Chapter 4
Chapter 4

Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis


Abstract
The analysis of a human thyroid SAGE expression library shows the presence of an abundant SAGE tag corresponding to the mRNA of thyroglobulin (TG). Additional, less abundant tags are present that can not be linked to any other known gene but show considerable homology to the wild-type TG tag. To study if these tags represent TG mRNA molecules with alternative cleavage, 3’RACE clones were sequenced. The results show that the three putative TG SAGE tags can be attributed to TG transcripts and reflect the use of alternative polyadenylation cleavage sites downstream of a single polyadenylation signal \textit{in vivo}. By screening more than 300,000 sequences corresponding to human, mouse and rat transcripts for this phenomenon we show that a considerable percentage of mRNA transcripts (human 44%, mouse 22% and rat 22%) show cleavage site heterogeneity. When analyzing SAGE generated expression data this phenomenon should be considered, since according to our calculations 2.8% of human transcripts show two or more different SAGE tags corresponding to a single gene because of alternative cleavage site selection. Both experimental and \textit{in silico} data show that the selection of the specific cleavage site for poly(A) addition using a given polyadenylation signal is more variable than was until now assumed.
Introduction

Polyadenylation of eukaryotic mRNA is characterized by the cleavage of the precursor-RNA and the addition of a poly(A) tail. Polyadenylation of a primary RNA transcript is suggested to have a function in RNA metabolism such as export of the mature mRNA, stability and recognition by ribosomes. The signals that determine the site of polyadenylation have been studied extensively in in vitro experiments. The upstream element AAUAAA together with a downstream U/GU-rich element are present in nearly all eukaryotic pre-RNAs and are proposed to regulate the complex machinery of proteins necessary to complete 3'-processing; for review see [1, 2, 3, 4]. An extensive in silico analysis of 3'-processing control signals showed a general similarity to published experimental findings [5]. The site of cleavage in the pre-RNA is determined by the position of the regulatory elements together with the nucleotide composition of the cleavage region. This region is located 11-24 nucleotides downstream from a AAUAAA element and 10-30 nucleotides upstream from a U/GU-rich element. The suggested preferred nucleotide at the cleavage site is ordered A>U>C>G [6]. Although not emphasized in literature, in general it is assumed that there is, in a given pre-RNA sequence, one preferred cleavage site used by the polyadenylation machinery although mutational analysis shows that in in vitro cleavage experiments neighboring nucleotides can be used [7, 8]. Many studies have been done into the use of alternative polyadenylation signals in 3'RNA processing. Different polyadenylation signals in the 3'untranslated region of pre-RNA molecules can be used to generate different mRNA transcripts [9]. In sequence databases (e.g. GenBank, Unigene) mRNA reference sequences are represented by the most abundant transcript from every gene since this transcript will have been preferentially cloned. When looking at a large number of 3'EST (expressed sequence tag) sequences corresponding to these genes often differences can be found resulting from alternatively spliced or polyadenylated mRNA sequences. A genome-wide study shows that 29% of human mRNAs displayed two or more polyadenylation signals able to start polyadenylation [10]. However, the selection of the actual nucleotide at which the pre-RNA is cleaved using a given polyadenylation signal has been a relatively untouched subject. In this study we exclusively focus on alternative polyadenylation as a result of heterogeneous cleavage downstream from a single polyadenylation hexanucleotide signal. We therefore term this event alternative cleavage site selection to avoid confusion with alternative polyadenylation caused by the usage of different polyadenylation hexanucleotide signals. When analyzing the data from a human thyroid expression profile generated using a SAGE (Serial Analysis of Gene Expression) library [11] we found different SAGE tags corresponding to alternatively polyadenylated thyroglobulin (TG) mRNA molecules. The mRNA sequences found here have different polyadenylation cleavage sites downstream from a single polyadenylation signal. In the analysis of growing amounts of SAGE generated expression data this effect could mean that if a SAGE tag sequence is located in the cleavage region of a mRNA transcript the resulting tag corresponding to this transcript can be heterogeneous. We studied the incidence of this phenomenon in silico in 3 mammalian genomes (human, mouse and rat) by screening their corresponding Unigene databases. Unigene provides a unique data source for obtaining information about variation in cleavage sites. Unigene clusters consist of grouped GenBank sequences that belong to the same gene and which therefore also include the sequences of alternative cleavage sites. Consequently, analysis of the sequences in such cluster reveals possible cleavage site variation for that particular gene. Additionally, we checked whether alternative cleavage sites could be accounted for in vivo by sequencing 3' RACE clones. The implications for analyzing SAGE generated expression data are discussed.
Material & Methods

SAGE library
The expression library was constructed from human normal thyroid tissue according to original protocol [12] and as described previously [11].

3' RACE PCR
1st-strand RACE cDNA was synthesized from the same thyroid mRNA as used in the SAGE library. As a primer oligo-dT
t with a 3’linker 5’GCATGCAGAATTCTGGATCC was used [13]. The linker sequence was used as a reverse priming site in subsequent amplification of the TG 3’ region. TG-specific forward primer TG-2F 5’GAGAAGATCTCCTAAGGCCTC was designed according to TG cDNA sequence (GenBank accession no. U93033) generating a PCR product of about 200bp. PCR amplification was performed using standard conditions, 2mM MgCl2 and 1' 95°C, 1' 55°C, 1' 72°C, for 30 cycles.

Cloning of 3' RACE TG-fragments
TG 3'RACE PCR fragments were cloned using the pGEMT easy vector (Promega). Ligation reaction was transformed to E. Coli (DH5α) cells and plated. DNA from TG-positive clones was isolated using the Wizard Mini-Prep (Promega).

Amplification of TG exon 48
Genomic DNA was isolated from the same thyroid tissue as used for the 3’RACE experiments using Trizol (Gibco-BRL) following manufacturer’s protocol. 100 ng of DNA was used in a PCR amplification of exon 48 using intron-located primers with M13 linkers attached. TG48-for: 5’AGAGAAGTCTAATCTGGCTTG, TG48-rev: 5’CTGGTGCTAAGGCTGCTCAT (GenBank accession no. AF080484).

Sequencing
TG 3' RACE clones and the exon 48 PCR product were sequenced with the Dynamic Direct cycle sequencing kit (Perkin Elmer) using M13 forward and reverse priming sites in the vector or PCR fragment. Samples were run on an ABI377XL Automatic Sequencer (Perkin Elmer) and analyzed using Sequence Analysis 3.0 software.

Sequence screening
For the in silico determination of alternative polyadenylation cleavage sites the Unigene databases for human, mouse and rat were used. Every sequence in a cluster was scanned for a poly(A) stretch and a poly(A) signal in the 5'-3'(+) or 3'-5'(-) orientation. Sequences that contained a poly(A) stretch with a minimum length of 5 nucleotides and included the AATAAA or ATTAAA signal within 50 bp upstream were selected for further analysis. Unigene clusters containing less than four 3'EST sequences were excluded because the following analysis needs at least 4 eligible sequences to be able to calculate a realistic result. For all selected sequences of a particular Unigene cluster the partial sequences between the poly(A) signal and start of the poly(A) stretch were aligned to identify alternative cleavage sites. Since the length of these partial sequences was only about 18 nucleotides on average, we allowed one mismatch during the pair-wise alignment of these sequences to account for sequencing errors. In case of two or more mismatches the sequence which was less frequent was removed from the set. Multiple mismatches in sequence alignments easily occur as a result of ESTs that were incorrectly classified in a Unigene cluster. Finally, from the remaining set of partial sequences, sequences that occurred only once were discarded for this set to avoid scoring of sequence errors. From the remaining sequences the number of alternative cleavage sites were counted.
Results

In the thyroid SAGE data [11] the sequence tag corresponding to the thyroid-specific mRNA thyroglobulin (TG) presents itself as 5'CGGTGAAAAA and is scored 210 times out of a total number of 10,994 sequenced tags. This tag partly spans the poly(A) tail. Two tags not corresponding to any known human mRNA, 5'CGGTGAAGCA and 5'CGGAAAAAAA, are present respectively 54 and 9 times. Because of the similarity of these tags to the most abundant (wild-type) SAGE tag we opted to clone TG 3'RACE fragments to determine whether these tags correspond to TG mRNA.

<table>
<thead>
<tr>
<th>SAGE tag</th>
<th>Freq</th>
<th>Sequence 3'RACE clone</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tg-tag1</td>
<td>210(77%)</td>
<td><img src="image" alt="Sequence" /></td>
<td>45 (73%)</td>
</tr>
<tr>
<td>CGGTGAAAAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tg-tag2</td>
<td>54(20%)</td>
<td><img src="image" alt="Sequence" /></td>
<td>9 (15%)</td>
</tr>
<tr>
<td>CGGTGAAGCA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tg-tag3</td>
<td>9(3%)</td>
<td><img src="image" alt="Sequence" /></td>
<td>5 (8%)</td>
</tr>
<tr>
<td>CGGAAAAAAA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Alternative cleavage sites in thyroglobulin mRNA. Results from SAGE and 3' RACE are summarised. Sequence for each variant is shown together with their relative abundance. SAGE-tag frequency is out of a total of 10994 tags; 3' RACE-clone frequency is out of a total of 62 clones.

As shown in figure 1, sequencing of 62 TG 3'RACE-clones results in 4 different polyadenylation sites. 45 correspond to the wild-type TG mRNA, so 5'CGGTGAAAAA (TG-tag1), 9 clones have a polyadenylation site 4 bp downstream and correspond to 5'CGGTGAAGCA (TG-tag2) and 5 clones show a polyadenylation site 3 bp upstream, 5'CGGAAAAAAA (TG-tag3). Additionally, 3 TG cDNA-clones with a polyadenylation site 16bp downstream of the wild-type cleavage site are present showing a fourth polyadenylation variant that is also represented by TG-tag2. The extra 16bp TG 3'UTR in these 3 clones correspond to the sequence as published for the last TG coding exon [14]. The 150bp sequence upstream of the polyadenylation site is identical for all 62 clones and to the TG genomic sequence showing no additional polyadenylation signal sequences. To exclude any mutations in the pre-RNA sequence downstream from the poly(A) site putatively responsible for the variance in polyadenylation, the last coding exon of TG from the thyroid tissue used in these experiments was sequenced. The genomic sequence of TG in this particular thyroid gland is identical to the published TG exon 48 sequence (Figure 2).
Within this sequence the relevant regulating sequences have been indicated as well as the 4 different polyadenylation cleavage sites in human thyroglobulin mRNA. The in silico screening as described in Material & Methods using Unigene databases of human, rat and mouse resulted in 302,975 sequences which passed the criteria out of a total of 3,037,962 sequences in the database. The selected sequences are represented in 9,625 human Unigene clusters, 4,424 mouse Unigene clusters and 5,092 rat Unigene clusters. The number of alternative cleavage sites in these sets of genes are summarized in table 1.

### Table 1. Number of cleavage sites as scored in the screening of sequences in human, mouse and rat Unigene databases. The table shows the number of genes with their respective number of cleavage sites as used in the 3'UTR sequence downstream from a single polyadenylation signal (see Material & Methods)

<table>
<thead>
<tr>
<th>Number of cleavage sites per gene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>5430</td>
<td>2754</td>
<td>963</td>
<td>326</td>
<td>115</td>
<td>26</td>
<td>14</td>
<td>5</td>
<td>4</td>
<td>9625</td>
</tr>
<tr>
<td>Mouse</td>
<td>3438</td>
<td>824</td>
<td>127</td>
<td>23</td>
<td>8</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4424</td>
</tr>
<tr>
<td>Rat</td>
<td>3982</td>
<td>953</td>
<td>122</td>
<td>27</td>
<td>7</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5092</td>
</tr>
</tbody>
</table>
The percentage of genes which encode mRNA molecules using more than one cleavage site in the pre-RNA sequence downstream from a single polyadenylation signal is 44% for human genes, 22% for mouse genes and 22% for rat genes. From these genes the majority uses only one alternative site and the maximum amount of alternative cleavage sites used in a single gene for human, mouse and rat is respectively 9, 6 and 6. As a control we looked at the results of our in silico analysis in the thyroglobulin Unigene cluster. The 2 most frequent variants out of the 4 that we scored by sequencing 3'RACE clones are accounted for by this method. To get an impression of the region of the pre-RNA sequence in which this effect takes place the distance of the alternative cleavage sites relative to the most abundant and therefore arbitrarily called wild-type cleavage site was calculated. The distance was divided into up- and downstream nucleotides. The maximal distance is 32 nucleotides from the wild-type cleavage site, while the alternative sites are mostly in the immediate neighbourhood (Figure 3).

![Figure 3](image)

Figure 3. Distance of alternative cleavage sites relative to wild-type cleavage site in Unigene clusters with alternative polyadenylation. The distance is divided into the number of up(-) and down(+) stream nucleotides (X-axis). The most abundant cleavage site is used as a reference and is depicted as 0. Frequency is given on the Y-axis. Arrows indicate human, mouse and rat data.

In human, mouse and rat the majority of alternative cleavage sites is located between 8 nucleotides upstream and 9 nucleotides downstream of the wild-type cleavage site. The number of downstream cleavage sites is higher than that of upstream cleavage sites in all three organisms. For human, mouse and rat the percentages up/downstream are respectively 43/57%, 34/66% and 33/67%. To investigate how often alternative cleavage sites can cause problems in the analysis of SAGE data the number of SAGE tags that are located in an heterogeneous 3'polyadenylation area were counted. In 267 out of a total of 9,625 human genes alternative SAGE tags because of alternative cleavage can be distinguished. This result means that 2.8% of all human transcripts screened are represented by at least two SAGE tags. For mouse and rat these percentages are 1.3% and 1.2%, respectively. The lower percentages for mouse and rat are in concordance with the lower percentage of cleavage site heterogeneity in these organisms.
Discussion

The results from the cloning and sequencing of human thyroglobulin (TG) 3′RACE fragments (Figure 1) prove the in vivo existence of mRNA transcripts with alternative cleavage sites. In cloning 3′RACE fragments a representation of the TG mRNA pool is found concerning the region of interest, namely the polyadenylation region. The abundance of the four different TG mRNA molecules is similar to the abundance of the respective tags in the SAGE expression library. After examination of the downstream pre-RNA sequence present in the genomic sequence of the last TG exon 48, a consensus pattern concerning the wild-type polyadenylation [6] could be found (Figure 2). The AAUAAA signal is present and for the three most abundant polyadenylation variants the distance to the poly(A) site is within the defined consensus distance (11-24bp). As for the infrequent fourth variant, with 28bp the distance to the poly(A) site may explain its low abundance. Following consensus downstream of the cleavage site a putative GU-rich site is present 18-23bp downstream of the TG wild-type poly(A) site, while another is present between 39-44bp from the wild-type poly(A) site. It seems that the first putative GU-rich region (TGTGTG) would be the most likely candidate responsible for the three most abundant cleavage sites, since the distance to this GU-rich region is 13-20bp. The fourth variant is only 6bp from this region and probably too close. It may use the second putative GU-rich region located 27bp from its cleavage site. This second GU-rich region may have a lower preference for the polyadenylation machinery since the expression of this last variant is lower compared to the three most abundant poly(A) sites probably using the most upstream GU-rich region. The results from the Unigene sequence database screen for variation in cleavage site selection indicate that a considerable percentage of mammalian transcripts (human 44%, mouse 22% and rat 22%) are present in two or more variants reflecting alternative cleavage sites used for polyadenylation. The numbers in this study are probably underestimated because of the strict selection criteria imposed on eligible sequence and Unigene clusters. This very strict clone selection scheme avoids too many false positives in the analysis but simultaneously increases the number of false negatives. If we look at the thyroglobulin Unigene cluster, two out of four variants were scored in our analysis, because not enough sequences from the less abundant variants were available. We therefore estimate the actual percentage of alternative cleavage site selection to be even higher. There is a distinct difference in the percentage of cleavage site heterogeneity between human mRNA on the one hand and rat/mouse mRNA on the other (44% versus 22% and 22%). A possible explanation for this could be that there is a selection on the more abundant of the mRNAs. Because of the strict selection criteria in the genome screen, all alternative cleavage sites that were scored in only one sequence were discarded to avoid counting sequencing errors. Gene clusters with many 3′EST sequences would therefore be scored for alternative cleavage sooner than gene clusters with a low number of 3′sequences. Since the number of human 3′EST sequences in GenBank is considerably larger than that of mouse or rat this could explain the difference.

The implications for analyzing SAGE generated data are obvious. SAGE tags which are generated near to the region of heterogeneous cleavage can show the same heterogeneity. In our analysis 2.8% of human transcripts show two or more alternative SAGE tags. When analyzing a typical SAGE library containing 20,000 unique tags this means that about 550 tags will correspond to transcripts with alternatively cleaved mRNA molecules. Because this phenomenon is especially prone to show up in abundant tags, SAGE data will have to be screened for this. The mechanism behind polyadenylation and its biological implications is a subject of many studies. Most studied however are focussed on alternative polyadenylation
using different polyadenylation signals. This form of alternative polyadenylation can even be associated with regulatory properties [9]. The phenomenon of alternative cleavage site selection using a single polyadenylation signal is not emphasized in literature. It is generally assumed that this is highly heterogeneous in yeast [personal communication; N.J. Proudfoot]. In yeast it is known that in contrary to mammalian polyadenylation there is a greater lag-time between cleavage and poly(A) addition giving 3'–5' exonucleases the possibility to remove nucleotides from the 5' cleavage product, thus producing different polyadenylated transcripts [3]. In our study we show that there is a higher percentage of alternative cleavage downstream than upstream of the wild-type cleavage site. This indicates that differences in mammalian mRNA polyadenylation cleavage sites can not be attributed to exonucleolic degeneration of the uncapped 5’pre-RNA molecule. A recent study in yeast polyadenylation identifies a novel factor (Nab4p/Hrp1p) which seems to control the cleavage site selection process [15]. If the specific cleavage site used by the polyadenylation machinery to add the poly(A) tail is regulated by a specific factor in yeast, considering the homology between yeast and mammalian polyadenylation, this can be expected to happen in mammalian cells also. From our results it seems however that a considerable part of mammalian mRNA transcripts are not subject to strict regulation of cleavage site selection by any protein. In summary, it seems that there is considerably more variation in cleavage site selection in mammalian mRNA transcripts variable than until now was assumed. This has implications for the analysis of SAGE expression data.

Acknowledgements

We thank Joost de Gast for writing part of the software used for the screening of the Unigene databases.
References