3 PROMOTER AND GENE EXPRESSION PROFILING ON HUMAN CHROMOSOME 20 CYTOBAND q12-13.2

This chapter describes computational efforts on identifying promoter elements in a 10 Mb region of 20q12-13.2 and assesses the success rate of the current prediction programs. Besides, I also attempted to uncover novel features to further improve the efficiency of such programs. Finally, I present the expression profile of the genes in this region obtained with the Affymetrix Gene Expression Arrays in two different cell lines and correlate these results with the profile of the promoters obtained *in silico*.

3.1 Overview of the region

As described in section 1.5.1, the selected chromosomal region at 20q12-13.2 spans 10,099,058 bp and has, in brief, 103 protein-coding genes, 36 pseudogenes and 43 processed transcripts annotated in the Vega genome browser (version 19). Based on the current annotation, 54% (56/103) of the protein coding genes have more than one transcript, and many of these annotated alternative transcripts are based on alignments with expressed sequences (cDNAs and ESTs) and many either be incomplete or represent aberrant transcripts. In total, there are 402 different transcripts of which 35% (142/402) do not encode a protein. Out of the remaining 260 coding transcripts, 177 are controlled by different promoters assuming that two transcripts utilize the same promoter if their TSSs are less than 150 bp apart from each other. Analysis focused on this set. In all instances of multiple transcripts of the same gene using the same promoter, the transcript that has more biological evidence (cDNAs, EST, promoter or transcription site predictions) was included in selected dataset. The schematic representation of the selection procedure for the 177 transcripts is given in Figure 3.1.

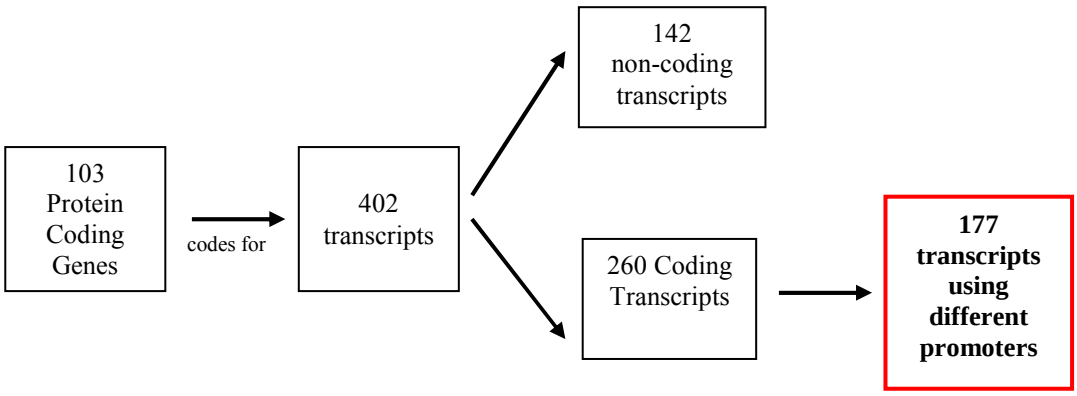

Figure 3.1 Schematic representation of the method for choosing 177 transcripts using different promoters.

The 177 selected coding transcripts are classified as representative transcript (RT) when they correspond to the longest 5' transcript, and alternative transcript (AT). It cannot be excluded that the annotation of both representative and alternative transcripts is incomplete.

3.2 Feature Predictions

The 177 candidate promoter regions (500 bp upstream and downstream of annotated TSS) were subjected to *in silico* analysis using CpG island, promoter and TSS prediction programs.

3.2.1 CpG islands

Promoter regions were searched for CpG islands using a program called "CpG Island Searcher" (Takai and Jones, 2002). Originally, CpG islands are defined as regions of greater than 200 bp with a %GC greater than 50% and an observed to expected CG dinucleotide (CpG) rate greater than 0.6 (Gardiner-Garden and Frommer, 1987). However, this software applies a set of more stringent conditions in order to exclude Alu repeats (see section 1.2.1). According to this program, 46% of the 177 promoters are associated with a CpG island. While 64% (66/103) of the RTs are associated with a CpG island, this percentage drops to 20% (15/74) for ATs.

CpG islands are also annotated in genome browsers such as Vega, Ensembl and UCSC and they all use the original definition of CpG islands. Vega and Ensembl use a program called "newcpgreport" (Micklem, 1999) to predict CpG islands. Newcpgreport finds CpG-rich regions of any length to find smaller CpG islands but it does not take into account the %GC of the region. Vega and Ensembl accept CpG islands found by this program only if their size and %GC are greater than 1000 bp and 50% respectively and the ratio of observed to expected number of CpG is equal or greater than 0.6. There are 97 CpG islands annotated in Vega Browser for the 20q12-13.2 and 75% (72/97) of them are associated with a TSS. Out of the selected 177 candidate promoters, 41% are associated with a CpG island according to Vega (Table 3.1). Three CpG islands annotated in Vega (associated with transcripts of GDAP1L1-001, GDAP1L1-006 and WFDC2-002) were not detected by "CpG island searcher" and both genes have restricted expression according to Unigene expression profiles (Wheeler et al., 2003). "CpG island searcher" detected 12 new CpG islands (10 out of 12 are associated with representative transcripts) and 75% of those transcripts are widely expressed according to Unigene expression profiles (Wheeler et al., 2003). That is in agreement with housekeeping genes being associated with CpG islands (Larsen et al., 1992). None of the newly added CpG islands contains ALU repeats and they are probably rejected by "newcpgreport" (used by VEGA and Ensembl genome browsers) since they are all smaller than 1000 bp.

The UCSC genome browser has more annotated CpG islands (118) in the region since it has a lower threshold for size (>300 bp) of a CpG island.

Type of Transcript	CpG island containing promoters		Total number of transcripts
	VEGA Genome Browser	CpG Island Searcher	
Representative Transcripts	58 (56%)	66 (64%)	103
Alternative Transcripts	14 (19%)	15 (20%)	74
All Transcripts	72 (41%)	81 (46%)	177

Table 3.1 Number of transcripts associated with CpG islands

Table A2 (Appendix A) lists CpG and GC content of a portion of the promoter regions (250 bp upstream and 50 bp downstream of TSS). A number of promoters (CDH22-001, C20orf142-001, ZSWIM3-001 and DNTTIP1-005) are associated with a CpG island despite their low CpG content since their CpG islands span downstream of their first exon.

3.2.2 Promoter Predictions

Promoter prediction in complex genomes is one of the most challenging tasks in computational biology for there are no known clear sequence signatures indicating its presence. PromoterInspector was the first prediction program with an acceptable sensitivity rate (Scherf et al., 2000). In a previous study, PromoterInspector was used to predict promoters in 20q12-13.2, where the program predicted 47.4% of the promoters correctly, producing 1.3 false predictions for every correct prediction, i. e. with a 43% sensitivity rate (Stavrides, 2002). Nowadays, several algorithms are available to predict promoters, nevertheless they all suffer from high false positive rates (Bajic et al., 2004). The program First Exon Finder (FirstEF) succeeds to predict a diverse set of promoters with a relatively low false positive rate (Bajic et al., 2004) (see section 1.2.6). Therefore, I chose FirstEF to predict promoter sites in the region.

FirstEF produced 198 predictions in the region and 68 (34%) of them are associated with the 5' of a coding transcript. The remaining predictions are not associated with 5' end of transcripts of protein coding genes. Interestingly, although 54% of the

promoter regions of RTs are associated with a FirstEF prediction, this percentage is only 19% for the promoter regions of ATs. In terms of its accuracy and sensitivity, FirstEF predicted 66% of the promoters correctly while producing around 2.4 false predictions for every correct prediction, which translates to 30% sensitivity.

FirstEF utilizes the compositional features of promoters such as presence of a CpG island, therefore it is relevant to explore the GC content of its predictions to see if there is any bias towards GC-rich promoters. Within the selected promoter set, 88% (60/68) of the FirstEF predictions were associated with a CpG island. The remaining eight FirstEF predictions overlapping with promoters not associated with a CpG island have a significantly higher GC content (p value <0.0005). This result shows the success bias of the predictor toward CpG-island associated promoters. Additionally, 42% of the false predictions are overlapping with internal splicing sites of the genes; this is also expected since first slice-donor sites are also used for predicting promoters by the program.

3.2.3 Transcription Start Site Prediction - Eponine

While promoter prediction programs aim to find “promoter regions” in the genome, they are unable to locate the exact start position of transcription. Eponine is developed to locate the TSSs in mammalian genomes by exploiting (i) CpG enrichments downstream of TSSs (ii) the presence of a TATA-box binding motif centred around 30 bp upstream of a TSS (Down and Hubbard, 2002). The model constraints are schematically described in Figure 3.2.

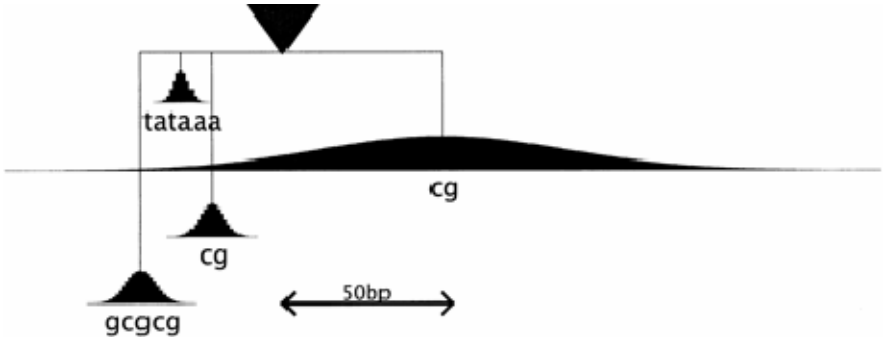

Figure 3.2. Sequence signals utilized by Eponine. Arrow head on the top marks the true TSS (taken from Eukaryotic Promoter Database (EDP) (Cavin Perier et al., 1998)). This figure is reproduced from reference Down and Hubbard, 2002.

Eponine predicted 267 TSSs in the region and 42 of them are located within 25 bp upstream or downstream of an annotated TSS. Out of these 42 predictions, 27 are associated with a coding transcript. As expected, all coding transcripts associated with an Eponine prediction contain a CpG island. Interestingly, only 4 of these predictions contain a TATA box binding motif (JASPAR model number M00980 (Sandelin et al., 2004)).

3.2.4 Overall Summary of Predictions

Out of the 177 candidate promoters, 23 are associated with all three prediction programs and 88 are not associated with any. Out of the 23 promoters with three predictions, 78% (18/23) correspond to a RT, while out of 88 promoters with no predictions, only 35% (57/88) corresponds to a RT. The prediction distribution over transcripts is shown in Figure 3.3.

Table A3 and Table A4 (Appendix A) list all the RTs and ATs in the region with the predictions associated.

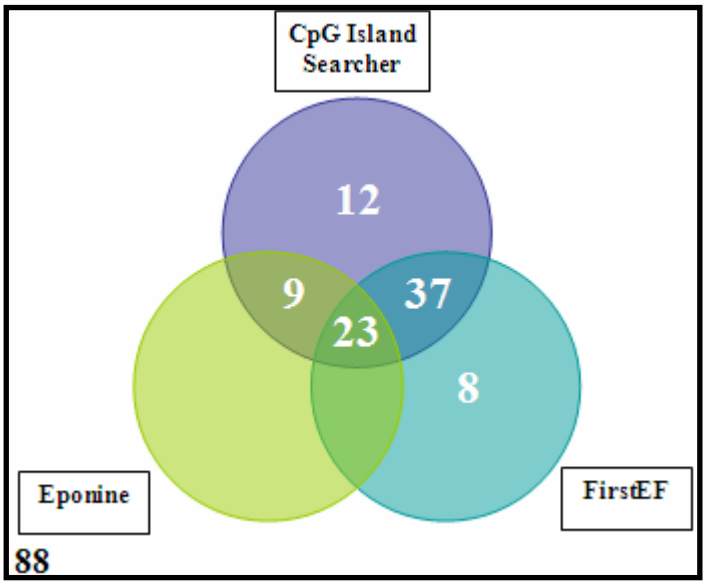

Figure 3.3. Number of promoters associated with predictions. There are 88 promoters not associated with any prediction.

In this study, promoters of 72 protein coding genes were detected either with the help of a CpG island or a promoter or TSS prediction program, but only 6 out of these 72 promoters were not associated with a CpG island. Promoters of 31 protein coding genes were not detected by any prediction programs; none of them is associated with a CpG island. These 31 genes have tissue specific expression according to Unigene Expression Profiles (Wheeler et al., 2003). These results clearly show the bias of the current promoter prediction programs towards CpG island associated promoters. Certainly, these prediction programs require additional signals to be able to detect promoters not associated with CpG islands.

3.3 Sequence and Structure Topology of Promoter Sequences

3.3.1 Sequence Topology of Promoter Sequences

Promoter sequences contain binding motifs which direct the recruitment of the trancription machinery onto it. The presence of these motifs may well set specific constraints on the sequence of promoters. Therefore, I investigated the frequency of nucleotides at each position of the promoter sequence relative to the TSS. To this end,

sequences of extended core promoters (100 bp upstream and 50 bp downstream of TSS) are aligned relative to the TSS. Then, the number of each nucleotide, represented by B (B → [A, C, G, T]), in each position ($f(B)_r$, Equation 3.1) is counted in all sequences as it is illustrated in Figure 3.4.

Figure 3.4 Schematic representation of calculating frequencies of each nucleotide at a given position (r) in N number of promoter sequences. The number of each nucleotide is counted at a given position (r) along the sequences (denoted by red box) to obtain the frequency of the nucleotide at that position.

$$f(B)_r = \sum_{i=1}^{N} g(B)_r \ \textbf{where} \begin{cases} \textbf{if} \quad (\mathbf{B_r})_N = B, & \mathbf{g(B)_r} = 1 \\ \textbf{else} \quad , & \mathbf{0} \end{cases} \qquad \text{Equation 3.1}$$

$f(B)_r$ → number of nucleotide B on position r

B → [A, C, G, T]

r → position in the sequence [1..R] where R is the sequence length and

N → number of sequences

Nucleotide frequencies are normalized by subtracting the observed frequencies of each nucleotide at any random site in the human genome (denoted by O(B), Equation 3.2). These observed frequencies of nucleotides A, C, G and T at any random site in human genome are 30%, 20%, 20% and 30% respectively (Wu et al., 2005).

$$F(B)_r = \left(\frac{f(B)_r}{R} \times 100 \right) - O(B) \qquad \text{Equation 3.2}$$

$F(B)_r$ → normalized frequency of nucleotide B on position r

O(B) → observed frequency of nucleotide B in the genome

Also, nucleotide frequencies in ten adjacent positions are averaged to smooth out the local fluctuations.

Figure 3.5 shows the percentage deviation of each nucleotide from its randomly observed frequency at each position in the sequences of the 177 putative promoters under investigation.

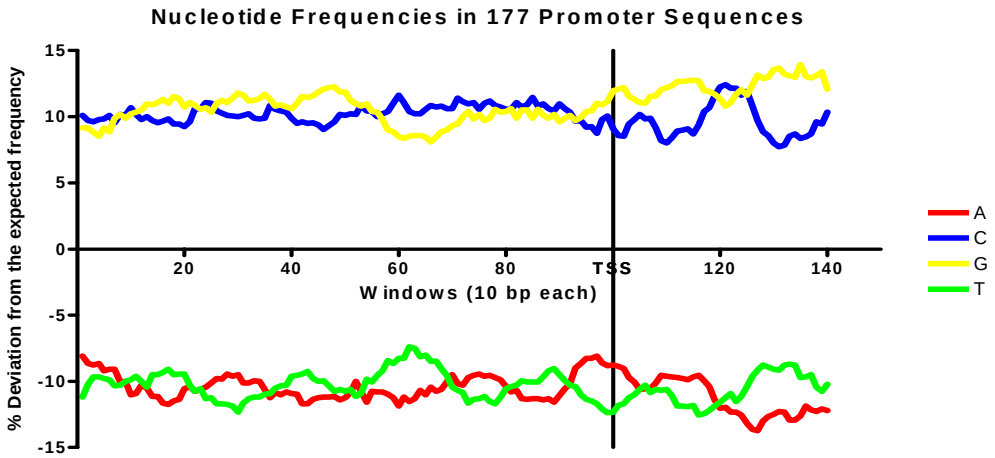

Figure 3.5. Frequency plots of each nucleotide relative to their distance to TSS in 177 promoters. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. Solid black line at x=100 marks the TSS.

Figure 3.5 shows that nucleotides C and G are observed around 10% more than their expected frequencies whereas there is a 10% drop in observed frequencies of A and T nucleotides in promoter sequences.

Promoters can be categorized into two broad classes; TATA-box enriched promoters, which have relatively well-defined initiation sites, and CpG rich promoters (Carninci et al., 2006). The 177 putative promoter sequences were divided into two groups according to presence or absence of a CpG island. For illustration purposes, I summed

the normalized frequencies of C and G nucleotides and then plotted for each class the summed nucleotide frequencies (Figure 3.6).

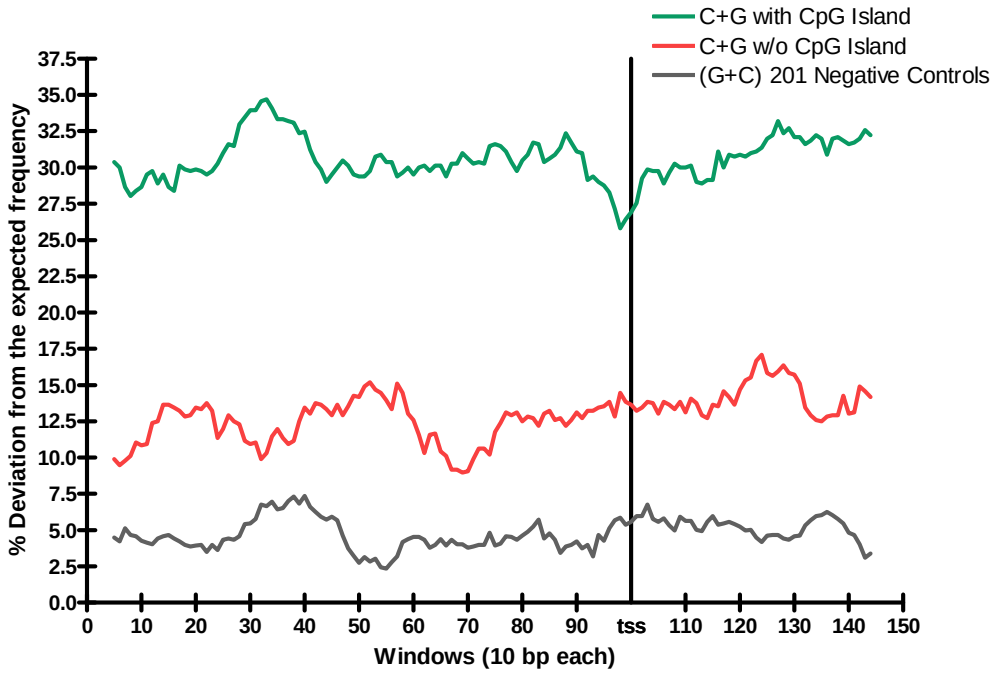

Figure 3.6. Summed nucleotide frequencies in promoters with, without CpG islands and 201 negative controls. Solid black line at x=100 marks the TSS. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations.

As expected, CpG island containing promoters have much higher GC content (30.6% ± 1.6%) as opposed to promoters without CpG islands (12.9% ± 1.7%). Both groups have however higher GC content compared to randomly selected negative controls (201) (4.8% ± 1.0%) as shown in Figure 3.6.

I also profiled 75 bp upstream and downstream of 3' UTR end of the 177 selected transcripts to investigate GC enrichment of such sequences (Figure 3.7).

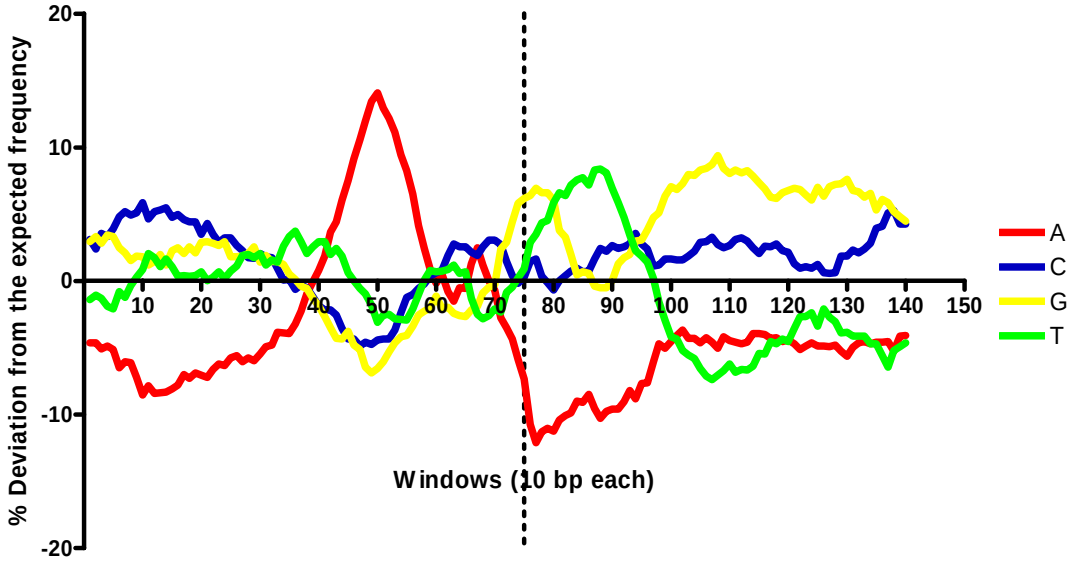

Figure 3.7 Frequency plots of each nucleotide relative to their distance to the 3' end of the 177 promoters. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations. Solid black line at x=75 3' end of the transcript.

As shown in Figure 3.7, the sequence profile of the 3' ends of the transcripts are different than that of promoter sequences. There is no C:G enrichment across the sequences and different peaks are observed; a strong A peak at around 50 bp upstream of the 3' end of the transcript and G peak at around 30 bp downstream of the 3' end of the transcript.

3.3.1.1 Sequence Topology and Prediction Programs

Promoter and TSS prediction programs seek for specific sequence motifs such as CpG islands, TATA-box or initiator sequence. In the investigated promoter dataset, 44% (77/177) of the promoters are associated with either a promoter (FirstEF) or a TSS (Eponine) prediction. A+T and C+G nucleotide frequency plots of the two subgroups i.e. promoters with or without associated predictions are shown in Figure 3.8.

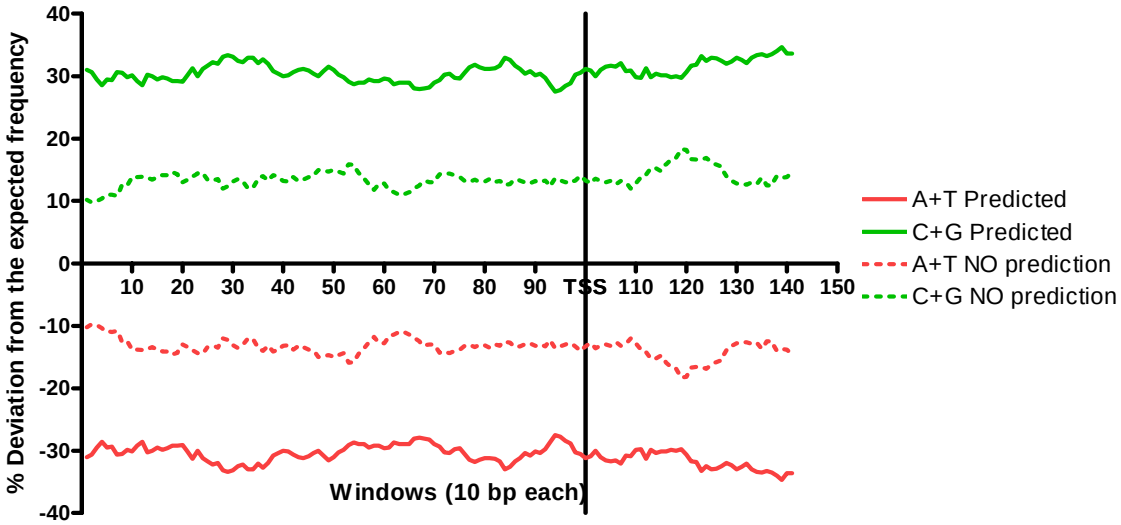

Figure 3.8 Summed nucleotide frequencies of 77 promoters associated with at least one prediction (dotted lines) and 100 promoters not associated with any prediction. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations.

Promoters predicted computationally are very GC-rich compared to the other subgroup. This is expected since 90% of the promoters with a prediction are associated with a CpG island. The similarity of the nucleotide frequency profiles between the two subgroups provides supports that the sequences without predictions are most likely false negatives.

A promoter with a CpG island has a higher chance to be predicted *in silico* since it has a distinct sequence feature. To see whether there are any sequence features in promoters that are missed by the prediction programs, I examined the nucleotide frequencies of promoters not predicted by any program. Additionally, I subdivided promoters that are not associated with any prediction according to their transcription class since 70% (72/103) of the RTs are associated with at least one prediction and/or a CpG island while only 23% (17/74) of the ATs are associated with at least one prediction and/or a CpG island

Figure 3.9 displays the summed nucleotide frequencies of RTs and ATs not associated with any prediction or a CpG island.

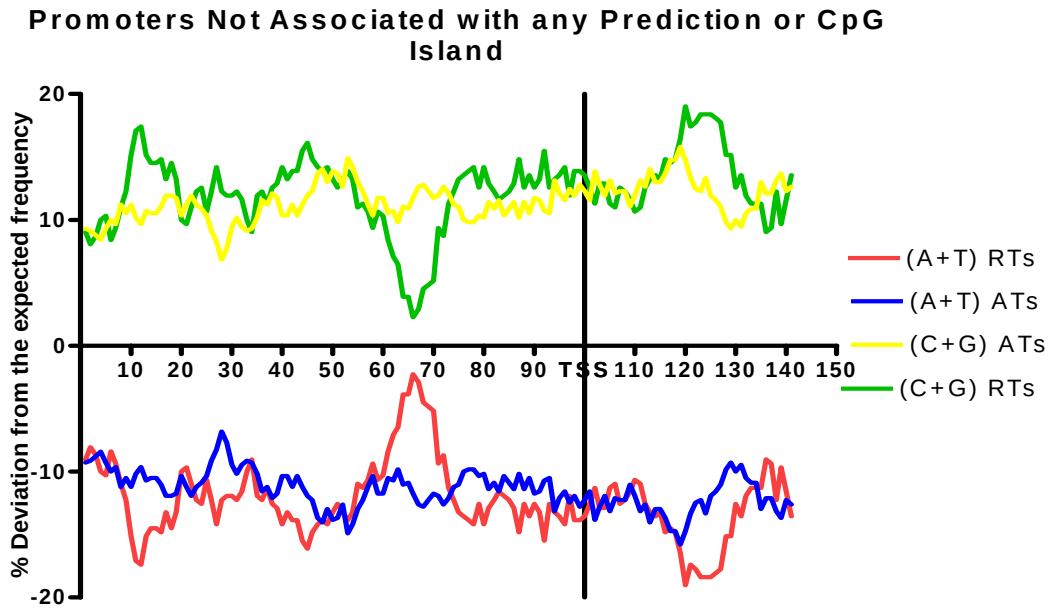

Figure 3.9. Summed nucleotide frequencies of RTs not associated with any prediction or a CpG island and ATs not associated with any prediction or a CpG island. Frequencies are averaged out in windows of ten nucleotides to eliminate local fluctuations.

The two transcript types have rather similar C+G and A+T content but points of clear difference are also visible. The A+T peak in the RTs plot around 35 bp upstream of the TSS indicates the presence of a TATA-box. There is no significant change in the A+T and C+G distribution between the two classes at 25 bp upstream and downstream of TSS. However, there is an increase in C+G content in the RTs 30 bp downstream of the TSS, which might be a sign for the downstream promoter element (DPE). For further analysis of the differences between the nucleotide frequency distributions between RTs and ATs with no prediction or CpG island, I fitted curves to each C+G nucleotide frequency distribution plot and subtracted fitted C+G curves of different transcripts respectively and plotted difference curve in Figure 3.10.

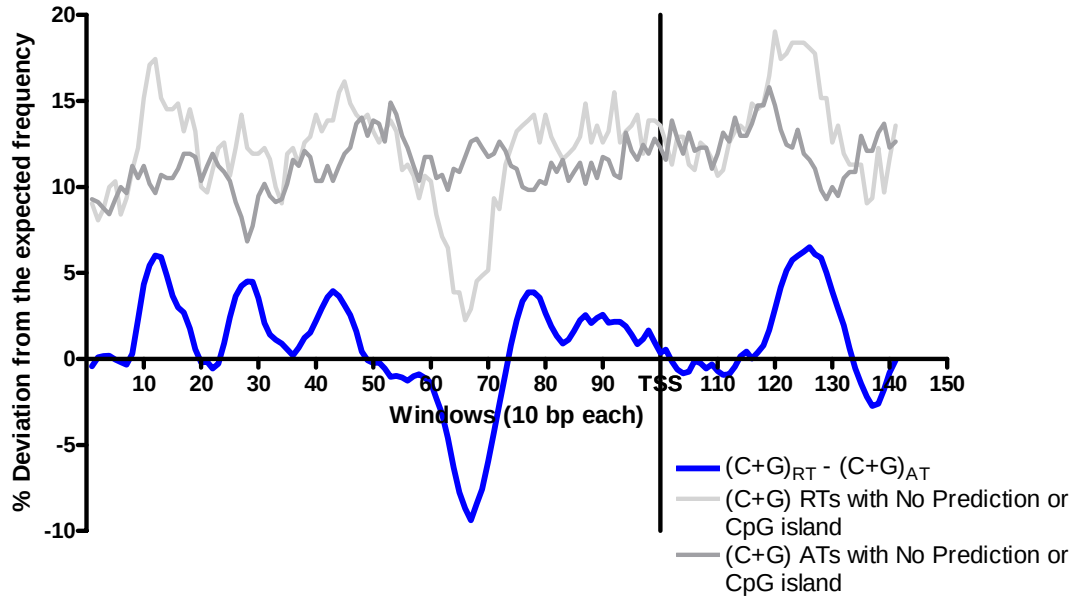

Figure 3.10 (C+G) plot of RTs and ATs not associated with any prediction are curve-fitted (grey curves) and subtracted from each other (blue curve).

The (C+G) peaks between 100 to 50 bp upstream of the TSS in Figure 3.11 may corresponds GC-boxes (consensus sequence, GGGCGGG) which serve as binding sites for the transcription factor Sp1 (Fukue et al., 2005).

In summary, promoters display a unique sequence pattern where (C+G) content is on average 20% higher from its expected frequency. Interestingly, (C+G) and (A+T) contents do not fluctuate greatly, which means that changes in the frequency of nucleotide C is mostly compensated by nucleotide G and changes in the frequency of nucleotide A is compensated by nucleotide T.

Computational efforts for predicting promoters or TSS seems quite biased towards (C+G) content of sequence as promoters with low (C+G) content cannot be detected by these programs (Figure 3.8). These programs certainly have room for improvement to detect more promoters on the sequence level, as Figure 3.10 clearly shows that representative transcripts which are not associated with any prediction or CpG island have distinct sequence features such as Sp1 and TBP binding sites as well as DPE. It is intriguing that although alternative transcripts not associated with any prediction or

a CpG island, do not have such distinct sequence motifs, they do retain the high GC content that may indicate promoter sequences.One can speculate that these promoters may be very tissue-specific and/or tightly-controlled, as they do not have clear binding sites for common activation factors.

3.3.2 Structural Topology of Promoter Sequences

Initiation of transcription is a complex process requiring a number of proteins acting co-operatively on the promoter region. DNA bending is a vital factor in protein binding and nucleosome positioning; a straight and rigid double helix cannot accommodate deformed structures which allows functional protein binding (Travers, 1989). It was therefore interesting to examine the bendability profile of the 177 putative promoter sequences in this study together with non-promoter sequences and finally correlate the bendability profiles of different types of promoters to structural features.

3.3.2.1 DNA Bending

DNA interactions with DNA I nuclease were used to understand the DNA bending as the binding of the nuclease is largely determined by the flexibility of the sequence towards the major groove (Brukner et al., 1990). It is possible to extract a measure of bending ability of a DNA sequence using experimental data produced by employing variety of DNA sequences and their corresponding DNase I cutting frequencies (Brukner et al., 1995). To this end, bendability figures for all possible trinucleotide sequences were produced as listed in Table 3.2; the measure is in arbitrary scale and higher values (less negative) means higher bendability towards the major groove (Brukner et al., 1995).

Sequence	Reverse Complemented	Bendability Score
AAA	TTT	-0.274
AAC	GTT	-0.205
AAG	CTT	-0.081
AAT	ATT	-0.28
ACA	TGT	-0.006
ACC	GGT	-0.32
ACG	CGT	-0.033
ACT	AGT	-0.183
AGA	TCT	0.027
AGC	GCT	0.017
AGG	CCT	-0.057
ATA	TAT	0.182
ATC	GAT	-0.11
ATG	CAT	0.134
CAA	TTG	0.015
CAC	GTG	0.04
CAG	CTG	0.175
CCA	TGG	-0.246
CCC	GGG	-0.012
CCG	CGG	-0.136
CGA	TCG	-0.003
CGC	GCG	-0.077
CTA	TAG	0.09
CTC	GAG	0.031
GAA	TTC	-0.037
GAC	GTC	-0.013
GCA	TGC	0.076
GCC	GGC	0.107
GGA	TCC	0.013
GTA	TAC	0.025
TAA	TTA	0.068
TCA	TGA	0.194

Table 3.2. Bendability scores of all possible trinucleotides. Higher values (less negative) translate to higher bendability towards major groove.

Bendability scores were produced for the 177 putative promoters using 100 bp upstream and 50 bp downstream sequence of the TSS. A bendability score was assigned to every trinucleotide along the sequence according to Table 3.2 and the scores of every 12 nucleotides (10 trinucleotides) were averaged out to smooth the local fluctuations as it is illustrated in Figure 3.11.

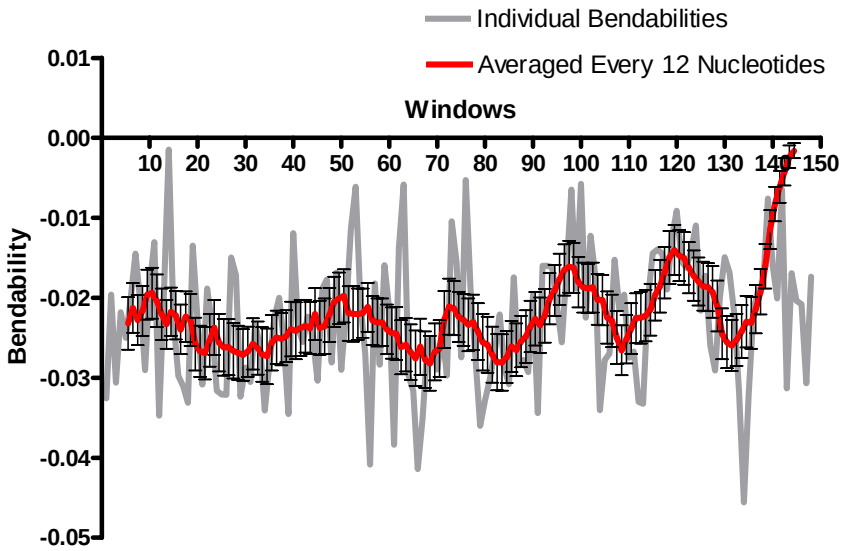

Figure 3.11. Average bendability score of every trinucleotide along the sequence of 177 promoter sequences (grey line) and bendability scores averaged at every 12 nucleotide (10 trinucleotides) shown with error bars (red line).

Figure 3.12 shows the bendability profile of 201 randomly selected intergenic and intra-genic human sequences (negative controls) alongside those of the 177 putative promoters.

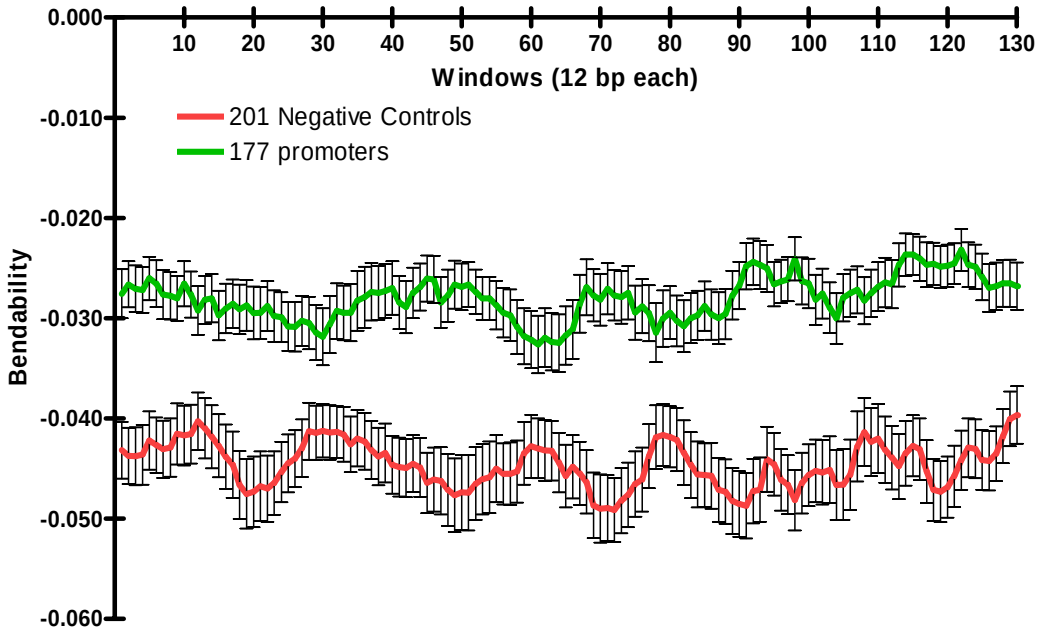

Figure 3.12. Bendability scores of 201 negative control sequences (red line with error bars) versus 177 promoter sequences (green line with error bars) averaged out in every 12 nucleotides.

Promoter sequences have a higher mean bendability (0.02704 ± 0.004546) compared to the negative control set (0.04302 ± 0.00687). The high bendability of promoter

sequences might offer a more flexible structure to DNA helix to accommodate protein binding events.

Open-chromatin and gene rich regions of the genome are enriched with GC-rich sequences, which can have higher bendabilities (Vinogradov, 2003). The bendability profiles of putative promoters associated or not associated with a CpG island are shown in Figure 3.13.

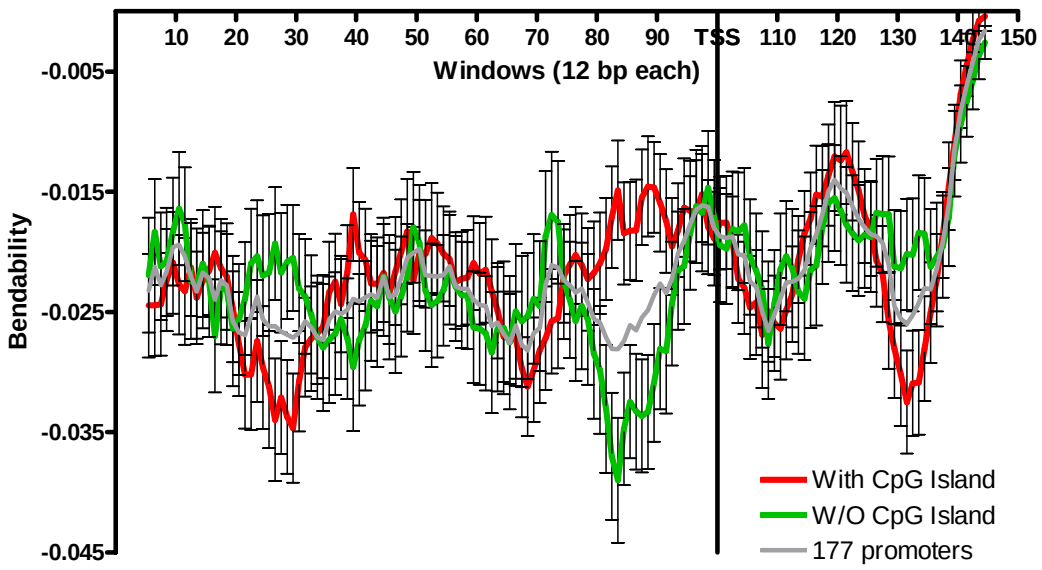

Figure 3.13. Bendability scores of 81 promoters associated with a CpG island (red line with error bars) and 86 promoters not associated with a CpG island (green line with error bars). Grey line indicates the bending profile of all the dataset.

The bendability profiles of the two sets overall do not differ greatly. However, there is a difference in bendability scores around -20 bp, which coincides with the position of TATA-box motif. However, analysis of a genome-wide set will be required to assess the significance of this observation.

Such different DNA conformations might help recruiting different sets of protein complexes or even different type of initiation complexes (Wieczorek et al., 1998).

Since there is a significant difference in nucleotide distribution of representative and alternative transcripts not associated with any prediction or CpG island (see Figure

3.10), I assessed the bending profiles of these two types of transcripts. The plots of bendability scores of the two transcript groups are presented in Figure 3.14.

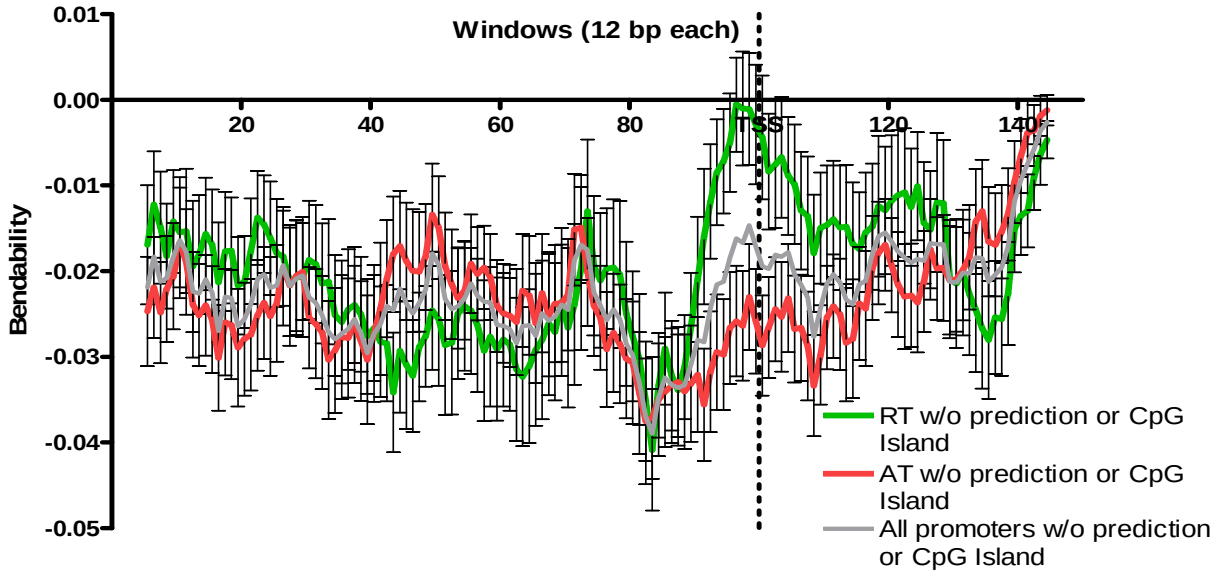

Figure 3.14. Bendability scores of RTs (green line with error bars) and ATs (red line with error bars) not associated with any promoter or TSS prediction or a CpG island and grey line denotes the bendability of all promoters sequences not associated with any prediction or a CpG island.

The two groups show significant differences in their bending profiles most notably around the TSS where RTs not associated with a CpG island or prediction have much higher bendability towards the major groove. This high bending might be the result of (A+T)-rich sequence just before the TSS found only in representative transcripts (see Figure 3.9).

From these analyses, we can conclude that there is a specific bendability profile in promoter sequences and this profile is affected by the presence of a CpG island. As expected, there is a correlation between the sequence content and the bending capacity of the sequence as higher bendability is observed in the presence of (A+T)-rich sequence profile in case of the promoters of representative transcripts not associated with a CpG island or prediction.

3.4 Expression Profile of 20q12-13.2 using Affymetrix Arrays

Affymetrix Human Expression Array U133 plus 2.0 was used to determine the expression status of genes in the 20q12-13.2 region on HeLa S3 and NTERA-D1 cell lines. These two cell lines were used to characterise experimentally the 177 putative promoter sequences by gene reporter assays (see Chapter 4) and chromatin immunoprecipitation assays (see Chapter 5). The arrays were hybridized with the total RNA extract of each cell line as described in section 2.3. A script was written to extract all the probes on the corresponding array perfectly aligning to the sequence of the investigated region. There were 280 probes aligning to 20q12-13.2 and 205 of them aligned to a coding transcript. There were 6 and 13 probes representing non-coding transcripts and pseudogenes respectively. The remaining 56 probes were either

- representing partial cDNAs
- ambiguous
- matching on the reverse strand of an annotation
- matching a partial gene, or exon prediction
- not matching with any annotated feature or experimental evidence

Affymetrix probes are classified into 4 groups according to the uniqueness of the probe (Ivanova et al., 2006). Probes that have a;

- "**_at**" suffix represent probes designed to detect a unique sequence of a single gene
- "**_a_at**" suffix represent probes designed to detect multiple alternative transcripts of a single gene
- "**s_at**" suffix represent probes designed to detect multiple transcripts of different genes
- "**x_at**" suffix represent probes that can cross-hybridize to unrelated sequences; these probes should be treated with caution.

I removed all the probes with an "x_at" suffix since some of the signal of this type of probes might be produced by an unrelated sequence. This left 188 unique probes.

3.4.1 Coding Genes and Transcripts

Out of 103 genes, 99 (96%) genes are represented by at least one probe on Affymetrix U133 Plus 2 expression array respectively. At the gene level, 47 (47.5%) genes were expressed in both cell lines and 29 (30%) genes are not expressed in any of the two cell lines used in this study. As shown in Figure 3.16, 17 genes are only expressed in NTERA-D1 and 6 genes are only expressed in HeLa S3 cell line. Therefore, NTERA-D1, a testis cell line, has a broader expression pattern as expected, since testis has already shown as one of the tissue source of one of the most complex and diversified transcriptome (Jongeneel et al., 2005). The expression status of all the genes is listed in Table A5 (Appendix A).

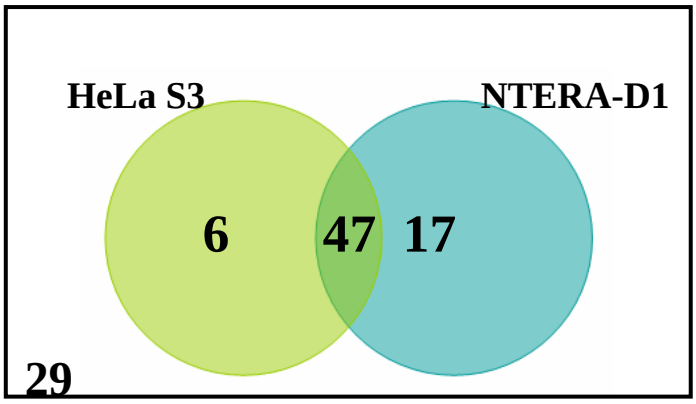

Figure 3.15. Expression profile of 96 genes represented by probes on Affymetrix U133 Plus 2 expression array, where 29 genes are not expressed by neither of the cell lines.

Table A6 (Appendix A) lists the transcripts that are represented by each probe on the expression array, unfortunately none of the probes was able to differentiate the expression of the alternative transcripts of the same gene.

3.4.2 Expression Profiles and CpG Islands

Out of the 99 genes represented by at least one probe on Expression Array, 57 contain a CpG island (see Table 3.3). Of those 57, 36 (63.2%) were expressed in both cell lines. Out of the 42 genes not associated with a CpG island, 11 (23.8%) were expressed in both cell lines while 21 (50%) were not expressed in either. In summary, 86% of the genes associated with a CpG island are expressed in at least one cell line whereas this figure is 50% for the genes not associated with a CpG island.

	Representative Transcripts Associated with a CpG Island	**Representative Transcripts NOT Associated with a CpG Island**
Expressed in both cell lines	36 (**63.2**%)	10 (**23.8**%)
Expressed only in HeLa S3	1 (1.8%)	5 (11.9%)
Expressed only in NTERA-D1	12 (21.1%)	6 (14.3%)
No expression	8 (**14.0**%)	21 (**50.0**%)
Total number of transcripts	56	42

Table 3.3. Expression Profiles of 96 representative transcripts associated or not associated with CpG islands

3.4.3 Pseudogenes and Processed Transcripts

There are seven probes aligning to processed pseudogenes and four of them gave high positive signals, but all these pseudogenes show sequence homologies with the ubiquitously expressed ribosomal protein genes. There are six probes designed to detect to expression of processed transcripts. One of these probes ("226835_s_at" aligning to RP4-686N3.3 processed transcript) showed a very high positive signal. However, this probe cross hybridizes the mRNA of E3 ubiquitin-protein ligase gene (NEDD4). So it is not possible to differentiate the signal since this gene is widely expressed.