criterion (E ≤ e-15)]. cDNAs for which more than one query contig was aligned were used to select the most appropriate query contig to represent that cDNA. Usually the longest one was selected, and others were discarded. Selected contigs were stored for a further filter described in Section 3.3. Sequences not aligning or showing a 1-to-1 alignment (only one query sequence aligning on the cDNA) were used as the source material for a second-stage alignment onto the mouse full-length cDNA collection. Again where more than one trout contig aligns onto a single mouse cDNA, retain the longest sequence and discard the others. Combine the retain sequences with those contigs not aligning or showing 1-to-1 alignment and store them for use in the next filter described in Section 3.3. Figure 14.1 (boxes b and c) shows how 26,400 trout sequences were trimmed back to 22,500 sequences using this technique. BLAST results for cDNA alignments were parsed into a tab-delimited table. Microsoft Excel was used to manipulate the data. The most appropriate sequences are selected by sorting hits’ IDs first and then query sequences’ lengths.

3.3. Collation of Resulting Gene Collections, Prediction of Optimised Oligoprobes and Submission to Third-Party Array Manufacturers

The probe capacity of the chosen platform defines the extent to which the sequence dataset must be trimmed back. At the time of our trout project, *Agilent* and *Oxford Gene Technology* offered slides containing 22,000 probes (though see Note 3). Other platforms may offer smaller numbers of probes, may require multiple probes per sequence or may require inclusion of extensive control probes or mismatch probes, all of which requires that the list of target sequences is reduced to match the array capacity. With this in mind, we recommend that you focus on genes possessing (i) a well-established identity through BLAST homology searching and (ii) a meaningful description of function, and by discarding sequences that perform least well in these respects.

3.3.1. Reducing the Set of Target Sequences to Match the Array Capacity

Sort the redundant sequence descriptions generated in Section 2.2 in an Excel spreadsheet using BLAST top-hit descriptions and their bit scores. If sequences share the same BLAST top hit, the one with the highest bit score for the top hit is retained, and others are discarded.
The sequence lists generated in Sections 3.1 and 3.2 can also be reduced by applying more stringent criteria for BLAST identification, or by selecting genes possessing a well-annotated identity or those which are associated with particular biological processes or functions of specific interest, such as apoptosis.

### 3.3.2. Prediction of Oligoprobes

Submit the collated list of sequences to the probe prediction package *ArrayDesigner* (http://www.premierbiosoft.com/dnmicroarray/index.html, see Note 9 for other suggestions). This generates a list of potential oligoprobes of the desired length with scores reflecting the likelihood of hybridisation performance. Select the predicted best-performing probe (but see in Section 3.4 for a suggested optimisation of probe design).

### 3.3.3. Submission of Probe Sequences to Array Fabricator

The FASTA-format sequences should be submitted to the manufacturer for fabrication. Increasingly this is via a web-based submission system such as Agilent's *eArray* (http://earray.chem.agilent.com/earray/).

### 3.4. Selection of Optimised Oligoprobes

While probe-design algorithms in current use are sophisticated, they do not capture all of the important features of oligoprobe behaviour. Consequently, there is no guarantee that the top predicted optimal probe actually performs the best under the experimental conditions. Thus, we suggest that users compare several of the predicted probe designs for a given test sequence, and to select the one that displays the best-performance characteristics experimentally. For this we suggest using two specific tests; the first compares each probe with its corresponding mismatched pair (the same sequence but with a few base substitutions introduced) to test for rejection of similar sequences, that is hybridisation specificity. The second test focuses on the rejection of non-specific binding based on the observation that it occurs much more rapidly (4 h) than specific hybridisation (24–48 h) and the two can be discriminated by testing binding after a short and a long period of hybridisation (16).

#### 3.4.1. Selection of Oligoprobes

From the oligoprobe design software, identify three 65-mer oligonucleotide designs for each of the test sequences produced in Section 3.3. For each of these probes, generate a mismatch version containing alternative bases at positions 10, 20 and 30 from the 3' end. This comprises 6 probes for each of the selected sequences. Finally, commission the fabrication of array sets containing each of the resulting 3 match and 3 mismatch arrays for the tests (see Note 10).

#### 3.4.2. Optimisation Tests

For the match/mismatched comparison, complete the hybridisations against a single, labelled cDNA preparation from the tissue of choice (see protocols below). The arrays should be scanned and the resulting .tiff files analysed using appropriate software (see...
Note 11). Intensity values for each perfect match (PM) and its mismatch pair (MM) should be introduced into an Excel spreadsheet and for each pair a simple measure of sequence specificity should be calculated, thus:

\[ R = \frac{PM - MM}{PM + MM} \]

The difference value divided by the sum generates a number between –1 and +1. Perfect rejection of the mismatch probe would give \( R = +1 \), and no rejection would give \( R = 0 \). To favour probes of equal discriminatory power but with increased spot intensity, we adjusted the factors by introducing an empirical weighting for spot intensity.

\[ R' = R + \left( \frac{R}{10,000} \right) \]

For the second test, hybridise two PM array sets against the same labelled cDNA preparation as used in the first test, one for a hybridisation period of just 4 h and the other for 48 h. From the analysed data, we again computed for each probe:

\[ T = \frac{(48 \text{ h} - 4 \text{ h})}{(48 \text{ h} + 4 \text{ h})} \]

again generating a number between –1 and +1 with similar properties as indicated for \( R \). This was also intensity weighted:

\[ T' = T + \left( \frac{T}{10,000} \right) \]

Combine the two tests by summing \( R' + T' \), which is then used to select the top-ranked probe using the SORT function in Excel (Microsoft). The selected probes are then collated into a final probe list for submission to the array fabricator for final production.

3.5. Hybridisation, Imaging and Data Acquisition; Technical Protocols

3.5.1. Methods

3.5.1.1. Synthesis of Aminoallyl-Labelled First-Strand cDNA

1. Combine 10 \( \mu \text{g} \) of total RNA, 2 \( \mu \text{l} \) of 100 \( \mu \text{M} \) oligo(dT)\text{20}, 2 \( \mu \text{l} \) of 20 \( \times \) aminoallyl-dUTP/dNTPs and molecular biology grade H\text{2}O to give a final volume of 26 \( \mu \text{l} \).

2. Incubate the mixture at 65°C for 5 min.

3. Cool the sample on ice for at least 1 min and then add the following: 8 \( \mu \text{l} \) of 5 \( \times \) first-strand buffer, 2 \( \mu \text{l} \) of 0.1 M DTT, 2 \( \mu \text{l} \) of molecular biology grade H\text{2}O and 2 \( \mu \text{l} \) of SuperScript III (200 units/\( \mu \text{l} \)).

4. Incubate the reaction at 50°C for 1 h.
5. Terminate the reaction by adding 20 µl of stop solution and incubating at 65°C for 15 min.
6. Add 60 µl of neutralising solution.

### 3.5.1.2. Purification of First-strand cDNA

1. Pre-heat an aliquot of molecular biology grade H₂O (pH 7.0-8.5) to 65°C.
2. Transfer the terminated and neutralised reverse transcription reaction to a 1.5-ml microcentrifuge tube and add 240 µl of buffer NT (2 volumes of NT to 1 volume of sample).
3. Load the mixture onto a NucleoSpin Extract II column placed in a 2-ml collection tube and centrifuge the column assembly at 13,000 rpm (maximum speed in a microcentrifuge) for 1 min at room temperature. Discard the flow-through.
4. Wash the column with 600 µl of 80% ethanol (instead of buffer NT3 supplied with the kit) and centrifuge as in Step 3. Discard the flow-through.
5. Wash the column with 200 µl of 80% ethanol and centrifuge as in Step 3. Discard the flow-through.
6. Remove residual ethanol from the column by centrifuging the assembly at 13,000 rpm for 4 min at room temperature.
7. Transfer the column to a fresh 1.5-ml microcentrifuge tube and add 40 µl of pre-heated molecular biology grade H₂O (65°C) to the centre of the silica membrane.
8. Incubate the column at room temperature for 1 min and then centrifuge the assembly at 13,000 rpm for 1 min at room temperature. Retain the eluate.
9. Add another 40 µl of pre-heated molecular biology grade H₂O to the membrane and repeat Step 8.
10. Prepare two identical aliquots of the cDNA preparation by mixing the eluate (~76 µl) and transferring half of the sample into another 1.5-ml microcentrifuge tube and vacuum centrifuge both aliquots until just dry.

### 3.5.1.3. Coupling of aminoallyl-labelled first-strand cDNA with Cy dye

1. Resuspend the dry first-strand cDNA (from Step 10, Section 3.5.1.2) in 5 µl of 0.1-M sodium bicarbonate.
2. Dissolve the Cy dyes in an 100µl of DMSO. Then add 5µl of the selected Cy dye to the cDNA solution and incubate at room temperature for 1 h in the dark.
3. Quench any uncoupled Cy dye by adding 6 µl of 4-M hydroxylamine and incubating at room temperature for 15 min in the dark.
3.5.1.4. Purification of Dye-labelled First-strand cDNA

1. Pre-heat an aliquot of elution buffer NE to 65°C.
2. Add 70 μl of molecular biology grade H₂O to the Cy3-labelled cDNA and 200 μl of buffer NT to the Cy5-labelled cDNA; i.e. the cDNA samples to be co-hybridised.
3. Combine the two samples and load the mixture onto a NucleoSpin Extract II column placed in a 2-ml collection tube.
4. Centrifuge the column assembly at 13,000 rpm for 1 min at room temperature. Discard the flow-through.
5. Wash the column with 600 μl of buffer NT3 and centrifuge as in Step 4. Discard the flow-through.
6. Remove residual ethanol from the column by centrifuging the assembly at 13,000 rpm for 3 min at room temperature.
7. Transfer the column to a fresh 1.5-ml microcentrifuge tube and add 50 μl of pre-heated buffer NE (65°C) to the centre of the silica membrane.
8. Incubate the column at room temperature for 1 min and then centrifuge the assembly at 13,000 rpm for 1 min at room temperature. Retain the eluate.

3.5.1.5. Hybridisation of Target to Rainbow Trout Oligonucleotide Microarray

1. Denature the dye-labelled target (from Step 8, Section 3.5.1.4) at 100°C for 2 min in the dark, then centrifuge the tube briefly at room temperature to collect the condensation.
3. Pipette the mixture onto a gasket slide inserted into an Agilent hybridisation chamber.
4. Place a microarray with the numeric barcode facing upwards (see Note 13) on top of the gasket slide and assemble the hybridisation chamber.
5. Hybridise overnight or over 2 nights (~ 40 h, see Note 14) at 60°C with rotation at 4 rpm.

3.5.1.6. Post-hybridisation Array Washing

1. Disassemble the hybridisation chamber.
2. Immerse the sandwiched slides (gasket slide plus microarray, with the numeric barcode side facing upwards) in 6 × SSPE, 0.005% N-Lauroylsarcosine and pry the two slides apart from the barcode end.
3. Wash the microarray in 6 × SSPE, 0.005% N-Lauroylsarcosine for 5 min at room temperature in the dark.
4. Wash the microarray in 0.06 × SSPE and 0.18% PEG 200 for 5 min at room temperature in the dark.
5. Dry the array by centrifugation at 1000 rpm for 5 min at room temperature in a microtitre plate centrifuge (e.g. Eppendorf 5804 with A-2-DWP rotor).

6. Scan the array.

It is beyond the scope of this chapter to provide full details of experimental design and the subsequent statistical analysis of the microarray data; however, we provide the following guidelines.

The power of microarray analysis depends critically on adopting an appropriate experimental design based on conventional statistical principles (17). There are several reviews and books describing this specifically within the context of microarray analysis (18–20). Defining an appropriate statistical design at the outset, including the number of independent biological replicates, establishes the scale and cost of the array experiment as well as the form of the subsequent analysis. There are two principle designs in the two-colour array format; (i) the reference-based design, in which each RNA sample is compared on the same microarray with a reference RNA that is common to all arrays being analysed. Typically, it is necessary to run two slides for each RNA sample, with the reversal of the fluorophores (“dye-swap”) used for the sample and reference to overcome potential bias in the performance of the dyes, and (ii) the ANOVA-based design (21), which does away with the reference sample and compares two different RNA samples on each array according to a scheme drawn up to maximise statistical power. This approach offers equivalent or greater statistical power compared with the reference-based method but using half of the microarrays, thus increasing cost-effectiveness.

There are 5 stages to the statistical processing of the data; (i) quality control, which involves the identification and removal of poor-quality spots, and, if desired, the subtraction of background values from foreground values of spot intensity, (ii) the within-slide normalisation of data to some standard value to overcome artefacts. In two-colour, ratio experiments, the individual spot values are usually normalised to the median value for all of the high-quality spots, and in some cases the variance may also be normalised. Secondly, early work using two-colour arrays indicated some intensity-dependent differences in the performance of the two fluors, so it is usual to apply the “Lowess” intensity-dependent correction (22), which also achieves the normalisation step just described. Stage (iii) consists of determining the significance value associated with different intensity values between replicated values from experimentally compared conditions. Stage (iv) is the application of some correction for multiple sampling, to overcome the fact that with so many statistical evaluations (i.e. one for each feature on the array), the conventional $P$ value of 0.05 provides many hundreds of potentially false-positive genes.
In stage (v) genes that are shown to be significantly different between compared conditions may then be subjected to intensive interpretation through the application of pattern searching and ontological profiling techniques. The needs of data analysis in both approaches are relatively demanding for most biologists and we strongly recommend that a statistician is consulted during both the design and analysis stages to get the best outcome from what might be an expensive experiment.

### 3.7. Conclusions

In optimising sequence datasets for construction of oligoarrays, the principle objective is to maximise the representation of genes of interest, while minimising redundancy. The number of platforms for array fabrication is increasing and the scale of gene representation is growing as feature sizes become smaller. Most importantly, many advanced commercial platforms now offer highly flexible, cost-effective routes to the small production runs, requiring only that a list of target sequences is provided and submitted online. As sequence resources expand worldwide, the means by which customised arrays can be designed for highly specific purposes, and without great expense, now lies in the hands of individual investigators. We illustrate these principles by showing how we designed and fabricated a high-density oligoarray for the rainbow trout, a non-model species but for which large numbers of ESTs, and an assembled, non-redundant assembly were available (see Fig. 14.2). This array has a much wider gene

![Fig. 14.2.](image)

**A** shows a detail from the array image generated after scanning. This shows adequate hybridisation intensities, and also the quality of the spot morphology and placement, and also that the background fluorescence was low compared with spot intensity. **B** shows a “self-self” plot in which both Cy-labelled channels of a single array sampled the same trout brain cDNA sample. The data points should be identical in each channel and thus should lie on the diagonal line. Dispersion reflects the technical errors inherent in the approach. **C** shows a typical experiment in which different Cy-labelled cDNA samples were used in each channel. Again the data broadly fit the diagonal line, but the spread of data is greater than in A reflecting differences between samples in a subset of genes. Both **B** and **C** illustrate LOWESS-normalised data.
4. Notes

1. In molecular biology, a hybridisation “probe” is defined as a fragment of DNA, usually of a known identity, which hybridises with complementary sequences present in the query DNA. In contrast to solution hybridisation in Northern analysis, microarray probes are immobilised onto the solid surface of the array as a series of discrete “features” or spots. The features are printed in a grid pattern or array. Usually each contains a different probe although it is entirely possible for the same probe to be printed on different features (i.e. control spots, repeat of specific probes of interest). The “target” is the complex mixture of labelled mRNAs isolated from the tissue of interest, which have been reverse transcribed to first-strand cDNA and fluorescence labelled. The denatured and labelled target in solution then hybridises with its complementary probe on the array surface.

2. Probe length strongly influences hybridisation performance (8). Probes of 65 mers avoid the lack of sensitivity of shorter probes (25–35 mer) without the extra costs of building longer 100–150 mer probes, and this length is the frequently adopted commercial platforms.

3. Array-fabrication technology is rapidly evolving. This is particularly influencing the number of features placed on the array and thus how many sequences can be represented. In mid-2007 Agilent launched a new array format containing 240 K probes, and they aim to develop a more advanced platform with 750 K features before the end of 2008. Nimblegen currently offer a platform containing 385 K features and expect to launch a 2.1 M array before the end of 2007. Both platforms can easily include probes for all known sequences for a particular species and the need to refine sequence collections might be much less important. But both of these platforms can configure each slide with multiple, separate arrays each provided with a separate hybridisation space. This means that a single slide can simultaneously analyse up to 12 different samples, and this reduces costs per array proportionately.
Of course each of these arrays would have a much smaller capacity, e.g. $4 \times 44$ K or $12 \times 120$ K, which means that there is still a need to reduce the number of gene probes to match the reduced capacity, as described in this chapter.

4. BLAST alignments are undertaken in a standalone PC, and PERL scripts are available to perform BLAST searches for large-scale sequences and to parse the results into tab-delimited text files. Download the scripts from webpage http://legr.liv.ac.uk/oligoarray/parser.htm and unzip the file, then 6 PERL scripts will be placed into a folder called “perlscripts”. Use the script “splitfasta.pl” to separate the sequences into individual files and use “blast.pl” to perform the BLAST searches.

5. Using a cluster of PCs offers considerable time saving. Thus, we use a Linux cluster (e.g. http://bioserv2.sbs.liv.ac.uk/~fishomics/) composed of 40 commodity Linux machines linked to a node PC. Rainbow trout ESTs of 50 K aligned against the nr database of GenBank takes just 3 h.

6. For established EST projects, pre-assembled gene sets may be located at public databases for multi-species collations (e.g. Gene Index Project, LEGR Data Centre, PartiGeneDB, TBestDB or NEMBASE) or for single species (e.g. funnyBASE). The Gene Index Project uses TGI Clustering Tools (23) (http://compbio.dfci.harvard.edu/tgi/software) to assemble ESTs taken from dbEST (GenBank) for over 30 species. UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) (24) is an alternative resource for species not included in The Gene Index Project. It provides the best-represented sequence from each sequence group rather than the assembled sequences for a group of overlapping sequences. Other full-length cDNA resources, such as human, rat, cow, etc. are also available on http://mgc.nci.nih.gov/, and the Xenopus Gene Collection (XGC) for Xenopus laevis and Xenopus tropicalis can be found at http://xgc.ncbi.nih.gov/. Information for other full-length cDNA projects can be found at http://www.ncbi.nlm.nih.gov/genome/flcdna/list.cgi.

7. Assembly involved clustering sequences into groups that represent genes. Sequences that are closely matched are then joined together to give an extended read, termed a “contig”. Sequences that fail to align to any other sequence are termed “singletons”. The entire assembled collection thus comprises both contigs and singletons.

8. Oligoprobes must be specified to the array manufacturer as either sense or anti-sense orientation. Which one depends on the method by which the fluorescence-labelled transcript or transcript products are generated.
9. Probe design predicts the oligonucleotide that maximises hybridisation to the sequence of choice yet avoids cross-hybridisation to related transcripts (9). A second issue is to achieve designs that have the desired hybridisation properties under the hybridisation conditions being used (temperature and solution/array surface chemistry). A well-used and popular open source package for oligodesign is OligoArray2, (25) currently available as version 2.1 (http://berry.engin.umd.edu/oligoarray2_1/). Agilent (http://earray.chem.agilent.com/earray/login.do) offer an on-line design capability tied to their fabrication platform. Oxford Gene Technologies (http://www.ogt.co.uk) and Nimblegen (http://www.nimblegen.com) can undertake the entire design and fabrication role on behalf of clients. There is rather less consensus on how to achieve validation and optimisation of array designs, though these issues are now being addressed (8).

10. As an example of this, we have collated a refined set of 22,000 rainbow trout sequences, each of which was submitted to probe design software. The three probes with the best-predicted performance were regarded as perfect match probes, for which we generated the mismatch equivalents. This gave a total of 66,000 PM probes and 66,000 MM probes. We then commissioned the fabrication of a complete probe set of 6 slides each with 22000 features, comprising 3 slides with the PM probes, and a further 3 slides with the MM probes.

11. Presently, there are two principal types of array scanners in the market, those using laser scanning and others using a CCD camera. A key issue in selection is to ensure that scanner resolution matches the feature sizes of the array. Low-density arrays can be scanned effectively with 10-μm resolution while the newer, high-density oligoarrays require at least 5-μm resolution. Image files generated by the scanning device should then be quantified using appropriate software. All commercial scanners are provided with image processing software of variable quality. Open source software for this can be found at www.tm4.org/spotfinder.html, and a contemporary package that can be used on third-party scanners can be purchased from www.cambridgebluegnome.com.

12. The protocols for all four steps are now well established, though there is some variation in precise conditions and technique between different array and scanner platforms. See www.microarray.org, pga.tigr.org/protocols.shtml, www.flybase.org and www.liv.ac.uk/lmf/protocols, or the websites of commercial platforms for full details. Several manufacturers market expensive hybridisation stations that circulate the hybridisation fluid across the surface of a limited number of microarrays, usually 4. To date we have not found them to offer any
great advantage over the simpler method of static hybridisation under a “liferslip” (www.eriemicroarray.com), a coverslip with raised edges, which holds a specific volume of hybridisation fluid. Slides and lifterslips can be mounted in inexpensive hybridisation chambers such as those marketed by Corning (“Costar”; catalog no 2551; www.corning.com) or Genetix (catalog no X2530; www.genetix.com).

13. Microarray slides from commercial sources may have a bar-code in the form of a stick-on label attached to the top side of the array. This is to aid the identification of the array when used in instruments equipped with infrared barcode readers.

14. It is standard to hybridise overnight, but the intensity of fluorescence can be enhanced by longer incubations, up to 48 h, to reach equilibrium.

References


Exploiting ESTs in Human Health

Sandro José de Souza

Abstract

Expressed Sequence Tags (ESTs) are fragments of cDNA clones. They correspond to the most abundant type of cDNA information available in the public databases. ESTs have been used for expression profiling, gene identification, characterization of differentially expressed genes, and identification of transcript variants among other utilities. In this review I will discuss the major features of the collection of ESTs available in the public domain giving a special emphasis on how this dataset has been used in studies about human diseases.

Key words: ESTs, expression profiling, alternative splicing, SNPs, differential expression.

1. Introduction

With the completion of the Human Genome Project, a major challenge facing biologists is the complete characterization of the human transcriptome. This extraordinary task involves (i) the identification of all human genes, (ii) the characterization of the transcriptome variability, (iii) the definition of the spatial and temporal expression patterns of transcripts and variants, and finally (iv) the association of transcripts and variants with complex biological phenomena, such as diseases.

In the last 15 years, studies involving the transcriptome have changed dramatically. The emergence of high-throughput technologies allowed the accumulation of an unprecedented amount of data. Among these technologies I should cite Expressed Sequence Tags (ESTs) (1), Serial Analysis of Gene Expression (SAGE) (2), Massively Parallel Signature Sequencing (MPSS) (3), and microarrays (4).
Although these datasets are per se very informative, their integrated use has proven to be the most effective way to extract more valuable information. The availability of the human genome sequence has been critical for the establishment of integrated knowledge bases. The genome sequence represents a scaffold on which transcriptome and other types of data are mapped and integrated.

Although the amount of EST data available is huge, there is a clear need for more data. Some tissues from mammalian species are underrepresented in GenBank as well as several pathological situations. As you will see in the next few pages, the informative power of ESTs is outstanding and I expect that the new sequencing technologies (discussed later here) will allow a deeper and wider coverage of the transcriptome through EST sequencing.

In this review I will focus on EST datasets available in the public domain with a special emphasis on how these datasets are being used in studies of human disease.

2. Expressed Sequence Tags

ESTs are single-pass sequences derived from cDNA clones. Since ESTs are derived from cDNA libraries, they can be used to fingerprint the gene expression pattern of the cell/tissue from which the library was constructed. The development of high-throughput sequencing allowed the group of Craig Venter in 1991 (1) to generate ESTs in large scale (5, 6). Those ESTs were primed to the polyA tail of the transcripts and were therefore called 3’ ESTs. The reason to prime the 3’end is both methodological (use of oligoT) and conceptual. The initial idea behind the EST approach was to generate a catalog of genes expressed in a cell. Therefore, all sequences had to be indexed to a specific position in the transcript, allowing the comparison of different messages and the identification of unique sequences. It soon became clear however that sequences derived from the opposite end of the transcripts would also be useful. Those sequences were called 5’ ESTs. In the 1990s, 5’ ESTs would correspond simply to the other end of the clone. It is very common therefore to find 5’ESTs in the middle of the transcripts due to cDNA clones that did not represent a full-length transcript. More recently, strategies have been used in large-scale sequencing projects that target the 5’ cap of transcripts (7). Therefore, these sequences are more reliable in defining the true 5’ end of a transcript. Also in the 1990s, the development of another strategy for generating ESTs significantly
increased the value of the ESTs dataset. The group of Andy Simpson in Brazil used low-stringent PCR in the development of ORESTES, a strategy that generates ESTs directed toward the central part of the transcripts (8). This methodology was the basis of a large-scale EST sequencing project (HCGP, described later). Several publications from this project emphasized the value of the ORESTES since they complement the dataset of 3’ and 5’ ESTs (8–11).

This and other EST-sequencing projects generated millions of sequences. A specific division of GenBank, called dbEST, was created to store EST data. The present release of dbEST (060206) contains more than 36 million ESTs. Table 15.1 shows the number of ESTs for ten of the most abundantly represented species. ESTs from human and mouse correspond to one third of all dbEST sequences. Due to the large-scale nature of EST sequencing, dbEST is extremely redundant. Several initiatives have attempted to cluster EST data into gene indexes. The goal of any gene index is to reduce the huge dataset into a manageable set of ESTs clusters. The major gene indexes are UniGene (http://www.ncbi.nlm.nih.gov/UniGene/), maintained by NCBI, the TIGR gene index (http://compbio.dfci.harvard.edu/tgi/), originally maintained by TIGR and now hosted at Harvard, and finally the STACK database (http://www.sanbi.ac.za/Dbases.html), maintained by the South African National Bioinformatics Institute. All databases, except STACKdb, maintain gene indexes for dozens of species. TIGR Gene Index and STACKdb discriminate between alternative splicing isoforms, a feature not provided by UniGene. An important

<table>
<thead>
<tr>
<th>Table 15.1</th>
<th>Ten species with large number of ESTs in dbEST release 060206</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>7,741,746</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>4,719,380</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>1,184,706</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>1,056,773</td>
</tr>
<tr>
<td>Xenopus tropicalis</td>
<td>1,044,182</td>
</tr>
<tr>
<td>Bos Taurus</td>
<td>1,039,059</td>
</tr>
<tr>
<td>Rattus sp</td>
<td>871,147</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>854,672</td>
</tr>
<tr>
<td>Zea mays</td>
<td>753,472</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>686,395</td>
</tr>
</tbody>
</table>
feature of STACKdb related to the subject of this review is a categorization of sequences. STACKdb is organized into 15-tissue based categories and one disease-based category. The disease category includes all sequences annotated as disease related. Although ESTs correspond to the vast majority of data in the gene-indexes, full-insert mRNAs are an important component in the clustering process. Since they correspond to a full-length high-quality sequence, these sequences are able to bridge ESTs that do not overlap. The number of these type of sequences is growing fast in the public databases because of many large-scale sequencing projects like the Mammalian Gene Collection (MGC) (12).

In principle, ESTs can be used to evaluate the level of expression of a given message. However, their use may be limited by the fact that many libraries are normalized, i.e., they are subject to a procedure that increases the frequency of rare messages to the detriment of more abundant ones. Furthermore, the last decade has seen a widespread use of other methodologies that are more quantitative than ESTs, like SAGE and microarrays.

ESTs have been also widely used to detect transcript variability, mainly alternative splicing [for a review see (13)]. Figure 15.1 illustrates this point. Almost all data on splicing variants were achieved through different methods involving sequence comparisons. The idea behind these methodologies is that splicing variants from the same gene share some common sequence and can therefore be grouped together. The presence of indels (insertions/deletions) in the alignment would suggest the existence of two splicing variants. ESTs are the major source of data in these studies aimed to catalog the splicing variants. Mironov et al. (14), for example, used EST contigs from the TIGR Gene Index to make an inventory of intron-exon structures in the set of known human genes. With the availability of the human genome sequence, several groups started to make more precise inferences about alternative splicing by simply mapping all cDNAs onto the genome sequence (15, 16). By using the genome sequence, one can specifically determine the exon/intron organization of a given cDNA sequence. The use of more precise algorithms of alignment, such as Sim4 (17), which take special attention to the exon/intron border and the consensus splicing signals, made these inferences more reliable.

Instead of comparing pairwise alignments, the state-of-art methodology today aligns all cDNAs against the genome, loads the cDNA coordinates into relational databases, and compares the borders among different cDNAs from the same gene. The approach used by our group (15, 18) is based on the following steps: (i) mapping of cDNAs onto the genome sequence using BLAT; (ii) refinement of mapping, for those cDNAs reporting two or more exons, using Sim4; (iii) grouping of all mapped cDNA into clusters; and (iv) identification of transcripts from the same cluster using alternative exon-intron borders. Besides the genome
sequence, another type of data that has been critical for a more reliable identification of splicing variants is the full-length cDNA sequences (classified as mRNAs in GenBank). These sequences per se report variants. For example, cDNAs from the Reference Sequence initiative (http://www.ncbi.nlm.nih.gov/RefSeq/) report splicing variants for a significant fraction of all human genes. Furthermore, this type of sequence greatly improves the clustering of ESTs since they act as a continuous scaffold on which ESTs can be mapped, serving as bridges between EST clusters. The user of the EST datasets described above, who is interested in searching for splicing variants, should be aware that the fragmented nature of ESTs in many cases precludes the identification of combinations of alternative intron/exon borders. This is restricted to those variants reported by the full-insert mRNAs described above.

Fig. 15.1. UCSC Genome Browser showing part of the transcript variability found in the EST dataset for the gene CD44. Exons are marked as short vertical lines joined together by horizontal lines (introns). Arrows in the introns indicate the orientation of the gene (plus or minus strands). Observe the rich variability of exon usage found within the nonredundant set of full-length mRNAs for this gene (entries marked in blue).
3. ESTs and Disease

ESTs were in principle used to evaluate gene expression profiles in diseases. While most of the interest is dedicated to cancer (see next section), other diseases are targeted as well. A very useful application of ESTs is positional cloning, i.e., the identification of particular genes involved in the phenotype of a given disease solely based on their genomic location. This location is usually deduced from population-based studies where that chromosome segment is somehow changed in affected individuals. The genes in that region are candidate disease-associated genes. After a genomic location is identified, ESTs and other types of cDNA information can be used to identify a disease-associated gene. Fundamental for this identification is the functional characterization of the candidate genes. For example, Katsanis et al. (19) used the whole collection of ESTs to identify positional candidates for genes involved in retinopathy. They have developed a series of computational tools to predict EST clusters likely representing genes preferentially expressed in retina. These cDNAs were then mapped onto the genome sequence and compared to mapped but uncloned retinopathy genes.

ESTs can also be used to identify genes differentially expressed in a given disease state. UniGene has a useful tool to evaluate if a given tissue and/or disease state is differentially represented in a cluster. In each UniGene cluster entry, there is a link for “Expression Profile” which shows a breakdown by tissue and by health state for all cDNAs in the respective cluster. It is important to mention, however, that genes differentially expressed are more reliably identified by other more quantitative approaches such as SAGE, MPSS, and microarrays. The reason for this lies in several factors including the use of normalized libraries for EST sequencing and the lower transcript coverage within EST libraries.

One interesting application of ESTs is for the identification of Single Nucleotide Polymorphisms (SNPs) in transcribed sequences. Although the low quality of EST sequence is a problem for this type of analysis, different algorithms (PolyPhred and PolyBayes, among others) have been developed for the electronic identification of putative SNPs (20, 21). The use of ESTs for the identification of SNPs has both advantages and disadvantages. The advantages include easy access to sequence data, SNPs located in the expressed part of the genome, and initial analysis depending only on computational approaches. However, many disadvantages can be identified including poor quality of sequence reads, small population coverage, and lack of noncoding SNPs (especially those located in promoters). The utility of SNP data mining by
using ESTs was shown in several examples including the identification of mutations in the GADD34 gene (22) and in several genes encoding members of the cytochrome P450 family (23).

Although the importance of ESTs has been discussed for many cases, they are not frequently used directly to identify potential drug targets. After a target is identified though, ESTs are widely used to characterize their splicing variants or putative homologs.

4. ESTs and Cancer

The amount of ESTs derived from cancer samples continues to grow in the public databases. Today more than 50% of all ESTs in dbEST correspond to sequences derived from a large variety of tumor samples. The emergence of large-scale sequencing projects has been crucial for this increase in information available for genes expressed in cancer. As recently compiled by Strausberg et al. (24), there are two major cancer-oriented genome projects that focus their efforts in the transcriptome: the Cancer Genome Anatomy Project (CGAP) and the Fapesp/Ludwig Human Cancer Genome Project (HCGP). CGAP’s main goal is the integration of genomics and cancer research. CGAP produces an index of genes expressed in normal and tumor tissues through the use of both EST sequencing and SAGE (2). The HCGP devoted its efforts (the project is now finished) for the generation of ESTs from a panel of tumors more frequently found in the Brazilian population. The project used a different strategy for the production of ESTs. As already discussed, the ORESTES approach generates sequences predominantly located at the center of the transcripts (25). This is achieved through the use of low-stringency PCR using arbitrary primers. Besides providing sequences from the central part of the transcripts, the low-stringency PCR reaction is expected to apply a normalization effect to the set of sequences.

The completion of the sequencing of the human genome has contributed significantly to cancer research and increased the value of the transcriptome databases. Transcriptome data, especially ESTs, have been crucial for the identification of human genes and the variability found at the transcriptome and genome levels. Identification of genes differentially expressed in tumors is one of the most important forms of mining the cancer-orientated databases. Several reports have been published in which such genes were successfully identified by EST-based analyses (26, 27).

Let us take the CT (cancer/testis) antigens as an example. CT antigens correspond to a class of human genes characterized by a unique pattern of expression. Genes belonging to this class are
predominantly expressed in testis, among normal tissues, and several types of tumor tissues (26). This can be easily observed if one takes one of the most known members of this family, MAGEB1. In the respective UniGene (http://www.ncbi.nlm.nih.gov/UniGene/) entry for this gene one can click on the “Expression Profile” link (within the “Gene Expression” box) and see that this gene is expressed predominantly in normal testis with most of the remaining ESTs coming from melanoma samples.

As mentioned before, ESTs are widely used to detect transcript variability, especially splicing variants. We know nowadays that alternative splicing is quite important in the pathogenesis of several human diseases. It is believed, for example, that around 15% of all human genetic diseases are caused by mutations in sequence elements important for constitutive splicing. Isolated cases of differential expression of splicing variants in cancer have been reported in the last ten years [for a review see (28)]. For example, BCL-x, an apoptosis regulator, has two splicing variants due to alternative donor sites in its exon 2. Only the longer form is differentially expressed in cancer (29). The most widely known example, however, is CD44, a cell-surface glycoprotein. Differential expression of several splicing variants of CD44 has been observed for a range of different tumors [for a review see (28)].

The increasing amount of cDNA libraries constructed from both normal and tumor samples has allowed the development of genome-wide strategies for the identification of tumor-associated splicing variants. Several groups reported genome-wide screening strategies searching for splicing variants differentially expressed in tumors (18, 30–32). Without exception, these authors made use of the huge amount of EST data available in the public databases to identify putative variants differentially expressed in tumors. The proportional frequency of cDNAs derived from distinct variants is an indication whether a given variant is differentially expressed in a library or in a pool of libraries. One of the major problems affecting this type of analysis is the identification of genes, not variants, differentially expressed in tumors. This happens because the computational and statistical methods used in the analyses are not sensitive enough to discriminate the expression level of all variants from a given gene. The critical issue is to discriminate transcription from splicing regulation. Available methods usually flag cases of differentially expressed genes as variants that are differentially expressed. More recently, we tried to overcome this limitation by using SAGE data to discriminate differentially expressed variants from differentially expressed genes (18). We first identified splicing variants that were overrepresented in EST libraries from tumor samples. Overrepresentation was based on the total number of ESTs for each tissue analyzed in either tumor or normal libraries. This generated a list of splicing variants putatively associated with tumors. After
experimental validation, however, we observed that this list was inflated with examples of genes overexpressed in tumors. To filter those cases, we identified those genes that were shown to be overexpressed in the same tumor tissue by SAGE. Our computational approach finally identified more than 1,300 splicing variants putatively associated with cancer. Experimental validation for a subset of these candidates was achieved for both tumor cell lines and patient samples.

If one is interested in searching for splicing variants uniquely or differentially expressed in cancer, many analyses are possible. One can, for example, search the datasets produced by the papers cited above (18, 30–32), which together comprise thousands of genes and variants. Another possibility is to look specifically at alternative splicing databases, such as the Alternative Splicing Database maintained by EBI (http://www.ebi.ac.uk/asd/) and the Hollywood RNA Alternative Splicing Database, maintained by Chris Burge’s group at MIT (http://hollywood.mit.edu/). Finally, if the user is interested in a specific gene or variant, he/she can look directly at the genome browsers available, such as those from UCSC (http://genome.ucsc.edu), NCBI (http://www.ncbi.nlm.nih.gov/mapview/), or Ensembl (http://www.ensembl.org/).

5. Use of Integrated Approaches to Study the Transcriptome

Although EST data have been extremely important in studies of several diseases, the integrated use of different types of data has proven to be more effective. Integration of EST data with the genome sequence has provided excellent results. The complexity of the human genome sequence makes the ab initio identification of genes an almost impossible task. The final validation of gene predictions and exact intron/exon boundaries can be established only by the utilization of transcribed sequences, especially ESTs. We have applied this integrative view of the transcriptome to identify new genes on chromosome 22 (11), 21 (33) and at the Hereditary Prostate Cancer Locus 1 (HPC1) on chromosome 1q25 (34). A similar approach was applied in a large collaborative project executed in the state of São Paulo-Brazil. In the “Transcript Finishing Initiative” we used the genome sequence as a scaffold for EST mapping and clustering (35). We then performed RT-PCR to bridge gaps between EST clusters that are likely derived from the same gene. This strategy proved to be an efficient way to discover new genes and splicing isoforms (35).
The integrated use of ESTs and genome sequence is also illustrated by the already mentioned work from Katsanis et al. (19). These authors used sequences available in dbEST to predict EST clusters representing genes predominantly expressed in the retina. These sequences were then mapped onto the human genome sequence and provided positional candidates for more than 40 retinopathies. This was possible due to the existence of genetic data associating diseases with specific chromosome anomalies or linkage studies associating a given marker to a specific disease trait.

We have also developed an *in silico* protocol for the identification of genes differentially expressed in cancer by using ESTs and SAGE (36). If there are enough EST data for a given tumor and its normal counterpart, these sequences can then be used to identify genes differentially expressed. This analysis is done by statistically testing the proportion of ESTs from a same gene derived from both normal and tumor samples. This analysis generated a list of candidates that guided subsequent approaches. SAGE was then used to corroborate the initial analysis done with ESTs. We used this protocol for a comparison of normal breast tissue with breast tumors. We were able to identify 184 genes differentially expressed in breast tumors. Around 82% of tested candidates were shown to be differentially expressed using other experimental procedures.

### 6. Future Perspectives

The development of next-generation sequencing technologies, as exemplified by three platforms that are already on the market (454, Solexa, and SOLiD), will certainly revolutionize biomedical sciences. It is expected that sequence-based expression profiles will again be the most used form for evaluating gene expression, in detriment of microarray. It is therefore expected that the number of ESTs in dbEST will grow at an even higher pace covering many biological situations including a large variety of different pathologies.

I envisage that in the next few years we will see a huge amount of genome and expressed sequence data from the same individual. This will allow deeper and more informative association studies linking genetic and epigenetic data to specific features of human disease.

### 7. Final Remarks

The major goal of this chapter was to provide the reader with a general overview of the richness of EST datasets regarding different features of human diseases. ESTs are important for the
discovery of new genes, for the identification of transcript variants, for the expression pattern of a given gene, and for the identification of putative SNPs, among other utilities. However, the user has to keep in mind that ESTs are products of a single-pass sequence reaction and are therefore enriched with artifacts. However, when integrated with other types of data, ESTs are extremely rich and are able to significantly contribute to our understanding of human diseases.

References


<table>
<thead>
<tr>
<th>A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI 3700/3730</td>
<td>154, 157, 158, 164, 165–166, 216</td>
</tr>
<tr>
<td>ABI format</td>
<td>23, 154, 156, 157, 158, 164, 165–169, 186, 199, 216</td>
</tr>
<tr>
<td>ABI tracefile</td>
<td>23, 199, 248</td>
</tr>
<tr>
<td>Absorbance ratio</td>
<td>135</td>
</tr>
<tr>
<td>Access permissions</td>
<td>253</td>
</tr>
<tr>
<td>Actuator arm</td>
<td>160, 161</td>
</tr>
<tr>
<td>Adapter sequence</td>
<td>170, 192, 200</td>
</tr>
<tr>
<td>Adapter trimming</td>
<td>192</td>
</tr>
<tr>
<td>Advantage 2 polymerase mix</td>
<td>157, 171</td>
</tr>
<tr>
<td>Affymetrix Genechip platform</td>
<td>290</td>
</tr>
<tr>
<td>β-agarase</td>
<td>75, 100</td>
</tr>
<tr>
<td>Agilent array</td>
<td>295</td>
</tr>
<tr>
<td><em>Alexandrium tamarense</em></td>
<td>110</td>
</tr>
<tr>
<td>Alkaline treatment buffer</td>
<td>55, 61</td>
</tr>
<tr>
<td>Alternative splicing</td>
<td>9, 209, 252, 313, 314, 318, 319</td>
</tr>
<tr>
<td>Alternative transcripts</td>
<td>5</td>
</tr>
<tr>
<td>Ambiguous basecalling</td>
<td>222</td>
</tr>
<tr>
<td>Ampicillin</td>
<td>74, 75, 91, 101, 111, 114, 133, 139</td>
</tr>
<tr>
<td>Amplification cycles</td>
<td>46</td>
</tr>
<tr>
<td><em>Arabidopsis</em></td>
<td>20, 27</td>
</tr>
<tr>
<td><em>Ancylostoma caninum</em></td>
<td>51, 52, 65, 155, 278</td>
</tr>
<tr>
<td>ArrayDesigner</td>
<td>299</td>
</tr>
<tr>
<td><em>Ascaris suum</em></td>
<td>25</td>
</tr>
<tr>
<td>Autoclaving</td>
<td>44</td>
</tr>
<tr>
<td>AutoFACT, see EST processing software</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Back extraction</td>
<td>58, 63, 91, 96, 98</td>
</tr>
<tr>
<td>Bacterial contamination</td>
<td>33, 167, 198</td>
</tr>
<tr>
<td>Basecaller</td>
<td>166</td>
</tr>
<tr>
<td>Basecall extraction</td>
<td>154, 176, 192, 199</td>
</tr>
<tr>
<td>Batman sequencing program</td>
<td>165, 166, 182, 183</td>
</tr>
<tr>
<td>Bead deposition device</td>
<td>175</td>
</tr>
<tr>
<td>Bead washing</td>
<td>38, 45, 60, 88, 95, 162, 172</td>
</tr>
<tr>
<td>Betaine</td>
<td>18</td>
</tr>
<tr>
<td>BigDye</td>
<td>133, 139, 156, 216</td>
</tr>
<tr>
<td>Binding buffer</td>
<td>36, 40, 55, 60, 74, 95</td>
</tr>
<tr>
<td>Binomial distribution</td>
<td>285</td>
</tr>
<tr>
<td>Binomial random variable</td>
<td>283</td>
</tr>
<tr>
<td>Binomial test</td>
<td>278, 284</td>
</tr>
<tr>
<td>Bioindicators</td>
<td>7</td>
</tr>
<tr>
<td>Bio-Linux</td>
<td>193, 197, 206</td>
</tr>
<tr>
<td>Biomek FX instrument</td>
<td>163, 164</td>
</tr>
<tr>
<td>BioPerl, see Perl</td>
<td>51, 54, 59, 73, 88</td>
</tr>
<tr>
<td>Biotin-hydrazide</td>
<td>50, 52, 54, 59, 66, 73, 87, 88</td>
</tr>
<tr>
<td>Biotinylolation buffer</td>
<td>54</td>
</tr>
<tr>
<td>BLAST annotation</td>
<td>210</td>
</tr>
<tr>
<td>BLAST (Basic local alignment Search Tool)</td>
<td>6, 7, 141, 167, 177, 178, 180, 184, 185, 192, 194, 195, 196, 201, 202, 204, 210, 211, 212, 213, 217, 218, 223, 227, 228, 231, 236, 249, 257, 258, 259, 262, 264, 266, 267, 272, 292, 294, 296, 297, 298, 306</td>
</tr>
<tr>
<td>BLAST bit score</td>
<td>201, 264, 268, 274</td>
</tr>
<tr>
<td>BLAST databases</td>
<td>193, 201, 206, 217, 227, 236</td>
</tr>
<tr>
<td>Blocking oligonucleotides</td>
<td>74, 93, 105, 112, 122 n.1</td>
</tr>
<tr>
<td>BME</td>
<td>71, 76</td>
</tr>
<tr>
<td>BNN132 cells</td>
<td>75, 101</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>25, 128, 131, 201, 210, 223, 290, 298</td>
</tr>
<tr>
<td>Cancer</td>
<td>316, 317, 318, 319, 320</td>
</tr>
<tr>
<td>CapTrapper</td>
<td>2, 154, 192</td>
</tr>
<tr>
<td>Cap structure</td>
<td>19, 50, 70</td>
</tr>
<tr>
<td>CapTrapper</td>
<td>19</td>
</tr>
<tr>
<td>Captropping, see CapTrapper</td>
<td></td>
</tr>
<tr>
<td>Carbohydrate-rich molecule contamination</td>
<td>143 n.16</td>
</tr>
<tr>
<td>Amplification</td>
<td>16, 19, 21, 22, 28, 33, 34, 35, 37, 41, 112, 113, 117, 119, 127, 130, 133, 136, 137, 146 n.36</td>
</tr>
<tr>
<td>Coverage and redundancy</td>
<td>278</td>
</tr>
<tr>
<td>Directional</td>
<td>14, 23, 35, 111, 132, 133, 139, 145 n.33</td>
</tr>
<tr>
<td>First strand synthesis</td>
<td>45 n.9, 90, 132</td>
</tr>
<tr>
<td>Normalization</td>
<td>15, 20, 21, 34, 35, 37, 42, 65 n.1, 70, 74, 91, 93, 104 n.20, 109–122, 190, 277, 281, 317</td>
</tr>
<tr>
<td>Quality assessment</td>
<td>14, 22, 135, 136, 137, 139</td>
</tr>
<tr>
<td>Redundancy</td>
<td>14, 15, 28 n.6, 29, 141, 190, 242, 253, 273 n.4, 278, 292, 294, 297, 304</td>
</tr>
</tbody>
</table>

325
CDNA libraries (continued)
  Second strand synthesis ........................................... 19, 91, 127
  Size distribution ............................................ 42, 43, 145
  Size fractionation ........................................ 34, 46 n.15, 137, 138, 146 n.35
  SL-TT ...................................................... 127, 130, 131, 144 n.21
  Source of material .......................................... 4, 15, 34,
  52, 65, 70, 75, 127, 130, 134, 141, 143
  Spliced leader ................................................ 19, 125, 126, 127, 201
  Starter ................................................................ 118, 119, 120
  Subtraction .................................................. 1, 21, 65 n70, 73, 91, 93, 104 n.20,
  105, 111, 117, 119, 120, 290, 303
  Suppression subtraction ................................... 15, 21, 290
  Transformation ............................................... 19, 20, 22,
  43, 44, 133, 139, 155, 158, 280, 282
Cell breakage .................................................. 34, 56, 77, 142
Cell cycle ................................................................ 8, 46, 120
Cellulose columns ............................................... 39
Centricep columns ............................................. 116
Checking RNA purity ........................................ 17
Chimeric sequences .......................................... 14, 248
Chloramphenicol ............................................... 38, 43
Chloroform extraction ....................................... 58, 89, 91, 92, 93, 94, 96,
  102, 144 n.22, 145 n.31
Chromatogram .................................................. 6, 191, 199, 205
Ciona intestinalis ............................................ 129, 141 n.1, 313
CLOBB, see EST processing software
CloneMiner cDNA construction kit ....................... 17, 18
Cloning and expression systems .......................... 18
Cloning vectors ............................................. 18, 45, 56, 64, 70, 75, 98, 122 n.1,
  145 n.33, 156, 200
  Lambda ....................................................... 18, 56, 64, 66 n.6, 75, 102
  pBluescript .................................................. 18
  pDNR-LIB ................................................... 18
  pDONR ...................................................... 18
  pSMART ...................................................... 18
  pSPORT ..................................................... 18, 133, 139
  pT3T7-pac ............................................... 112, 122 n.1
  TOPO-TA .................................................. 145–146 n.33
Cluster3.0 ....................................................... 260, 270, 271
Cluster assembly ............................................... 192, 207–210, 247
Cluster identifier ............................................. 207, 208
Clustered sequences ........................................ 207, 306 n.7
Cluster size distribution ................................... 29–30
Clusters of orthologous groups of proteins (COG) ..... 250,
  254 n.4
Cluster structure ............................................... 213, 247
Coding regions ............................................ 23, 35, 69, 130, 222, 223, 224, 225,
  229, 230, 234, 237
Codon bias .................................................... 14, 224
Codon usage table ......................................... 225, 229, 232, 238 n.6
Cold start reaction ......................................... 86, 104 n.10
Colony picking ............................................... 159, 161
Competent cells ............................................ 20, 38, 43–44, 155, 158
Compute cluster ............................................ 236, 260, 272
Condition-specific genes ................................. 120
Configuration file ........................................ 174, 197, 198, 206, 210, 212,
  217, 230, 233, 236, 237, 238 n.6
Consensus sequences, see Contigs ........................ 318
Constitutive splicing ........................................ 168
Contact file .................................................. 168
Contaminant database ..................................... 167, 184
Contaminating sequence .................................. 14, 65,
  78, 127, 167, 177, 184, 192, 198–200
Contigs ....................................................... 5, 209, 222, 224, 228–232, 237, 241, 292,
  294, 295, 296, 297, 298, 306, 314
Cost effectiveness ........................................... 4, 14, 303
Ct analysis ................................................... 116
CPAN ...................................................... 193, 226, 235, 236, 260
Creator SMART cDNA library construction .......... 17,
  20, 35
Cre-lox mediated excision .................................. 75, 98, 102
Cre recombinase ............................................ 75, 102
Cross_match, see EST processing software
Cross-hybridisation ......................................... 307
CTAB ....................................................... 71, 73, 78, 79, 86, 87
Cupiennius salei ........................................... 196, 199, 203, 204, 205, 207, 211,
  212, 213, 218
Cy3/Cy5 labelling ........................................... 302
D
  Database design .......................................... 245–246
  Databases, see EST databases
  Data processing pipeline .................................. 242, 249
  Data submission ......................................... 154
  DbEST, see EST databases - dbEST
  DbEST submission ........................................ 168, 196, 198, 202, 204
  DE-81 chromatographic paper .......................... 56
  Decoder ................................................. 223, 226, 228, 229, 230, 232, 234, 235,
  237, 238
  DEPC ...................................................... 44, 71, 72, 77, 79, 81, 132, 142 n.7
  DH10B cells ................................................ 75, 102, 117
  DH5α cells ............................................... 38
  Dialysis ..................................................... 45, 75, 98, 99
  DIANA-EST, see EST processing software
  Differential expression .................................... 318
  statistical tests .......................................... 278
  splicing variants in tumors ............................ 318–319
  Directional clones ....................................... 14, 23, 111
  Directory structure ..................................... 169, 203, 209, 225
  Discovery rate .......................................... 23, 110, 120, 278–279
  Disease associated genes ............................... 316
  Dispo-column .......................................... 72, 79
  DMSO ...................................................... 18, 294, 301
  DNA driver .............................................. 95, 112, 113, 115, 118
  DNA ligase buffer ....................................... 132, 133, 138, 139
  DNA polymerase ....................................... 18, 104 n.12, 112, 127, 133,
  136, 138, 142, 146
  DNase treatment ........................................ 132, 135, 143 n.15
DNTP mixture ...........................................37, 56, 63, 72, 74, 84, 85, 96, 103–104, 112, 115, 132, 135, 170–171, 293
Driver concentration ..................................104–105, 116–118
DSN/DSN enzyme, see Duplex specific nuclease
DSN storage buffer..................................37, 42
Duplex specific nuclease.............................20–21, 37, 42
Dye swap..................................................303
Dynabeads .................................................157, 172, 173

**E**

*Echinococcus granulosus* .................7, 126, 128, 131, 133, 137, 143 n.16, 145 n.30, 259, 261, 264
Efficiency of transformation .........................114
Electroelution ...........................................38, 45, 46
Enzyme commission (EC)...........................250, 254
Equilibrated phenol ....................................53, 98
*Escherichia coli* strain C56 ..................75, 101, 600
EST databases .............................................241–255
dbEST.....................................................2, 4, 5, 6, 8, 22, 153, 166, 167, 168, 178, 180 n.1, 190, 193, 196, 198, 201, 202, 204, 207, 218, 251, 254, 292, 293, 295, 306 n.6, 313, 317, 320
ESTIMA ....................................................6, 245, 254
Gene Index Project ..................................292, 293, 296, 306 n.6
Nemagene .................................................6
NEMBASE ...............................................6, 7, 191, 213, 245, 306 n.6
PartiGeneDB ...........................................7, 191, 245, 254 n.1, 306 n.6
STACKkb ..................................................313, 314
TBestDB .....................................................6, 242, 243, 245, 246, 247, 248, 249, 253, 254 n.1, 306 n.6
UniGene ....................................................7, 119, 120, 306 n.6, 313, 316, 318
EST-Ferret, see EST processing software
EST mapping ..............................................314, 319
EST processing software .........................5, 158, 189–219, 243, 248
AnnotBr ..................................................191, 213
AutoFACT ................................................6, 243, 244, 249, 250, 251, 252
CLOBB....................................................192, 194, 208, 209, 254
Cross_match ..............................................192, 195, 198 n.5, 199, 201
d2-cluster ..............................................254 n.2
Decoder....................................................223, 226, 229, 230, 232, 234, 235, 237, 238
DIANA-EST .............................................223, 254
Dust .......................................................167, 199
EMBOSS ...............................................226
ESTAnnotator .................................6–7, 26, 130, 189, 196, 201, 213, 250–251, 257
ESTAP.....................................................252
EST-Ferret .............................................292
ESTScan2 ..............................................223
ESTWISE ..............................................223

**F**

False positive ..................................131, 167, 201, 278, 283, 284, 303
FASTA file format ......................184 n.12, 198, 200, 204, 206, 208, 209, 210, 217–218, 261, 262, 273, 292, 293, 299
First strand buffer .........................37, 54, 57, 72, 84, 132, 135, 157, 170, 293, 300
First strand cDNA synthesis, see cDNA libraries
Fragment size ...........................................43, 46
Frame shift errors ....................................237
Frequency of occurrence .......................110
FTP .......................................................166, 168, 169, 178, 180, 231
Full length enriched ..........................127, 130, 141 n.3
Functional annotation ......................6, 247, 249, 252, 257, 269, 270, 292
Funding ..................................................26, 27

**G**

GC10 competent cells ................................155
Gel electrophoresis ..................13, 39 n.9, 42, 44 n.1, 46 n58, 62, 91, 92, 102, 146 n.35, 171
Gel extraction kit .........................37, 41, 42, 46 n 14
Gene discovery .................2, 4, 7, 9, 13, 23, 49, 69, 109, 110, 119, 120, 130, 189, 277, 279, 281
Gene discovery rate, see Discovery rate
Gene finding .................................8, 130, 221

**H**

**I**

**K**

**L**

**M**

**N**

**O**

**P**

**Q**

**R**

**S**

**T**

**U**

**V**

**W**

**X**

**Y**

**Z**
Translation initiator................................................ 126, 130
Trans-splicing, see Splice leaders
Treeview.......................................................... 260, 270–271
Trehalose.........................................49–51, 54–55, 57, 62, 69–70, 72,
84–85, 170
134–135, 142–143
Trypanosomatids............................................. 127, 128, 129
Trypsin treatment ............................................................. 78
Tumor samples .............................................. 252, 317, 318, 320

U
UniGene, see EST databases
Uniprot..........201, 210–213, 217–218, 220, 223, 228,
254, 257, 293, 296
Unique ESTs, see Singletons
Uniref................................................................. 201, 210,
213, 228–229, 232, 250, 254
Unknown chemistry................................................ 199, 216
User path............................................................ 193, 198, 214
UTRs (Untranslated regions) ................................ 119, 125,
126, 127, 130, 222, 296

V
Vector trimming.................................................... 167, 200, 219
Visualization of EST data ........................................ 242

W
Washing buffer..................................................... 72, 80
Washing solution.............................................. 55, 61, 73, 88–89
Washington university genome sequencing center........ 28, 155
Web design.......................................................... 245, 253
Web interface................................................... 196, 241, 242, 243, 245, 246,
252, 253
Web server ......................................................... 211, 244
Web tools......................................................... 241–255
WellMate......................................................... 156, 159, 160

Y
Yield, measuring.................................................... 73, 86,
277–279, 286

Z
Zeldia punctata ....................................................... 271