3 Resources for transcriptome analysis

With the increasing number of completely sequenced genomes, the understanding of transcriptional control is one of the major challenges in current research. The first step towards this goal is to *identify* the transcribed regions of the genes. This chapter is about computational approaches firstly to identify these transcribed regions and secondly to evaluate the regions that relate specifically to a subset of cell-types. The resources described are microarrays, Serial Analysis of Gene Expression (SAGE) and Expressed Sequence Tags. Special attention is drawn towards our method of choice, the EST data.

3.1 DNA microarrays

Microarrays exploit the ability of an mRNA molecule to hybridize specifically to the DNA template from which it originated. Frequently, oligonucleotides or cDNA molecules corresponding to different transcripts/genes are spotted at different positions on the array. By measuring the amount of mRNA bound to each site on the array, a single experiment leads to the estimation of the expression levels of the respective transcripts/genes.

3.1.1 Hardware

DNA Microarrays are small, solid supports onto which the sequences from thousands of different genes are immobilized at fixed locations. The supports themselves are usually glass microscope slides but can also be silicon chips or nylon membranes. The DNA is either spotted or synthesized directly onto the support. It is important that the gene sequences on a microarray are attached to their support in an orderly fashion. This location of each spot on the array is used to identify a particular transcript/gene sequence.

3.1.2 Experiment

Microarray experiments are based on hybridization probing, a technique that uses nucleic acid molecules as mobile probes to identify sequences that are able to base-pair with one another. This technique is mainly applied for distinguishing the expression of a set of transcripts in two different environments. The mRNA samples isolated from the two different cell conditions are used as templates to generate fluorescence-labeled cDNA. Different fluorescent labels (red and green) are used for the two cell types for which the expression patterns are intended to be compared. The two samples are then mixed and incubated with a microarray immobilized with mRNA/oligonucleotide molecules. The labeled molecules bind to the sites on the array, in amounts roughly proportional the expression level of the transcript in the respective cell type. After the hybridization step, the microarray is placed in a *scanner* that consists of lasers and a camera. The fluorescent labels are excited by the laser, and the microscope and camera work together to create a digital image of the array. This digital image is then used to calculate the red-to-green fluorescence ratio. This ratio is an estimate of the relative expression levels of the respective transcripts in the two cell types with respect to each other. As an example illustration, Figure 3.1 shows the experimental microarray procedure to distinguish the expression pattern of two different yeast strains.

3.1.3 Applications

Microarray technology offers applications on both genomic as well as transcriptomic levels. Some of these applications are briefly described below.

1. *Gene Expression variations* This application of microarrays is on the level of transcripts. Microarray expression analysis determines the level at which a certain gene is expressed. In this case, the immobilized DNA is cDNA derived from mRNA of known genes. The cDNA libraries associated to a certain tissue/disease is then contrasted with another tissue/disease.

2. *Exon Based Chips* With the knowledge that alternative splicing is a frequent phenomenon, gene expression analysis is evolving into exon expression analysis. This is facilitated by the advancement of chip technology that allows for more spots in the same space. This enables the inclusion of probes for individual exons/ exon junctions on the array, thereby allowing the detection of alternative splice events. Using this approach, tissue-related or tumor-related differential expression patterns of individual transcripts (Johnson et al. (2003)) are analyzed.

3. *Tiling arrays* With the enhancement of chip technology, it is now possible to build microarrays with small overlapping sequences (around 20 nucleotides) covering the entire chromosome or parts of chromosomes. These microarrays called

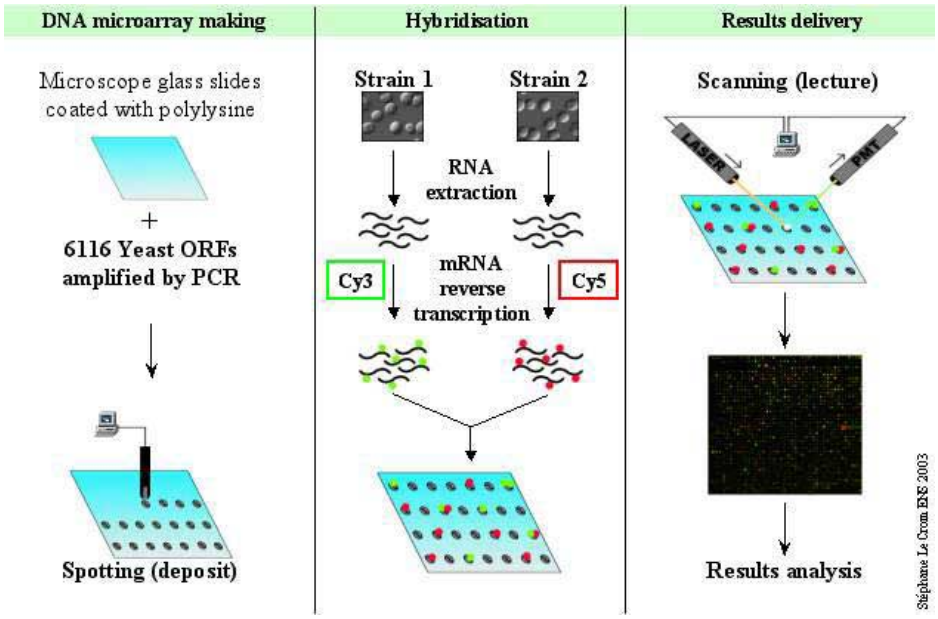

Figure 3.1: Illustration of a microarray experiment. The schematic shows the three steps of microarray analysis. In the first step, the yeast mRNA is immobilized on the microscopic glass slides. The next step consists of labeling cDNA derived from two different mRNA libraries with two different fluorescent dyes (red and green). This is followed by the hybridization of labeled cDNA libraries onto the spotted microarray. In the final step the red-green image of the microarray is scanned, digitally processed and evaluated for differential expression patterns. Image reproduced with permission from http://www.transcriptome.ens.fr

tiling arrays are interrogated with mRNA samples isolated from different cell types. The analysis leads to global detection of the transcribed parts of genome (Bertone et al. (2004), Schadt et al. (2004)). Another application of tiling arrays is to detect large parts of DNA that are lost or relocated due to mutations within the DNA repair genes (*Comparative Genomic Hybridization*).

3.2 Serial Analysis of Gene Expression (SAGE)

Serial analysis of gene expression (SAGE, Ryo et al. (1999)) is another experimental technique designed to gain a direct and quantitative measure of gene expression. Unlike microarray-based approaches, SAGE does not require prior sequence information. In SAGE, mRNAs isolated from a certain cell type are enzymatically processed to generate tags of a certain length corresponding to the 3' ends of the transcripts. Several such 3' tags are then concatenated and sequenced. Assuming no experimentally induced biases, the number of 3' tags per transcript correlate with the copy number of the respective mRNA in the sample. This leads to an estimate of expression levels of transcripts in the particular cell type. The methodology is briefly described below. For detailed description see the review by Yamamoto et al. (2001).

3.2.1 Methodology

The first step of SAGE (Figure 3.2) is to cleave the mRNAs using restriction enzyme *NlaIII* (anchoring enzyme) and bind them to streptavidin beads. The binding is accomplished via the poly A sequence on streptavidin beads corresponding to the 3' ends of the mRNAs. Therefore the bound sequence block would correspond to the sequence between the most 3' N1aIII restriction site and the polyA signal on the mRNA. After that the pool of bound mRNA fragments is divided into half and ligated to two different linker sequences which contain restriction site for another enzyme *BsmF1* (tagging enzyme). The tagging enzyme cleaves with the inclusion of 12 nucleotides (or 14 nucleotides using different set of enzymes, Ryo et al. (1999)). After a blunt end ligation step, the sequence between two linkers is amplified using primers corresponding to the linker sequences. Subsequently, the linker sequences are removed by again cleaving using the anchoring enzyme. The concatenated transcript containing parts of different mRNAs is then cloned and sequenced. Applying SAGE protocol, the elucidation of gene expression profile of a particular cell type is facilitated which would otherwise require several cloning and sequencing experiments.

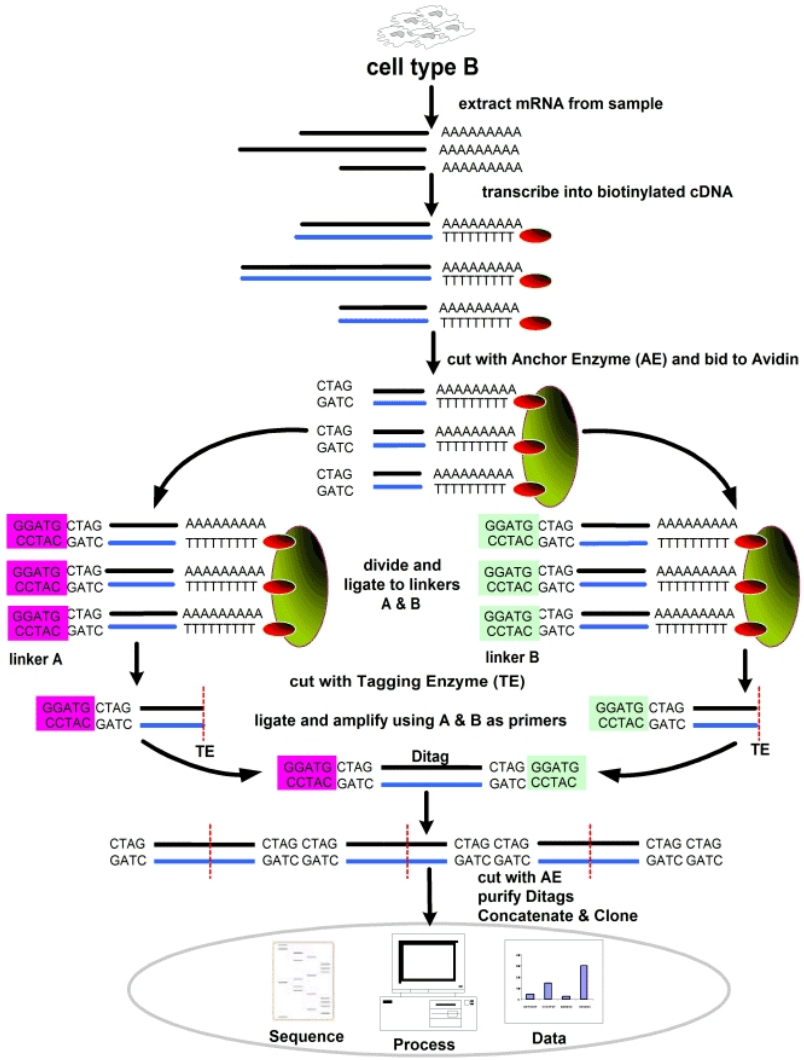

Figure 3.2: Schematic of SAGE procedure. The first step of SAGE is to cleave the mRNAs using restriction enzyme *NlaIII* (anchoring enzyme) and bind them to streptavidin beads. After that the pool of bound mRNA fragments is divided into half and ligated to two different linker sequences (A and B). The linker sequence contain the restriction site for another restriction enzyme *BsmF1* (tagging enzyme), which cleaves with the inclusion of several nucleotides. After a blunt end ligation step, the sequence between two linkers A and B is amplified using primers corresponding to the linker sequences. Then, the linker sequences are removed by again cleaving with the anchoring enzyme. The concatenated transcript continuing parts of different mRNAs is then cloned and sequenced. Image reproduced with permission from http://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Expression/exp82.html

3.2.2 Applications

SAGE procedure has been applied to study a variety of biological phenomena. Two categories of these applications are described below.

1. *Differential gene expression* SAGE technology has been widely applied to estimate the number of transcripts in different cell types. Chen et al. (1998) have analyzed changes in gene expression patterns in rat mast cells. Several of the differentially expressed genes revealed in that study are known examples for mast cells e.g. genes for preprolaxin, mitogen-activated protein kinase, etc.. Another similar study (Hashimoto et al. (1997), Hashimoto et al. (1999a), Hashimoto et al. (1999b)) involved comparison of SAGE profiles of human monocytes and their differentiated descendents, macrophages and dendritic cells which are often hard to distinguish pathologically. In this study, apart from several similarly expressed genes reflecting their functional similarity, some differential gene expression was also observed for genes like monocyte-derived chemokine, legumain etc..

2. *Cancer studies* Cancers exhibit differential expression of a large number of genes when compared to the normal state. SAGE procedure has been applied to study the characteristics of several cancer types like lung cancer (Hibi et al. (1998)), colorectoal cancer (Polyak et al. (1997)), colon/pancreas cancer (Zhang et al. (1997)).

3.2.3 Limitations

The main limitation of SAGE technology is that the procedure fails to differentiate between transcripts that are either alternative isoforms or are subject to alternative initiation of transcription (alternative promoters: Zhang et al. (2004)). The alternative promoters are now distinguishable using a recent technique called cap analysis gene expression (CAGE). CAGE uses cap-trapper full-length cDNAs (Carninci and Hayashizaki (1999)), to the 5' ends of which linkers are attached (Shibata et al. (2001)). The method allows high-throughput identification of sequence tags corresponding to 5' ends of mRNA at the cap sites Shiraki et al. (2003).
Another limitation of SAGE is the length of the tag, which (12-14 nucleotides) is insufficient to uniquely delineate all different 3' ends of mRNAs. Ambiguity in the mapping is further increased if the sequencing error rate is accounted.

3.3 Expressed Sequence Tags (ESTs)

An independent resource to analyze gene/transcript expression is the Expressed Sequence Tag (EST) data. ESTs are cDNA sequences (usually 200 to 500 nucleotides long) that represent parts of the expressed transcript. These are generated by isolating mRNAs from different cell types, tissues or organs. This section describes the experimental protocol of EST generation followed by a description of some of the EST-based resources.

3.3.1 Generation of ESTs

In a first step, mRNAs are isolated from the cell. Since the mRNA molecule is very unstable outside the cell, the enzyme *reverse transcriptase* (RT) is used to reverse transcribe the mRNA into complementary DNA (cDNA). Due to inefficient reverse transcription or lack of time for long mRNAs, the cDNAs are usually truncated to a few hundred nucleotides (Figure 3.3). These cDNAs are then cloned into a vector (host cell). The population of host cells containing the cDNAs is called a cDNA library. Subsequently, individual clones are randomly picked from the cDNA libraries and (partially) sequenced from either end to produce 5' as well as 3' ESTs. The frequency of picking a clone for sequencing is proportional to the copy number of the corresponding cDNA in the cDNA library. Therefore, this translates into a rough correlation between the number of EST sequences with the expression levels of the respective transcript.

3.3.2 Normalization of cDNA libraries

As described in the previous section, ESTs are generated by sequencing cDNA, which itself is synthesized from the mRNA molecules in a cell. Due to immense variations in the concentrations of various mRNA molecules, the likelihood of detecting lowly expressed transcripts is reduced considerably compared to the abundant transcripts. In order to facilitate detection of such rare transcripts the procedure of normalization (Bonaldo et al. (1996)) has been quite effective.
Figure 3.4 shows a schematic illustration of the procedure. The double stranded plasmid DNA is first linearized and used as a template for synthesis of RNA in vitro. In a parallel process, the same double stranded plasmid DNA library is converted to single stranded circles. The synthesized RNA and the single stranded circular DNA are then hybridized. This process is repeated and the number of repetitions along with the time for hybridization implicate the level of normalization performed. This

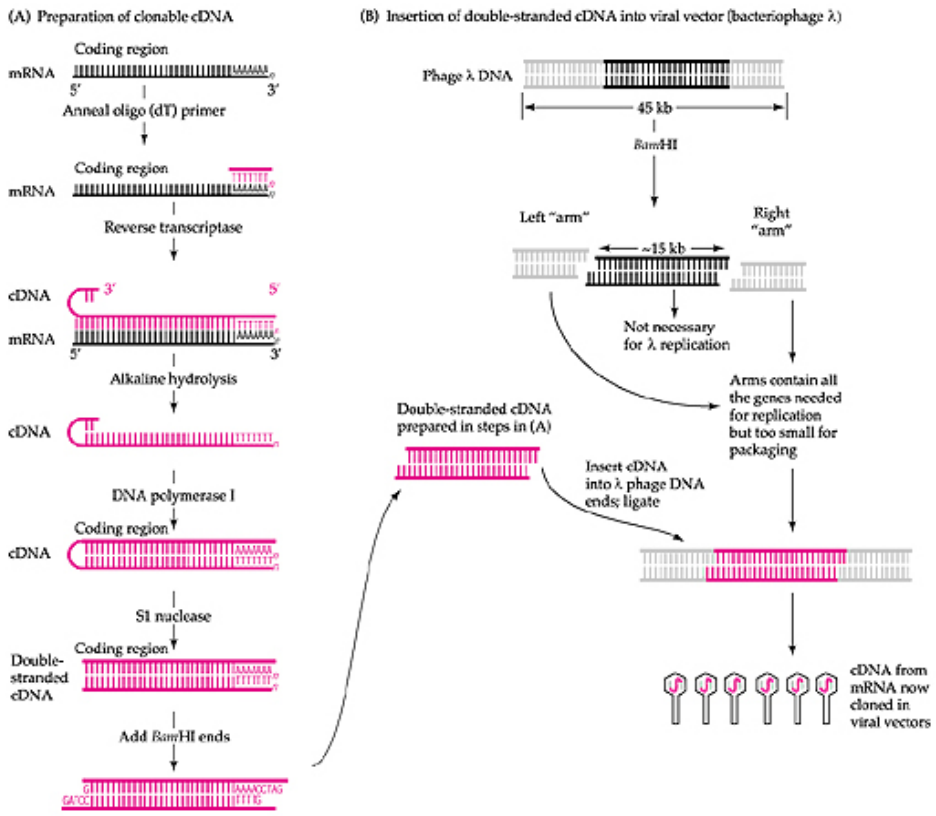

Figure 3.3: Generation of cDNA libraries from mRNAs. The figure illustrates the synthesis of (partial) cDNAs from mRNAs using the poly-T primers. Image reproduced with permission from Gilbert (2003).

is measured in terms of a Cot value, which is the product of cDNA concentration and hybridization time. For any particular transcript the hybridization between the corresponding synthetic RNA and the DNA is proportional to the concentration of the corresponding cDNA in the plasmid library. This leads to the enrichment of the rarer transcripts in the unbound cDNA. Extrapolating the inference to the EST data, the ESTs derived from such normalized libraries are likely to include rare transcripts that are not usually represented in ESTs derived from non-normalized libraries.

3.3.3 Clustering of the EST data

There has been a rapid growth in the number of ESTs generated (currently 4 million human and 3 million mouse ESTs) as well as the organisms covered (*dbEST*: Boguski et al. (1993)). The redundancy in EST sequences enables the clustering based on overlapping sequence elements, thereby grouping the ESTs belonging to the same gene. One of the main public resources providing the clustered EST data is the UniGene (Wheeler et al. (2003)). Quoting the National Center for Biotechnology Information (NCBI), *UniGene is a system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster is generated to contain sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.* Interestingly, several hundred thousands of the ESTs in the UniGene clusters represent uncharacterized genes. This emphasizes the need for detailed investigation of these clusters.
UniGene clusters have been evaluated by several research groups for a variety of gene-related studies. Combining the gene-based clustering of EST data with the cell type annotation of EST data leads to estimation of gene expression patters. As a further extension, groups of genes showing similar EST-based expression profiles across several tissues have been found to be a path of common pathways (Ewing and Claverie (2000)).

3.3.4 Assembly of the EST clusters: GeneNest

Although the UniGene clusters provide a handle on genes and their expression, they lack information related to different transcripts of the gene. It is now a known fact that via a process of *alternative splicing*, a single gene transcribes into multiple transcripts (Section 2.1.2). Since these transcripts are observable on the level of mRNA, this information is captured in the cDNA libraries and hence in the EST sequences. Assembly of the ESTs in a cluster splits the cluster into different contigs. Parts of these contigs differ from each other and potentially represent alternative splice events. The assembly of all ESTs as well as mRNA sequences is performed by the software GeneNest (Haas et al. (2000)). All cDNA/mRNA sequences related to an organism

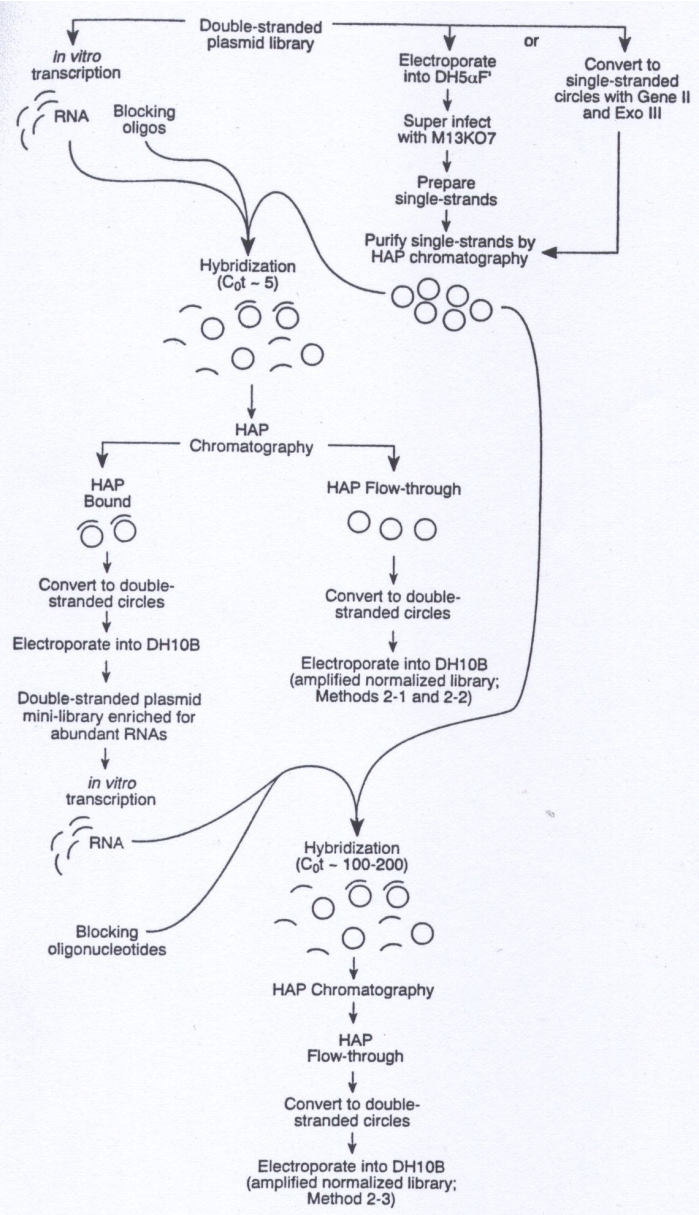

Figure 3.4: Normalization of cDNA libraries. The diagram shows the first description of normalization process using HAP chromatography. Double stranded plasmid DNA is linearized and used as a template for synthesis of RNA in vitro. In a parallel process, the same double stranded plasmid DNA library is converted to single stranded circles either in-vivo or in-vitro. The process of hybridization can then be performed with different levels of Cot that determines the reduction in the amount of more prevalent mRNAs. Image reproduced with permission from Bonaldo et al. (1996).

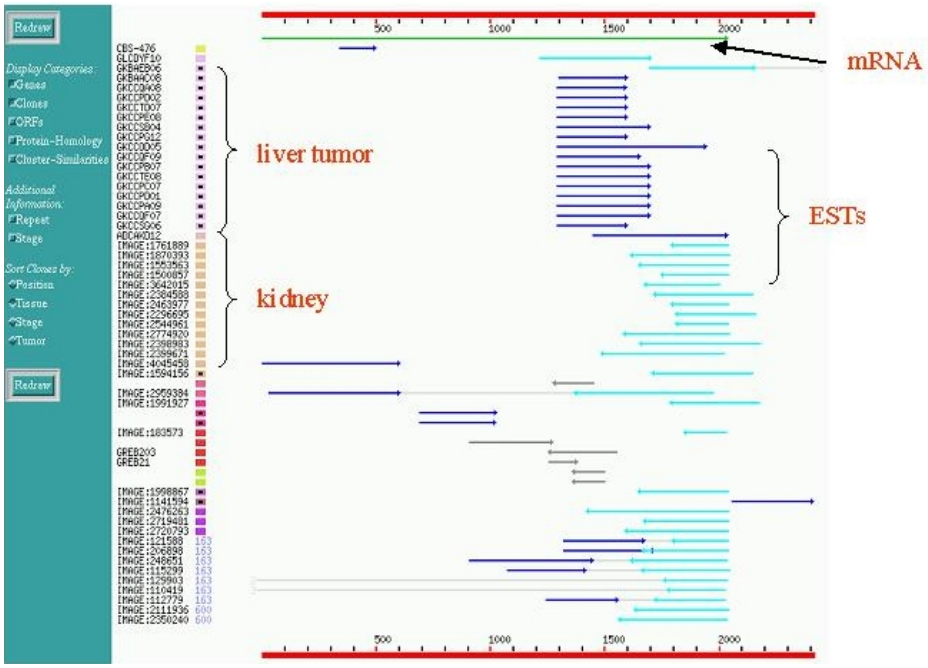

Figure 3.5: The GeneNest visualization. The figure illustrates the assembly of a single contig of a gene. The green arrow represents the mRNA and the other arrows represent EST sequences. The colored blocks on the left side represent the tissue annotation of each sequence, including the tumor annotation. The horizontal red bars on the top and bottom of the assembly represents the consensus sequence derived from the assembly of the mRNA and EST sequences. Image reproduced with permission from http://genenest.molgen.mpg.de

are extracted either directly from the EMBL (Stoesser et al. (2001)) database or from an already clustered UniGene (Wheeler et al. (2003)) database. A pre-processing step includes vector clipping, repeat annotation and marking of regions of low sequence quality in order to restrict processing to data of high quality. In further steps, these sequences are clustered and all members of each cluster are assembled into one or more contigs. A schematic view of an assembled cluster as presented on the GeneNest web site is shown in Figure 3.5. Currently, the database comprises of the gene indices of man (based on UniGene), mouse, Arabidopsis thaliana and zebrafish.

3.3.5 Genomic mapping of the EST data

The assembly of EST sequences in the GeneNest database separates alternative transcripts of a gene from each other. However, it is the mapping of the EST data to the genome that allows to differentiate between constitutive and alternative parts

of transcripts. The resources that map the EST data to the genome use different methodologies, some of which are described below.

1. *SpliceNest* The SpliceNest database (Figure 3.6) is a resource that deals with the problem of splice event predictions. It involves mapping of EST consensus sequences to the genome sequence assembly. The consensus sequences in this case are derived from the GeneNest database (Section 3.3.4).
   In SpliceNest all contigs derived from GeneNest are mapped to the chromosomes using a two step process. In the first step, a suffix tree based search program vmatch (upgrade of Reputer: Kurtz et al. (2001)) is used to infer the genomic regions of the potential mapping. For each cluster, one or more alignment regions are defined based on the matches of all its contigs.
   In the second step, for each cluster and each alignment region, all contigs are aligned to the corresponding genomic regions using the spliced-alignment-program sim4 (Florea et al. (1998)). The advantage of using a specialized gapped alignment program over the usual BLAST is that the inclusion of splice site definition improved the identification of exact exon-intron junctions.

2. *Alternative Splicing Annotation Project (ASAP)* The ASAP database also derives alternative splicing events by mapping expressed sequence tag data from UniGene (ftp://ncbi.nlm.nih.gov/repository/UniGene)to the genomic assembly derived from NCBI (ftp://ncbi.nlm.nih.gov/genbank/gbhtgXXX.seq.gz).
   As a first step, the clustered EST data is assembled to derive consensus sequences (Irizarry et al. (2000)). For constructing the consensus sequences, a maximum likelihood traversal of the EST-mRNA alignment is generated by using dynamic programming. These consensus sequences eliminate the minority features like sequencing errors, sequence differences due to paralog contamination, unaligned ends and inserts due to chimeric sequences or un-spliced introns. The genomic mapping of these consensus sequences is performed via two successive BLAST runs (Altschul et al. (1990)). The first high stringency BLAST is used to locate the genomic location of the gene which is followed by a low stringency BLAST to detect the exon boundaries (Modrek et al. (2001)). These (alternative) splicing predictions are embedded into a web-database, called Alternative Splicing Annotation Project (ASAP: Lee et al. (2003)). The web resource is accompanied by a graphical interface (Figure 3.7) for detailed investigation of individual splice events.

3. *Alternative Splicing Database (ASD)* The Alternative Splicing Database (ASD) (Thanaraj et al. (2004)), is another resource for identification of (alternative) splice events. There are two major differences that separate the ASD project from the ASAP/SpliceNest project.

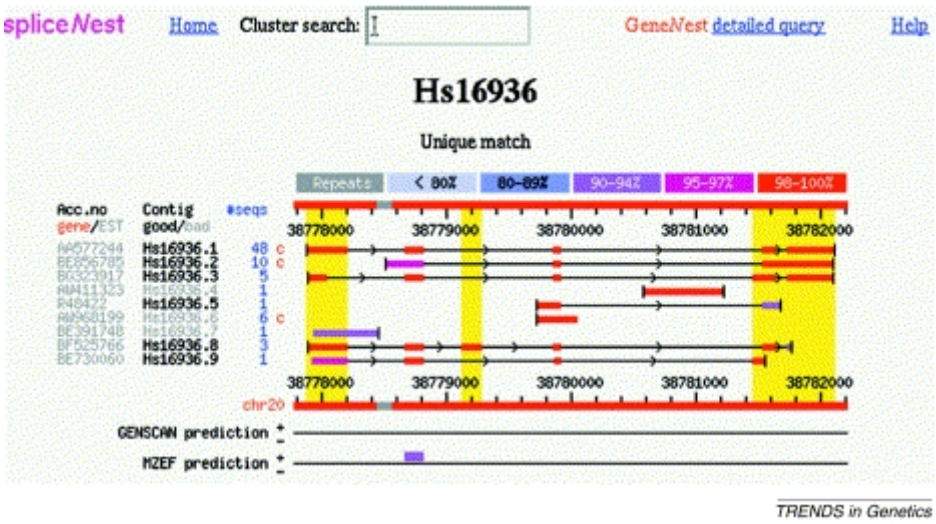

Figure 3.6: SpliceNest visualization of the gene structure. Every consensus sequence (contig) is represented in a row of colored boxes (exons) connected by black lines indicating the putative intron. Arrows symbolize introns bounded by the canonical splice-site consensus (GT/AG). Yellow bands indicate possible alternative splicing. Three different types of alternative splicing are highlighted for the unknown gene sketched in the picture: alternative donor/acceptor sites (left and right band) and an alternative exon. The thick red bar above and below the alignment represents the chromosomal sequence. Grey intervals indicate repeats. To the left of the alignment, information about each consensus sequence (e.g. accession number of a sequence contributing, name of the contig or number of sequences underlying a consensus sequence) is displayed. Below the bottom genomic sequence, predicted exons from *GENSCAN* and *MZEF* are shown for comparison. The colors reflect the confidence value supplied by each program. In this example, MZEF predicts only a single exon, GENSCAN predicts terminal exons slightly outside the chromosomal region visualized. Image reproduced with permission from Coward et al. (2002).

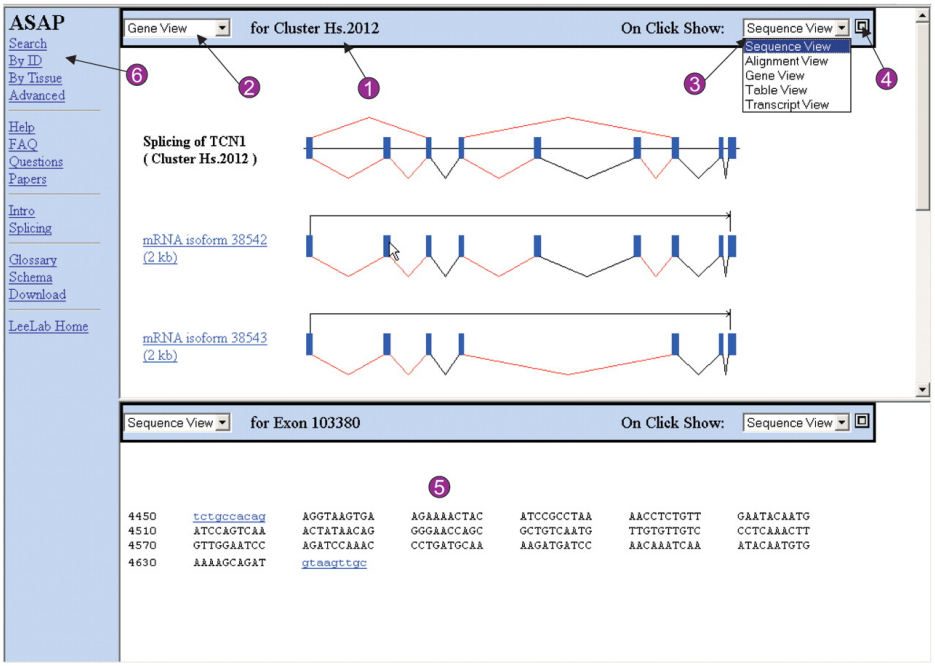

Figure 3.7: The ASAP's geneview. Each view has a title bar (tag #1); left menu (tag #2) that switches between various views for the current object; right menu (tag #3) that controls what view is shown when the user clicks on a feature; and maximize/split button (tag #4) that toggles between maximizing the view to fill the whole browser window, or splitting it into two views so the user can click on a feature in the upper view and see its detailed results in the lower view (tag #5). New searches, help, and additional information are available from the navigation bar (tag #6). Image reproduced with permission from Lee et al. (2003).

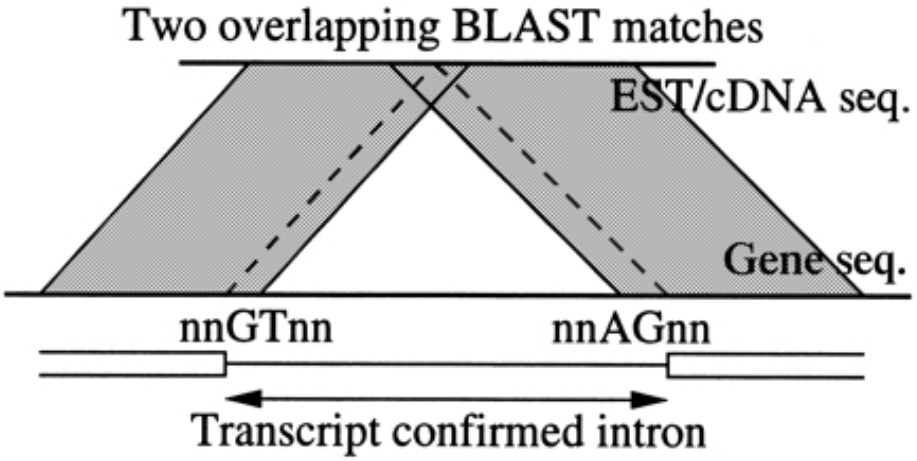

Figure 3.8: The ASD approach: transcript confirmation for introns. Identifying a transcript-confirmed intron by looking for transcripts covering an intron boundary in a gene-transcript alignment. Image reproduced with permission from Clark and Thanaraj (2002).

First the ASD project is generated by mapping the reference sequences (mRNAs) to the genes. The genes and transcripts extracted from Genbank (Benson et al. (2004)) are aligned under stringent conditions. Subsequently, only those introns are included for which both the 5' and 3' exons are confirmed by an EST or an mRNA.These are tagged as *transcript confirmed* splice events (Figure 3.8). Since the methodology involved mapping of transcripts to the genes and not the entire genome (as in ASAP), the ASD is intrinsically biased towards analysis of known genes (Clark and Thanaraj (2002)).

The second and perhaps the more distinguishing feature of the ASD project is that it includes a manually curated database of alternative splice events (AEDB: Alternative exon database, Stamm et al. (2000)). It is a collection of experimentally verified alternative splice events published in peer reviewed journals. The collected data includes the nucleotide sequences of alternatively spliced exons as well as the reported biological properties, including tissue specific expression, developmental regulation, alternative exon function and association with diseases. Figure 3.9 shows the graphical visualization of splice events in ASD project.

Although the three tools described use different approaches to derive this mapping, the final results may not be very different for most genes. However, all of the described databases might include contamination in the form of cDNA that correspond to un-spliced mRNA (pre-mRNA). This is turn results in sequences that look like genomic DNA. Furthermore, in the alternative splicing databases, sometimes incomplete alignments are discovered as alternative splice events. As a further development of

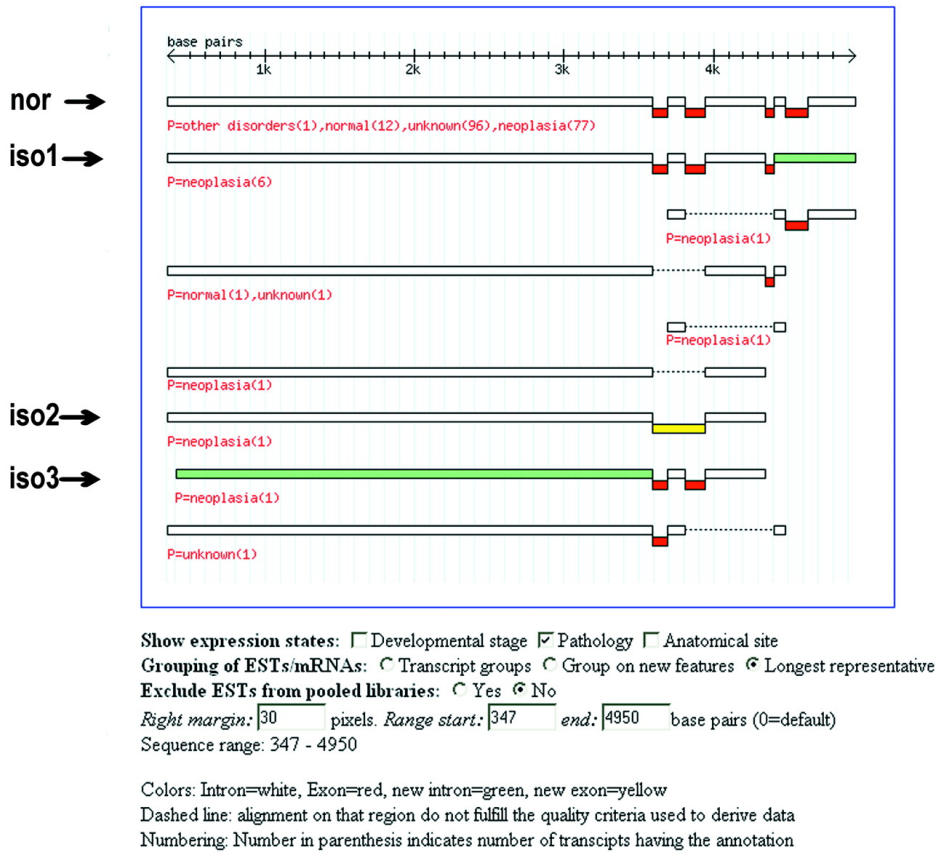

Figure 3.9: The ASD visualization of splice isoforms for the human C2F gene. Transcript sequences are grouped onto splice patterns, which are then displayed (color coding distinguishes introns from exons, and normal from alternative forms). The first pattern corresponds to the normal form of splicing as reported in the EMBL Database; patterns that are labeled iso1-3 correspond to the observed alternative splice patterns. The splice pattern labeled iso1 illustrates a skipped exon event, iso2 illustrates an intron retention event and iso3 illustrates an intron isoform event. Note that, for an exon to be shown in the view, both ends must have been confirmed. Transcript sequences can be grouped in three different ways. Each of the derived splice patterns is annotated for the expression state, with the annotation for a splice pattern obtained by consolidating those of its component multiple transcripts. Image reproduced with permission from Thanaraj et al. (2004).

the SpliceNest database, this thesis introduces computation of a reliability measure to these (alternative) splicing predictions (Chapter 4).

3.3.6 Expression levels using EST data

ESTs are generated by random sequencing of cDNAs which in turn are usually derived from mRNAs of annotated cell types. This tissue annotation used in conjunction with the clustered and assembled EST data links the expression informations to the genes and transcripts respectively.

Expression levels in GeneNest

The GeneNest resource provides the assembly of all ESTs as well as mRNA sequences. The description of ESTs in the database includes the categorization of the tissue-source into different tissue and tumor types. This annotation is used to compute significant expression of genes in different tissue categories in the form of p-values (Haas et. al., in preparation).
In a pre-filtering step, the tissue-distribution observed in a particular cluster is compared with the tissue-distribution across all clusters to derive significant clusters. For smaller clusters (less than 200 EST sequences), simulations are performed to compute the cluster entropy which is translated into a p-value. For the remaining clusters, a chi-square measure with Bonferroni's correction (Worsley (1982)) is used to derive clusters showing significant difference in expression patterns as compared to the overall distribution. For each of the significant clusters, the global distribution of EST clones per tissue is used to compute the likelihood of observing a given number of ESTs from a particular tissue in a cluster of given size. This is calculated using the function *pbinom* of the statistical package *R*.
The resulting p-values form a measure to describe genes that are significantly expressed in a particular tissue. Table 3.1 summarized the list of tissues with significantly expressed genes in GeneNest database.

Expression levels in SOURCE database

SOURCE (Diehn et al. (2003)) defines a set of relationships between different types of genomic data (viz. Ontology data, OMIM annotations, Swissprot annotations), together with the gene expression information derived from ESTs. For all Unigene EST clusters, an estimate of expression levels for all tissues per gene is computed in the following steps.
The fraction of number of clones of a certain tissue in the cluster and the total number

Tissue	Significant human genes (specific)	Significant mouse genes (specific)
adrenal gland	11 (5)	7 (0)
bone	11 (0)	9 (0)
brain	155 (45)	227 (46)
colon	7 (3)	93 (12)
ear	13 (1)	19 (2)
eye	49 (28)	74 (47)
gall bladder	6 (0)	
head	13 (1)	32 (3)
heart	24 (2)	36 (1)
kidney	20 (2)	166 (43)
liver	106 (26)	212 (66)
muscle	114 (27)	26 (1)
pancreas	66 (29)	82 (29)
pineal gland	5 (0)	-
testis	70 (37)	384 (281)

Table 3.1: Tissues with specifically expressed genes (via GeneNest database). The table contains a listing of all tissues for which some genes are significantly expressed. The numbers in parenthesis indicate the number of specific genes. Tissues like brain, liver and testis contain the largest number of tissues, while there are exceptions like kidney for which numbers are quite different.

of clones linked to that tissue is called an *expression frequency* for that tissue. Subsequently, a *normalized abundance* per tissue is calculated by dividing the expression frequency of that tissue by the summation of all the expression frequencies for that cluster. This normalized abundance score is used as a measure for deriving EST-based expression patterns in the SOURCE database.
The database also includes gene expression data derived from independent resources like microarray experiments. This enables comparing expression patterns of a gene across different experimental platforms. The relational definitions inherent in SOURCE allow detailed investigation of gene attributes together with their corresponding expression levels.

Tissue/tumor-specific transcripts in ASAP

The Alternative Splicing Annotation Project integrates the detection of alternative splice isoforms (Section 3.3.5) with the tissue-related annotation present in the EST data. This leads to the computational prediction of expression patterns in individual transcripts as opposed to all transcripts of the gene. The tissue-wise count of transcript-specific ESTs with respect to a random background distribution was attributed as the expression level (Xu et al. (2002); Xu and Lee (2003)). Transcripts that are significantly over-represented by ESTs derived from a single tissue are usually defined as being tissue-specifically expressed.
However, for the identification of statistically significant transcripts in ASAP, the effects of normalization of cDNA libraries is ignored. As discussed in Section 3.3.2, the procedure of normalization enriches the cDNA libraries for low abundant transcripts and therefore disturbs the correlation between the number of ESTs and the expression level. As a result, the computed statistical significance scores in ASAP are likely to overestimate specificity in the rarer transcripts.