6 T-STAG: An integrated portal for EST-based transcriptome analysis

This chapter presents the T-STAG (Tissue-Specific Transcripts And Genes) pipeline, our infrastructure for EST-based analysis of human and mouse transcripts. All relevant design issues as well as corresponding software implementations are addressed in this chapter. To elucidate the biological context, some example applications of the database have been discussed. A paper describing the T-STAG resource is in press (Gupta et al. (2005)).

6.1 The T-STAG database

As discussed in chapters 3, 4, and 5, there are several resources providing EST-based tools both for prediction of (alternative) splicing events as well as for the prediction of tissue-specific transcripts/genes. However, absence of a link between these related datasets limit biological applications.
Our resource and web-interface T-STAG, is a portal that integrates our alternative splicing predictions, the predicted expression patterns both of genes and individual isoforms as well as the man-mouse orthology relationships. In combination with the features for combining/contrasting the genes expressed in different tissues, T-STAG implicates important biological applications like the detection of differentially expressed genes in tumors, the retrieval of orthologs with significant expression in the same tissue etc. Additionally, our refined categorization of ESTs according to the normalization of cDNA libraries allows to search for putative low abundant transcripts.

6.1.1 Content

The following previously unlinked resources are integrated via the T-STAG database.

1. *Gene expression estimates:* The EST clusters (genes) and the annotation of EST libraries is derived from GeneNest database based on UniGene build 161 (Aug. 2003) for human and UniGene build 118 (Dec. 2002) for mouse (Wheeler et al. (2003)). The tissue distribution of ESTs in a cluster relative to random

background is translated into numerical estimates (p-values) of the likelihood of observing such a tissue distribution by chance (Section 3.3.6). Therefore, a low p-value for a given gene-tissue pair reflects either *significant* and/or *specific* expression of the gene in the respective tissue.

2. *Predicted (alternative) splice isoforms:* The constitutive as well as alternative splice forms are included together with the estimated confidence values for the predictions (Chapter 4).

3. *Tissue/tumor specific transcripts:* The GeneNest (Haas et al. (2000)) consensus sequences are mapped to the genome sequence (Human - April 2003 freeze of HUGO & Mouse - February 2002 freeze from the Mouse Genome Sequencing Consortium) and alternative isoforms are predicted with confidence values, using the EST coverage and splice signal indicators as a measure of reliability (SpliceNest: Coward et al. (2002), Gupta et al. (2004a)). Parts of these putative transcripts that are specifically covered either by ESTs related to a single tissue or only by ESTs derived from tumor related libraries are then labeled as tissue or tumor-specific splice events respectively (Gupta et al. (2004b), Chapter 5).

4. *Man-Mouse orthologs:* The human and mouse protein sequences are downloaded from the Ensembl database. Pairs of protein sequences with the best bidirectional BLAST alignment scores are defined as orthologs. The corresponding mRNA sequences are then inferred using TBLASTN of protein sequences with the respective reference sequences, thereby providing a link to the UniGene clusters.

These resources are integrated in a relational database infrastructure described in Section 6.1.2.

6.1.2 Database Design

The genomic alignment of GeneNest clusters stored in table `bounds` forms the heart of th T-STAG database (Figure 6.1). Only those clusters for which a genomic mapping is available are stored in the database. The alternative splicing information together with the corresponding confidence values is contained in the table `alternative_mod`. Table `expression` and table `tissue` contain the estimates of tissue-specific genes and tissue-specific alternative isoforms, respectively. In concordance with the table `tissue` another table `tumor` (not shown in schema) stores the data for tumor-specific alternative isoforms. Similarly, there are separate tables to delineate genes that are significantly expressed in normalized and/or tumor related libraries. All such combinations are stored in tables `expression_tumor`, `nor_expression`, `nor_expression_tumor`, `unnor_expression` and `unnor_expression_tumor`. The man-mouse orthology relations are stored as one-to-one mapping of unigene clusters for human and mouse.
Finally, the tables `est`,`info` and `symbol` store the EST mapping information, the EST

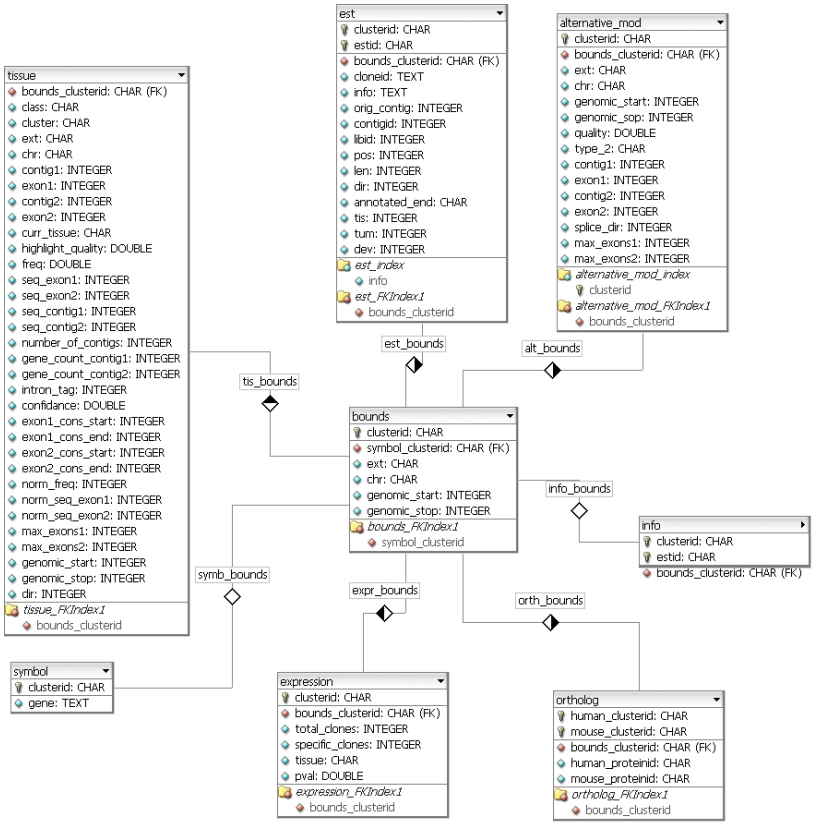

Figure 6.1: T-STAG database schema. Tables and main relations are shown. Tables `expression_tumor`, `nor_expression`, `nor_expression_tumor`, `unnor_expression` and `unnor_expression_tumor` are similar to the table `expression` and are therefore not shown due to space constraints. Another table `tumor` similar to table `tissue` is also omitted.

annotation and gene annotations respectively. These tables enhance the information content of the web output display.

6.1.3 Web-interface

The interface (Figure 6.2) is user-friendly and flexible with possibilities to define cut-offs (p-values for gene expression, quality values related to alternative splicing) based on individual applications. Additional restricted datasets based on individual applications can be generated by providing keywords and/or chromosomal location, thereby enabling queries like "*All kinases* expressed in human and mouse brain". The output of T-STAG is a batch of genes together with annotations, related tissues, etc.. This can either be downloaded as text or can be viewed as HTML. The HTML output provides tight links to the visualization tools, GeneNest (EST resource and visualization)

as well as SpliceNest (gene structure and alternative splice visualization), which allows a detailed inspection of candidate genes and transcripts (Figure 6.3).

6.2 Applications of the resource

6.2.1 Tissue-specific genes and splice isoforms

Tissue-specific regulation of gene/isoform expression is known to play critical functional roles as in case of many known genes like complement regulator CD46 (Russell et al. (1992)) or phosphodiesterase PDE7 (Bloom and Beavo (1996)). Additionally, tissue-specificity is also being associated with certain general mechanisms (tissue-specific RNA surveillance: Bateman et al. (2003)). Tissue-specific genes identified via the T-STAG database frequently include several known ones, as in the case of eye specific genes, in which 19 out of the top 20 have already been described to be functionally related to eye (Eg. rhodopsin, crystallin, opticin etc.). A similar evaluation performed for alternative isoforms also reveals several already known tissue-specific splice events among the top ranking matches. However, isoform-wise annotation is not as frequent as for entire genes. Nevertheless, most of the known genes containing putative kidney-specific transcripts are experimentally described to contain kidney related isoforms (*AFP*: Poliard et al. (1998); *SLC22A8*: Sweet et al. (2002); *WNK1*: Delaloy et al. (2003); *GLS*: Modi et al. (1991)). The so far un-annotated tissue-specific genes/isoforms with significant EST evidence need further investigation and possibly in-silico annotation.

6.2.2 Rare genes/alternative isoforms and disease related genes/isoforms

The EST data provides an estimate of the low-abundant genes/isoforms by the virtue of the differing protocols of EST generation. Due to the inherent over-representation of rare transcripts in normalized libraries (Bonaldo et al. (1996)), isoforms and genes that are represented only by such libraries are likely to be lowly expressed. This property of normalized libraries is utilized in the T-STAG database to filter out those transcripts that are likely to be lowly expressed. A large fraction of the tissue-specific alternative isoforms are observed to be such lowly expressed ones (Gupta et al. (2004b)). These low abundant transcripts may still have crucial functions as for one of the alternative isoform of gene *WNK1*, in which an alternative promoter controls the expression of a kidney-specific and kinase defective isoform (Delaloy et al. (2003)).

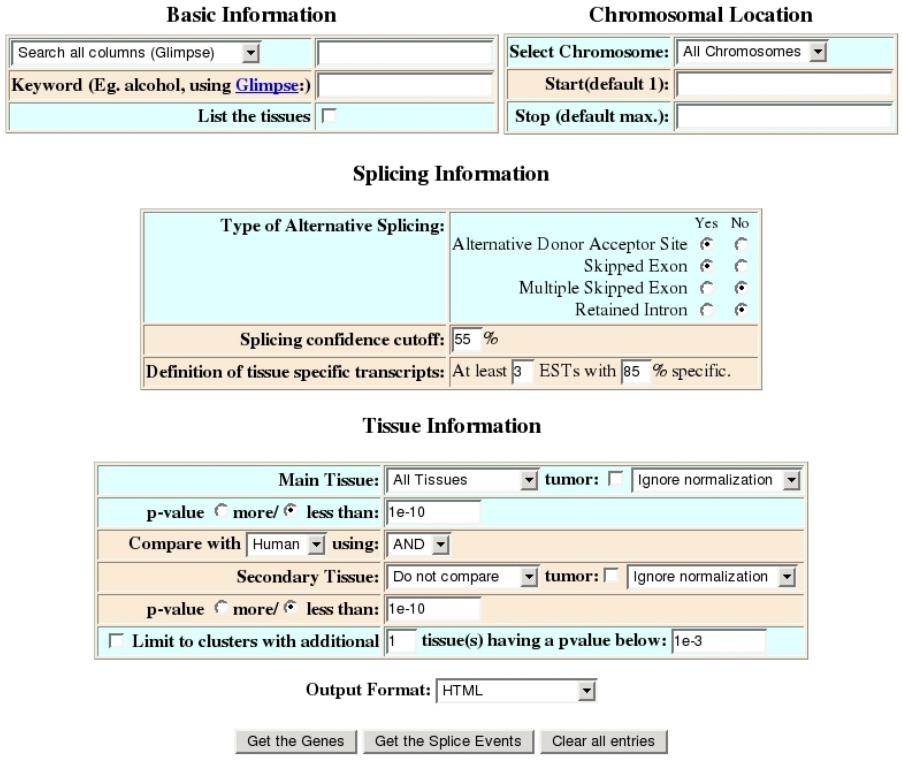

Figure 6.2: The T-STAG query interface. The interface is arranged in three main sections: *1) Basic information and Chromosomal location:* Various gene ids, accessions, keywords and chromosomal location can be specified. *2) Splicing information:* This can be used to select types of splicing or define a quality cutoff for alternative splice prediction depending on the particular application. *3) Tissue Information:* In this block the user can specify the tissues of interest. Information related to second tissue can be used to specify additional tissues in which the candidate genes *should (not)* be expressed. The organism can be switched in order to look at human and mouse orthologs. To limit the selection to specifically expressed genes, the number of additional tissues with significant expression can be restricted.
Finally, the output can be viewed as an HTML page or downloaded as a tab-delimited file.

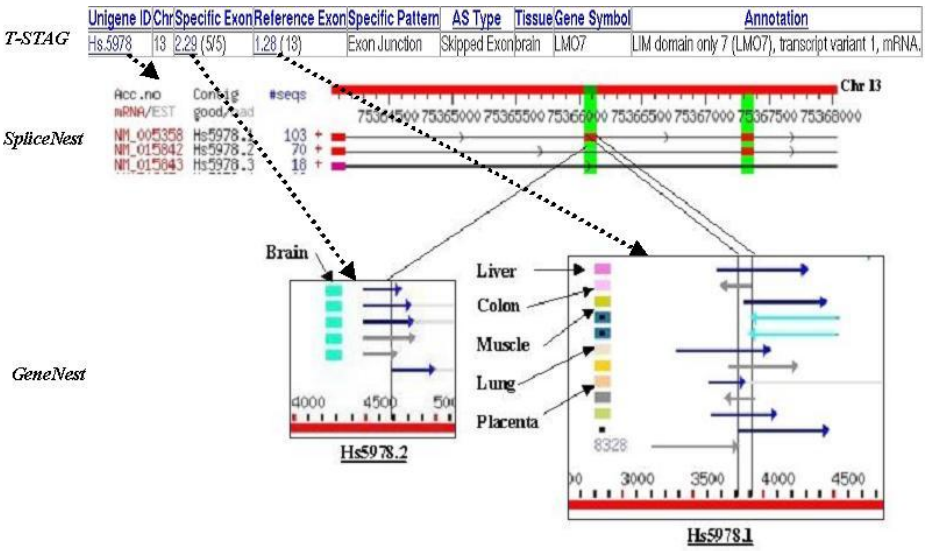

Figure 6.3: The hyper-linked output of T-STAG. The top part of the figure shows the T-STAG html output for the brain specific splicing of gene *LMO7*. The second level of the figure is a visualization of the gene in SpliceNest, showing parts of three transcripts with exons displayed as red blocks, connected by lines representing introns. The middle exon of the top transcript (Hs5978.1) is missing in the second transcript (Hs5978.2) and is therefore highlighted as an alternative splice event (green bar). The boundaries corresponding to this exon as well as the corresponding intron are visualized as vertical lines in the GeneNest database (3rd level: left and right box respectively). Both regions are covered by several ESTs depicted by horizontal arrows with corresponding tissues encoded in colored rectangles towards the left of each EST. Upon comparing the tissue distribution of these alternative regions it is evident that the middle exon of transcript Hs5978.1 is covered by ESTs derived from several tissues, while the corresponding exon junction that lacks this middle exon, in transcript Hs5978.2, is represented by ESTs derived from brain only, thereby revealing this as a brain specific splice event.

6.2.3 Comparison of expression patterns among genes

Due to our annotation of tumor and disease associated EST libraries, the T-STAG database also allows the retrieval of genes/isoforms that are *significantly* expressed in tumor or disease related tissues. However, in tumor cells an overall loss of control is observed in different parts of regulation machinery (Corn and El-Deiry (2002); Malumbres and Carnero (2003)), thereby leading to a large number of genes with abnormal expression levels. Therefore, in case of tumors only those genes are informative that show significant *differential expression* in tumors as compared to the normal cell types. In order to detect such *differentially expressed* genes, the predicted tumor-specific genes can be contrasted with another set of genes that are significantly expressed in the respective healthy tissue. This can be achieved by using the *subtraction* feature of the T-STAG database. Several of the top ranking genes revealed in this fashion are already known cancer related genes. In brain tumor for example, 6 of the top 10 genes have already been described to be tumor-associated. These include some genes which are suggested as tumor markers (*OLIG1*, Lu et al. (2001); *CRF*, Reubi et al. (2003)).
Alternatively, by using the *addition* feature of T-STAG, anatomically or functionally related tissues can be grouped together. Grouping heart and muscle reveals six genes with significant expression in both tissues. This set of genes include *titin* which is already known to play a critical role both for heart (Granzier et al. (2003)) and for skeletal muscle (Siebrands et al. (2004)). In addition, seemingly non-related pairs of tissues might also have biologically meaningful genes in common. In case of eye and pineal gland for example, we identified a group of genes (*CRX*, *OTX2* and *PDE6*) which are already annotated to be functional in both tissues (*PDE6*: Holthues and Vollrath (2004)). Furthermore, *OTX2* is a known transcription factor which regulates the expression of the gene *CRX* both in eye and in pineal gland (Nishida et al. (2003)), thereby hinting towards the existence of a common functional/regulatory pathway in these tissues. Some of the remaining genes in the dataset, most of which are currently annotated to be functional only in eye (viz. *RCV1*, *RTDBN*, potassium voltage-gated channel etc.) are therefore potential candidates that may be regulated by the same molecular mechanism.

6.2.4 Background definition for tissue-specific transcripts

With respect to the analysis of individual isoforms, the *addition* and *subtraction* features of the T-STAG database can be applied to categorize the tissue-specific isoforms into those that share significant expression in a certain tissue with other transcripts of the same gene and those that differ. The tissue-specific splice events observed in genes which are predicted to be specific to the same tissue are likely to be regulated via tissue-specific transcription factors (Odom et al. (2004), Pikkarainen et al. (2004)). On the other hand, a tissue-specific isoform detected in a gene that is expressed in

several tissues is possibly regulated by specific factors affecting splicing, in a tissue-specific manner (Hanamura et al. (1998)). We observe a large number of tissue-specific transcripts for both categories e.g. 187 human brain-specific transcripts potentially undergo specific alternative splicing while 91 specific transcripts are likely to be the consequence of common regulation of all transcripts of the respective gene.

6.2.5 Evolutionarily conserved expression patterns

The integration of orthology data with expression data enables the retrieval of *evolutionarily conserved* expression patterns in mouse and human. This provides an additional and more stringent schema for defining orthologs. Alternatively, the emergence of expression in additional tissues may reflect evolution of novel functions.
In addition, orthology relationships in T-STAG would facilitate integration with promoter annotation resources that use conserved non-coding blocks for detection of transcription factor binding sites (CORG: Dieterich et al. (2002)). This would help identifying sets of genes for which similar regulatory patterns lead to similar expression patterns.