5 Expression patterns of (alternative) transcripts

Every cell in an organism contains the same genomic identity. Expression of different subsets of these genes confer unique properties to each cell type. This diversity in cell types arising from a single genomic sequence points to a complicated regulation machinery controlled at multiple levels. Several genes have been described to exhibit differential splicing patterns for different tissues (E.g. *PDE1C* Yan et al. (1996); *IRF-3* Karpova et al. (2000)) that result either in alternative proteins or affect the regulation of the respective gene product Jin et al. (2003). This chapter focuses on the application of EST data to reveal specifically expressed transcripts corresponding to either of these mechanisms. This is followed by a comparison of the EST-based predictions with results from lab experiments.

5.1 Classification of cDNA libraries

The basis of our work is the tissue/tumor annotation of ESTs is GeneNest database (Section 3.3.4, Haas et al. (2000)) and the quality prediction of alternative splicing (Chapter 4, Gupta et al. (2004a)), visualized in the SpliceNest database (Coward et al. (2002)). The procedure of normalization is described below which is applied as a criterion for classification in our analysis. As described in Section 3.3, ESTs are generated from cDNA libraries corresponding to different tissues. Notably, a large fraction of cDNA libraries are constructed with a modified protocol called *normalization* (Section 3.3.2). This process of normalization reduces the ratio of abundant to rare transcripts which in turn facilitates the discovery of low expressed transcripts. However, this facilitation of gene discovery adds constraints on the EST data while conferring gene expression estimates based on EST counts. This problem is further complicated due to the different levels of normalized libraries available for different tissues. This difference in the levels of normalization introduces a bias in the EST counts which needs to be accounted for when analyzing tissue-related gene/isoform expression. Therefore, in our procedure the information related to the type of cDNA library is included. This leads to a better estimate of transcript/gene expression levels.

5.1.1 Methodology

The cDNA libraries of the GeneNest database are semi-automatically categorized into non-normalized, normalized/subtracted and PCR-based libraries by screening for the appropriate keywords in the original annotation of the respective EMBL database entries. All libraries for which none of the keywords are found are defined as being non-normalized. PCR-based libraries like those derived by ORESTES PCR are not used for the current analysis. Additionally, to avoid miscounting caused by PCR amplification, ESTs of the same library and with identical start/end positions in the alignment are treated as a single sequence. Since the level of normalization of different libraries may differ depending on the number of rounds of subtractive hybridizations performed, the normalization level is also extracted (measured as Cot or Rot: Sagerstrom et al. (1997)). This is limited by extent and clarity of the respective annotation entries. Increasing Cot-values hereby reflect the enrichment of clones derived from low abundant transcripts in the respective cDNA library. Besides the categorization of cDNA libraries according to the construction methods used, we further split these groups into libraries derived from healthy or disease tissue. Finally, ESTs of the four groups of cDNA libraries (healthy/non-normalized, healthy/normalized, disease/non-normalized, disease/normalized) are either analyzed separately or data of normalized and non-normalized libraries are combined.

5.2 Tissue/tumor-specific transcripts via GeneNest and SpliceNest

In the approach used by Xu et al. (2002) for detecting tissue/tumor-specific transcripts (Section 3.3.6), the effect of normalized cDNA libraries was ignored. Ignoring normalized libraries may affect the reliability of statistical estimation of tissue-specific expression levels. To resolve such biases, EST data related to normalized cDNA libraries are excluded from analysis in several computational approaches that aim at predicting tissue-specific expression (Megy et al. (2002); Schmitt et al. (1999)). In contrast, we propose to dynamically include or exclude the ESTs derived from normalized libraries. Our computational prediction is complimented with experimental validation of the predicted tissue-specific isoforms via RT-PCR across 40 tissue samples. The comparison of computationally predicted tissue specific expression patterns with experimentally derived expression patters allows the inference of the predictive potential of the EST data as well as the effect of normalized libraries on such predictions.

5.2.1 Prediction approach

Alternative splice isoforms in the SpliceNest database are revealed by aligning EST consensus sequences (putative transcripts) related to one gene to the appropriate genomic sequence. Significant differences in the boundaries of the putative exons are interpreted as alternative splicing events. Tissue-specificity is subsequently derived using the counts of ESTs per splice isoform.
For all exon-exon-boundaries that define a certain splice isoform the annotation of ESTs covering the respective boundary is evaluated. Isoforms overrepresented by ESTs from particular tissue are tagged as putative tissue/tumor specific splice isoforms. Several parameters (e.g. number of ESTs from a particular tissue, number of ESTs from other tissues, number of associated mRNA sequences etc.) are computed for these isoforms and finally stored in a relational database system. The refined set of tissue and tumor specific variants is then generated by setting the requirement of at least 3 ESTs in both alternative forms. Figure 5.1 describes such a prediction using GeneNest and SpliceNest visualizations. Since the counts of ESTs per tissue-specific splice event were frequently below 5, we considered it inappropriate to apply statistical methods as were used by Xu et al. (2002).
The predictions revealed 427 genes each contributing at least one potential tissue-specifically expressed variant. These variants show specificity for 28 different tissue types, where brain, testis and placenta account for approximately half of these transcripts (Table B.1). Many of these genes (n=210) exhibit isoforms that were exclusively detected due to ESTs derived from normalized libraries. These form a significant fraction (p-value: 8e-19) of the total genes that show tissue specific transcripts, since the number of ESTs derived from normalized libraries (896,645) is only 30% the total EST count (3,084,576) in tissues for which tissue specific isoforms exist. Similar prediction strategy was applied to predict the tumor-specific isoforms. The number of genes with transcripts exclusively expressed in tumors was relatively large (1120).

5.2.2 Experimental verification

A set of putative tissue specific (n=16) and disease-related (n=4) alternative splice events was arbitrarily selected for RT-PCR experiments. PCR primers were generated on the alternatively spliced exon as well as on either side of the event (Figure 5.2) using the primer design software GenomePRIDE (Haas et al. (2003)).
Subsequently RT-PCR experiments on 40 different tissue samples were performed by D. Zink at the German Cancer Research Center (see Appendix B for the experimental protocol and list of tissues). Gels were then manually examined for exact size, genomic contamination and the tissues in which the transcripts are observed.

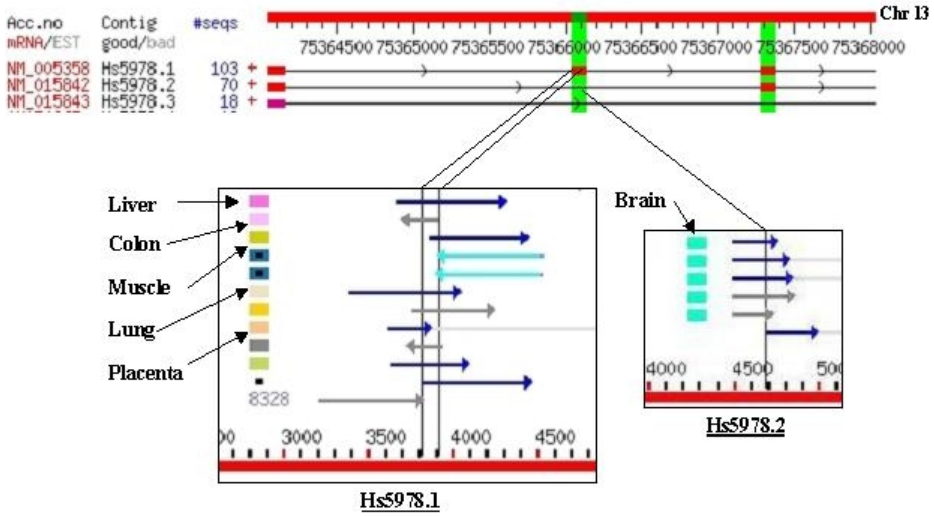

Figure 5.1: Detection of brain specific splicing in gene *LMO7*. The top part of the figure is a visualization of gene *LMO7* in SpliceNest, showing parts of three transcripts with exons displayed as red blocks, connected by lines representing introns. The middle exon of the top transcript (Hs5978.1) 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 (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.

Tissue	Variants (all)	Variants (normalized)	EST (all)	EST(normalized)
testis	100	72	106562	61837
brain	81	39	359489	177668
placenta	52	38	211830	107714
liver	34	0	137349	0
white blood cells	30	15	255381	114322
eye	23	9	171958	79794
pancreas	21	5	185558	12648
stomach	19	0	115672	0
prostate	14	13	120406	33429
kidney	11	8	137449	72555
lung	9	5	275104	135440
muscle	8	0	71895	2634
tonsil	7	0	18576	1324
skin	6	0	177156	0
adrenal gland	5	0	13819	0
heart	4	4	56901	36609
breast	4	4	118252	21791
uterus	4	3	218445	30445
development	4	0	12616	0
blood	2	0	12445	0
spleen	2	0	16014	0
fibroblast	1	1	14004	12393
pineal gland	1	0	6222	0
pituitary gland	1	0	8812	0
artery	1	0	16314	0
marrow	1	0	35408	0
ovary	1	0	93270	10918
nervous	1	0	117669	31143
Total	447	216	3084576	896645

Table 5.1: Tissues for which tissue specific transcripts are predicted. The table contains a listing of all tissues for which specific transcripts exist along with the number of ESTs related to individual tissues. Also, the ESTs derived from normalized libraries and the specific variants predicted via such ESTs are also listed.

5.2.3 Evaluation of tissue-specificity

Out of the 20 isoforms tested experimentally, 15 isoforms could be successfully verified in some tissue (Table 5.2). The remaining five variants are either likely to resemble rare transcripts according to the respective library construction protocol, or as in case of a disease-specific isoform (Hs.272688), the appropriate tissue sample was not available for experimental testing. Only four of the isoforms predicted based on the basis of normalized libraries could be validated using the standard RT-PCR conditions. For five additional isoforms a more refined protocol had to be applied in order to detect bands of significant strength. More sensitive PCR conditions frequently revealed expression in more tissues indicating low expression of the isoforms in these tissues. These results show the tendency of normalized libraries to be enriched for low-abundant transcripts. The predicted expression of the isoforms in a single tissue could not be confirmed for half of the variants analyzed (standard conditions). However, the isoforms were always detected to be expressed in the tissue that was originally predicted by our software. The observed expression pattern of the 'unspecific' isoforms ranges from expression in only a few, sometimes related tissues (LMO7 Putilina et al. (1998): brain, eye, testis, Figure 5.2; HRD1: brain, eye, thymus, salivary gland, kidney) to ubiquitous expression (MRPL42, ISGF3G). Those variants that were validated to be specifically expressed frequently originate from testis. Increasing the sensitivity of the RT-PCR revealed another testis-specific variant. At the same time the variants of the genes WNK1 and SCML1 were no longer defined as being tissue-specifically expressed since they were now also detected in a few additional tissues (Table 5.2: isoform 11 & 12). Consistent with previous work Gupta et al. (2004a) our approach of combining computational and experimental validation yields a high success rate in predicting the existence of splice variants. In line with the expected general enrichment of clones derived from lowly expressed transcripts in normalized cDNA libraries our experimental results confirm the expression of the predicted low abundance transcripts. Consequently, those isoforms that could not be validated experimentally may also reflect real biological signatures of extremely rare transcripts since they are often represented just by heavily normalized libraries (Cot 230, CIDE-A + Hs.48396). While the methods used in the construction of normalized libraries (PCR amplification, subtraction, size selection) increase the sensitivity of the detection of transcripts they unfortunately disturb the rough correlation between the expression level of a transcript and the observed number of related clones that is usually maintained in non-normalized libraries. Therefore, in these cases, the larger number of ESTs found for a specific transcript will profess to deal with a higher expressed transcript, also implying a higher confidence in the prediction although the sequences may be derived from the same although amplified clone.

#	Gene	Unigene	EST Evidence	ESTs	Cycles	Isoform	Specific	Comment (Most sensitive PCR)	Norm. Level
1	Unknown	Hs.112250	testis	3	39	+	+		
2	ISGF3G	Hs.1706	stomach	10	39	+	-	Ubiquitous	
3	MRPL42	Hs.112110	stomach-lymph	5	39	+	-	Ubiquitous	
4	SGN3	Hs.6076	testis	3	39	-	?		
5	PC326	Hs.279882	testis	9	39	+	+	testis (Tureci et al. (2002))	Rot-5
6	LMO7	Hs.5978	brain	5	39	+	-	brain,testis,eye(?) (Putilina et al. (1998))	
7	HRD1	Hs.274122	brain	3	39	+	-	brain,eye,3 others	
8	Unknown	Hs.24119	pancreas	4	39	+	-	approx. 10 tissues	Cot-20
9	BCLG	Hs.11962	testis	4	39,78	?,+	+,+		Cot-5
10	RBPMS	Hs.80248	placenta	4	39,78	?,+	-,-	Ubiquitous	
11	SCML1	Hs.109655	testis	12	39,78	?,+	+,-	approx. 6 tissues	Rot-5
12	WNK1	Hs.432900	kidney	3	39,78	?,+	+,-	Digestive system (Delaloy et al. (2003))	Cot-25
13	NY-CO-10	Hs.23557	testis	3	39,78	-,+	?,+		Cot-5
14	Unknown	Hs.169100	testis	3	39,78	-,-	?,?		Rot-5
15	Unknown	Hs.48396	breast	4	39,78	-,-	?,?		Cot-230
16	CIDE-A	Hs.249129	breast	4	39,78	-,-	?,?		Cot-230
17	KCNAB2	Hs.298184	tumor	29	39	+	-	Ubiquitous	
18	SNRP70	Hs.174051	stomach ascites	25/26	39	+	-	Ubiquitous	
19	RAB1	Hs.227327	tumor	39/95	39	+	-	fetal tissues, ovary (Wang et al. (2003))	
20	Unknown	Hs.272688	tumor	12	39,78	-,-	?,?	relevant tumor sample absent	

Table 5.2: RT-PCR validation results for tissue and disease-specific splice isoforms. RT-PCR validation results for tissue and disease-specific splice isoforms. The experiments are categorized into three groups viz. tissue specific isoforms predicted via ESTs related to non-normalized libraries (1 to 4), tissue specific isoforms predicted only via ESTs derived from normalized libraries (5 to 16) and disease-specific isoforms (17 to 20). For some of the variants represented by normalized libraries, standard PCR did not reveal the isoforms. However, five of these isoforms were detected using refined PCR conditions. The experiments frequently validated the isoforms and the tissue type, but the predicted specificity was rarely verified.

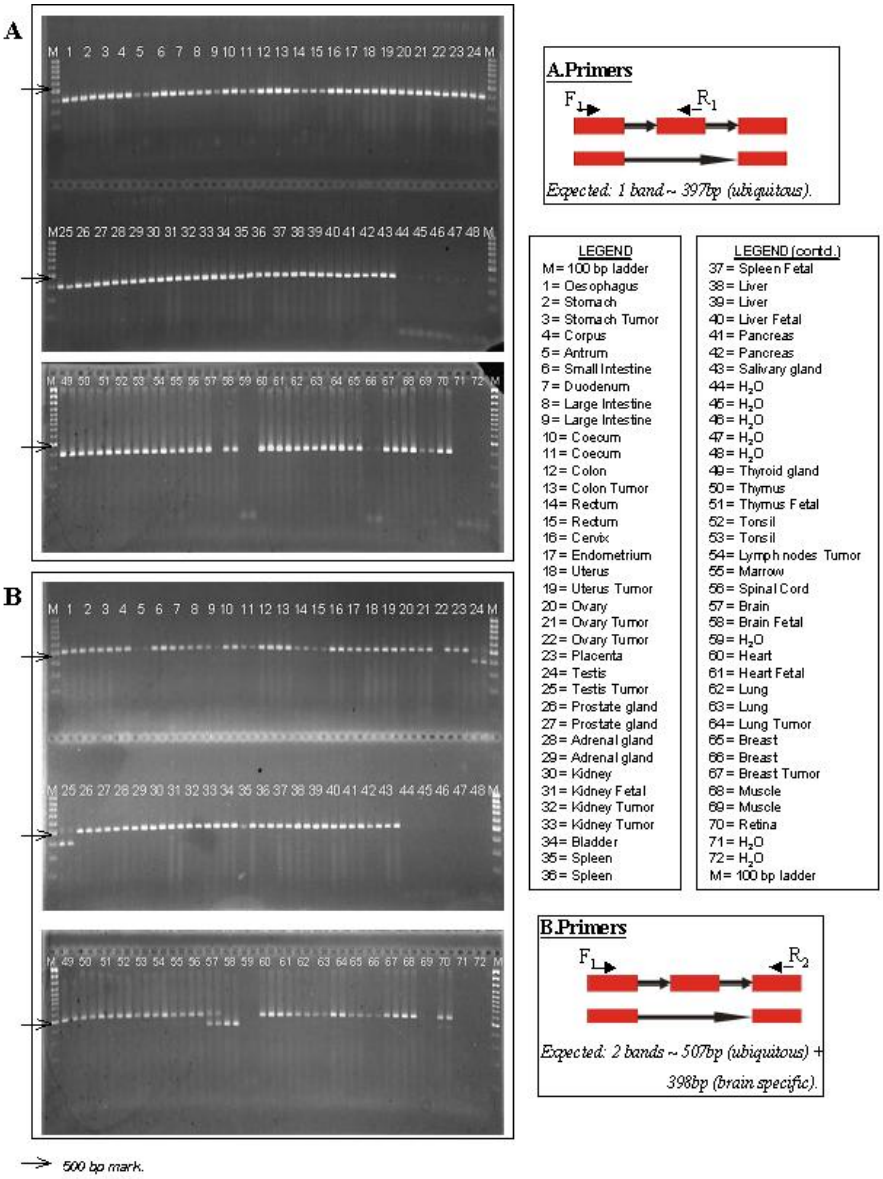

Figure 5.2: RT-PCR validation experiment of a putative brain-specific isoform.(A) The additional exon is detected in all tissues (primers F1,R1). (B) The primer pair F1-R2 located on exons flanking the extra exon results in two products where the shorter one is observed in brain, testis and eye (weak band). The predicted brain-specific expression pattern is, in fact, not specific.

5.2.4 Evaluation of tumor-specificity

For the disease-related transcripts also specificity was not observed in the experiments. Out of four such transcripts (Table 5.2: isoform 17-20), two were ubiquitously expressed although the large number of ESTs covering these variants suggested a high significance of the prediction. The tumor associated isoform described by Wang et al. (Wang et al. (2003)) was observed to be expressed in several fetal tissues along with ovary.
Therefore, in the context of tumors, our data shows that the predicted tumor-specific expression of isoforms derived from ESTs usually tends not to reflect the experimentally validated expression pattern. Rather it suggests expression in a collection of different tissues although the large number of related ESTs derived from tumor would imply a high confidence in the EST based prediction. Since tumor cells often show an up-regulation of a larger number of transcripts involved in various pathways (Corn and El-Deiry (2002); Malumbres and Carnero (2003)) the tumor-specific transcripts predicted based on the EST data may just reflect this general de-regulation of gene expression. The large number of predicted tumor-related isoforms further supports this hypothesis. Nevertheless, some transcripts detected via EST data may still serve as potential tumor markers like in case of the gene PRAME (Matsushita et al. (2003)) where the EST data as well as the experimental data suggests specific expression in testis and in a variety of different tumors (Figure 5.3).

5.3 Conclusions

Overall, ESTs are an extremely powerful tool to reliably unravel alternative transcripts independent of the level of expression. The functional relevance of the low abundant transcripts is not yet clear, especially if the isoforms do not affect the coding sequence. These isoforms may either be related to processes like nonsense-mediated decay (NMD: Hillman et al. (2004); Lewis et al. (2003)) or they might be some kind of non-functional leakage of the splicing machinery. Nevertheless, since many lowly expressed genes are already known to have important regulatory functions (Hao et al. (1994); Wieder et al. (1997); Geerlings et al. (2003)) this may also hold true for a not yet defined fraction of the alternative isoforms we detected via normalized libraries. In contrast to the prediction of the existence of isoforms, the task of predicting their expression pattern is much more error-prone since EST data always covers only a subset of potential tissues with variable sensitivity. The fuzzy terminology of tissue-specific expression that is frequently used to describe significant expression in a discrete tissue or a set of tissues, is therefore strongly biased by the sensitivity of computational and experimental methods (SCML1; WNK1: Delaloy et al. (2003)). Therefore computational prediction coupled with large scale experimental approaches as described by

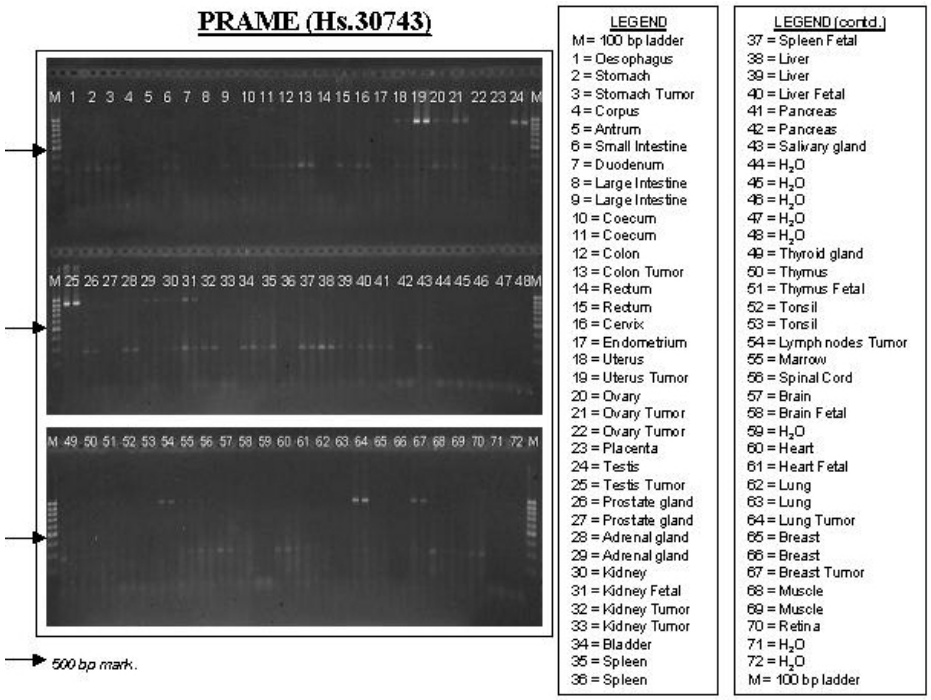

Figure 5.3: RT-PCR amplification of a 928 bp long region of gene *PRAME*. The corresponding bands are observed only for testis as the normal tissue along with several tumors (uterus, ovary, testis, lymph node, lung). In some cases a larger band is also seen, which corresponds to genomic contamination (checked by size of the band) in these tissue samples.

Johnson et al. (2003) are required for efficient delineation of tissue specificity.
Besides, the definition of specificity may also depend on the regulatory network that mediates tissue-specificity. While isoforms expressed in testis are specifically expressed in a more strict sense, other isoforms are expressed in a small set of (not necessarily related) tissues eventually pointing to alternative regulatory mechanisms acting with different stringency, e.g. involving transcription factors (Phiel et al. (2001), Naiki et al. (2002)) and/or DNA methylation (Ariel et al. (1991); Bergman and Mostoslavsky (1997)).