An *In Silico* Investigation into the Discovery of Novel *Cis*-acting Elements within the Intronic Regions of Human *PAX7*

*Maika G. Mitchell [1,2], Melanie Ziman [1]

[1] School of Exercise, Biomedical and Health Science, Edith Cowan University, Perth, Western Australia 6027,
[2]Email: blackmam@mskcc.org

[2] Sloan Kettering Institute (Memorial Sloan Kettering Cancer Center), New York City, New York 10021, USA

Abstract: PAX3 and PAX7 are homologous paired box family members expressed during early neural and myogenic development. Assays of mRNA expression have proven conclusively that PAX3 and PAX7 transcripts are present in embryonal and alveolar rhabdomyosarcoma, neuroblastoma, Ewing's sarcoma, and melanoma cell lines; the tumor-specific expression patterns correspond to expression patterns in corresponding embryonic cell lineages. The intronic regions of the PAX7 gene were analyzed using computational DNA pattern recognition methods. Several potential cis-regulatory motifs were identified in this investigation and one in particular that was common to both PAX7 and PAX3 and also to NF1, could have implications for the role of PAX7 in Alveolar Rhabdomyosarcoma and may be the cornerstone to more exciting, unique scientific investigations. **Methods:** In Silico biology methods are currently used in the pharmaceutical industry as an antecedent to wet chemistry and bench work. Here we employed several public online and offline programs/databases as tools to investigate the nucleotide sequences of the PAX7 gene. **Results:** Several potential cis-acting elements within the intronic regions of PAX7 were discovered through in silico biological methods. Transcription factors that could bind to these elements have also been identified and their association with cancer ascertained. Interestingly one cis element is found within a 155 bp sequence in intron 8 of PAX7 that surprisingly, is also found within intron 10 of PAX3 and is also found conserved within intron 23 of the NF-1 gene. Discussion: The use of In Silico Biology methods represent new, faster, cost-efficient techniques to identify novel regulatory elements that provide areas for more in depth in vitro investigations to confirm their functional effects. [Nature and Science. 2006;4(3):69-85].

Keywords: *PAX7*; *PAX3*; *cis* regulatory elements; *NF-1*; conserved sequences; ERMS; ARMS

Abbreviations and notations: TSS, transcription start site; *Cis*-acting element; ARMS, alveolar rhabdomyosarcoma; ERMS, embryonal rhabdomyosarcoma; NF-1, Neurofibromatosis factor 1; bp, base pair; TF, Transcription Factor; RD, Rhabdomyosarcoma; TSS, transcription start site

INTRODUCTION

Pax genes derive their name from the paired box gene region which encodes a highly conserved Paired DNA binding domain. Paired domains are found in all members of the Pax family. There are four classes of *PAX* genes based not only on sequence but on genomic organization. Genes within a given class have intron/exon boundaries and encoding regions in common. *Pax3* and *Pax7* are closely related paired box family members expressed during early neural and myogenic development and have been implicated in the development of specific myogenic and neurogenic cell lineages (Glaser, et al., 1994; Relaix et al., 2004). Pax proteins are thought to function primarily by binding to enhancer DNA sequences and modifying the transcriptional activity of bound downstream target genes (Chi et al., 2002). Assays detailing human *PAX7* and *PAX3* mRNA expression show conclusively that transcripts of these genes are present in Alveolar Rhabdomyosarcoma, Embryonal Rhabdomyosarcoma, neuroblastoma, Ewing's sarcoma, and melanoma cell lines and reveal tumor-specific expression patterns that correspond to those in corresponding embryonic cell lineages (Goulding, et al.,1991; Bennicelli et al., 1993; Macina et al., 1995; Schulte et al., 1997; Barr et al., 1999; Mercado et al., 2005).

The search for new regulatory elements in unreported regions of the *PAX7* gene would lead to a better understanding of the oncogenic pathways that activate this gene. In this study, the eight introns of *Homo sapiens PAX7* were scanned for new regulatory elements which may affect tumorigenesis in humans. We monitored the DNA repeat patterns of the eight introns in the time period from August 2005 to March 2006 to determine if updates to the NCBI database would affect the predicted outcomes for each query. The results for the eight introns of *PAX7* remained consistent in several repetitions of the experiment. We analyzed the intronic regions of this gene using computational DNA pattern recognition methods. We report here that several potential *cis*-regulatory motifs were identified in this way. The possible significance of all identified *cis* motifs for the *PAX7* gene were investigated using various web and offline databases that employ similar statistical tests and parameters. Moreover, transcription factors likely to bind to the elements were identified and the association of these transcription factors with tumour cell function determined.

Specifically, one newly identified *cis* element was found in a conserved region of intron 8 of *PAX7*; the same sequence containing the *cis* element (ctccaccc) was also found in alternative intron 10 of *PAX3* and in the 3' region (in intron 23) of the tumor suppressor gene *Neurofibromatosis factor 1* (*NF-1*). The presence of this element has not previously been reported in association with the intronic regions of *PAX7* up to the date of submission of this publication (June 2006). The region of the *NF-1* gene that is present in intron 8 of *PAX7* and identified here by *in silico* data mining methods, has recently been linked to Embryonal Rhabdomyosarcoma (ERMS) (Hadjistilianou et al., 2002; Oguzkan et al., 2006) and Alveolar Rhabdomyosarcoma (ARMS) (Woodruff et al., 1993; Dei Tos et al., 1997) and confirmed as a significant role player in carcinogenesis in a recent publication using fluorescent-labeled microsatellite markers. (Oguzkan et al., 2006).

The conserved sequence containing the *cis* regulatory element identified in intron 8 of *PAX7*, may have arisen by insertion of a regulatory element in all three chromosomal regions or by homologous recombination between chromosome 1 (*PAX7*), chromosome 2 (*PAX3*) and/or chromosome 17 (*NF-1*). As a result of the similarly conserved intronic sequences present in all three genes, it is possible that the genes are similarly regulated by common transcription factors (TFs) proposed to bind to the common *cis* elements. Experimentation and *in vitro* studies that may prove the biological importance of these gene sequences, is currently underway.

MATERIALS AND METHODS

DNA Data-mining

Definition of *Cis* Elements

Before collecting meaningful data for the basis of this research, the first course of action for data mining is to define a *cis* element and second, to test several software applications that can be used to find *cis* elements. A *cis*-acting element controls the initiation, or the rate of transcription and translation of genes that reside on the same chromosome as itself. *Cis* elements contain the following features (Park et al., 2003):

A short consensus sequence (≥ 8 base pairs long);

No fixed location but usually 100-200 bp upstream of the transcription start site or within 10 kb upstream or downstream or within intronic regions of a gene;

Can be located in a promoter or act as an enhancer or silencer;

It is assumed that a specific protein binds to the element and the presence of that protein is spatially and temporally regulated;

One consensus sequence is usually sufficient to confer a regulatory response but the sequence may be present as one of several consensus sequences close together or it may be present as tandem repeat units.

Knowledge of new *cis* elements in the intronic regions of *PAX7* (theoretical *cis* elements to be validated later by *in vitro* studies), may lead to a better understanding of the factors that lead to over-expression of the gene with resultant increased tumorigenicity (Lewin et al., 2000).

The Selection of Databases used for the Computational Portion of the Research

Computational queries were performed against known, validated segments of sequences in order to quantitate a threshold of accuracy for all future evaluations of data. Automatic e-mail updates for the *PAX7* sequences for *Homo sapiens* (as well as for the *Homo sapiens PAX3* and *NF-1* gene) was set up within the National Center for Biotechnology Information database (NCBI, http://www.ncbi.nlm.nih.gov).

To determine functional significance, that is, biological properties of the newly identified *cis* elements, their position within sets of conserved sequences was determined by computational sequence alignment, a fundamental means of detecting biologically significant patterns in genes (Lewin et al., 2000). Multiple alignment methods were used to locate and align exons or introns of DNA in an attempt to locate and align similar subsequences. This approach has often been used to look for transcription factor binding sites in similarly regulated promoters (Liu et al., 1995; Frith et al., 2004; Chen et al., 2005).

Programs Used for *In Silico* Investigations

The introns of *PAX7* were scrutinized for *cis* elements with eight separate programs which denote promoter areas, transcription factor binding sites and/or transcription start sites (TSS), DNA patterns, global and local nucleotide alignment and tandem repeats in submitted sequences. The names and functions of the programs used are:

1) Mreps: (http://bioweb.pasteur.fr/seqanal/interfaces/mreps.html) - Mreps is a software package for identifying tandem repeats (patterns that appear >1x in a given sequence) in DNA sequences.

2) CLC Free Workbench version 2.2 by CLC bioA/S: (http://www.clcbio.com) - The alignment software illustrates the conservation of all sequence positions below aligned sequences. The height of the bars in the view reflects how conserved that particular position is in the aligned sequence. If one position is 100% conserved the bar will be at full height. The

software uses a progressive alignment algorithm (Oguzkan et al., 2006) in order to create multiple alignments.

3) DNA Pattern Search – Softberry: (http://www.softberry.com/) - This program searches for significant patterns in the set of sequences.

4) PROSCAN Version 1.7 Web Promoter Scan Service: (http://bimas.dcrt.nih.gov/molbio/proscan/) - Predicts promoter regions based on homologies with putative eukaryotic Pol II promoter sequences. The site is serviced and maintained by Dr. Dan Prestridge at the Advanced Biosciences Computing Center, University of Minnesota.

5) Promoter 2.0 Prediction Server: (http://www.cbs.dtu.dk/services/Promoter/) – Promoter 2.0 predicts transcription start sites of vertebrate PolII promoters in DNA sequences. It has been developed as a frequently updated database of simulated transcription factors that interact with sequences in promoter regions. It builds on principles that are common to neural networks and genetic algorithms. The site is serviced and maintained by Steen Knudsen at The Center for Biological Sequence Analysis at the Technical University of Denmark.

6) TSSG - Recognition of human PolII promoter regions and transcription start sites from Softberry: (http://www.softberry.com/) - TSSG is the most accurate mammalian *cis* element prediction program.

7) CLC Gene Workbench v. 1.0.1 by CLC bioA/S: (http://www.clcbio.com) - Applying the Pattern Discovery helps identify unknown sequence patterns across single or multiple DNA and protein sequences. The discovery method is based on advanced hidden Markov models.

8) TRANSFAC® 7.0: (http://www.gene-regulation.com/pub/databases.html#transfac) - is a database of eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles. This database was used to compare DNA patterns discovered during the data-mining stage of this research with known *cis*-acting elements and identify the transcription factors most likely to bind to them.

For each program, fasta files of the eight introns of *PAX7* were pasted into each program. The output was saved into a word document and/or a portable data file (PDF) for scrutiny and review.

RESULTS

1. Prediction of novel *cis* regulatory regions by computer scans of intronic regions of *PAX7*:

The list of novel *cis*-acting elements found in each intron of *Homo sapiens PAX7* was created by analysis of:

1) The location of the *cis*-acting element in the sequence compared to the locality of known/previously identified *cis*-acting elements;

2) Comparison of the results from the different DNA patterning software programs;

3) Location of pattern to the proximity of start and stop codons or to exon/intron boundaries;

4) Comparisons of DNA patterns to those that exist for other *cis*-acting elements on the TRANSFAC Database v.5.0 (http://transfac.gbf.de/TRANSFAC/index.html).

Below are the results for each intron of *PAX7* obtained as a result of scanning the intronic sequences of *PAX7* using the software, databases and search criteria specified above. Table 9 contains a summary of the list of proposed *cis* elements for each intron of *PAX7* and table 10 lists the transcription factors that bind to each proposed *cis* elements.

PAX7 INTRON ANALYSIS: INTRON 1

PAX7 intron 1 patterns found by Softberry Pattern Search Software: Found 5 pattern (s)

1) *Pattern 1*, Length = 10, 552bp - 561bp
AAATAATAAT

2) *Pattern 2*, Length = 9, 552bp - 560bp
AAATAATAA

3) *Pattern 3*, Length = 10, 553bp - 562bp
AATAATAATT

4) *Pattern 4*, Length = 10, 554bp - 563bp
ATAATAATTA

5) *Pattern 5*, Length = 10, 1322bp – 1331bp
AATATAAAGT

Proscan: Version 1.7
Cis element region predicted on forward strand at 1094bp to 1344bp
TATA found at 1323bp, Est.TSS at 1353bp

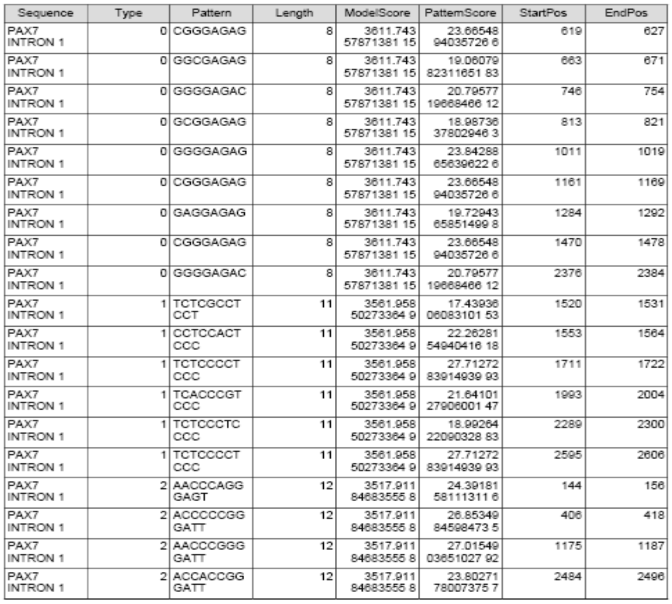
Softberry TSSG**:**
2 promoter(s) are predicted .
Promoter position: 1350 LDF: TATA box at 1324bp
TATAAAGT
Promoter position: 492 LDF: TATA box at 463bp
TTTATATG

Promoter 2.0 Prediction Server:

Position (bp)	Score	Likelihood
600	0.638	Marginal prediction
1000	0.561	Marginal prediction
2000	0.629	Marginal prediction

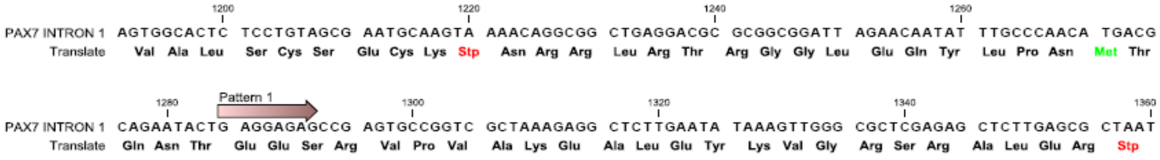
CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 1**)**

Table 1: CLC Gene Workbench v.1.0.1. Pattern Discovery Search



Sequence	Type	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 1	0	CGGGAGAG	8	3611.743 57871381 15	23.66548 94035726 6	619	627
PAX7 INTRON 1	0	GGCGAGAG	8	3611.743 57871381 15	19.06079 82311651 83	663	671
PAX7 INTRON 1	0	GGGGAGAC	8	3611.743 57871381 15	20.79577 19668466 12	746	754
PAX7 INTRON 1	0	GCGGAGAG	8	3611.743 57871381 15	18.98736 37802946 3	813	821
PAX7 INTRON 1	0	GGGGAGAG	8	3611.743 57871381 15	23.84288 65639622 6	1011	1019
PAX7 INTRON 1	0	CGGGAGAG	8	3611.743 57871381 15	23.66548 94035726 6	1161	1169
PAX7 INTRON 1	0	GAGGAGAG	8	3611.743 57871381 15	19.72943 65851499 8	1284	1292
PAX7 INTRON 1	0	CGGGAGAG	8	3611.743 57871381 15	23.66548 94035726 6	1470	1478
PAX7 INTRON 1	0	GGGGAGAC	8	3611.743 57871381 15	20.79577 19668466 12	2376	2384
PAX7 INTRON 1	1	TCTCGCCT CCT	11	3561.958 50273364 9	17.43936 06083101 53	1520	1531
PAX7 INTRON 1	1	CCTCCACT CCC	11	3561.958 50273364 9	22.26281 54940416 18	1553	1564
PAX7 INTRON 1	1	TCTCCCCT CCC	11	3561.958 50273364 9	27.71272 83914939 93	1711	1722
PAX7 INTRON 1	1	TCACCCGT CCC	11	3561.958 50273364 9	21.64101 27906001 47	1993	2004
PAX7 INTRON 1	1	TCTCCCTC CCC	11	3561.958 50273364 9	18.99264 22090328 83	2289	2300
PAX7 INTRON 1	1	TCTCCCCT CCC	11	3561.958 50273364 9	27.71272 83914939 93	2595	2606
PAX7 INTRON 1	2	AACCCAGG GAGT	12	3517.911 84683555 8	24.39181 58111311 6	144	156
PAX7 INTRON 1	2	ACCCCCGG GATT	12	3517.911 84683555 8	26.85349 84598473 5	406	418
PAX7 INTRON 1	2	AACCCGGG GATT	12	3517.911 84683555 8	27.01549 03651027 92	1175	1187
PAX7 INTRON 1	2	ACCACCGG GATT	12	3517.911 84683555 8	23.80271 78007375 7	2484	2496

Figure 1



```
                  1200                  1220                   1240                   1260
PAX7 INTRON 1  AGTGGCACTC  TCCTGTAGCG  AATGCAAGTA  AAACAGGCGG  CTGAGGACGC  GCGGCGGATT  AGAACAATAT  TTGCCCAACA  TGACG
    Translate   Val Ala Leu  Ser Cys Ser  Glu Cys Lys Stp  Asn Arg Arg  Leu Arg Thr  Arg Gly Gly Leu  Glu Gln Tyr  Leu Pro Asn  Met Thr

             1280                  1300                   1320                   1340                   1360
PAX7 INTRON 1  CAGAATACTG  AGGAGAGCCG  AGTGCCGGTC  GCTAAAGAGG  CTCTTGAATA  TAAAGTTGGG  CGCTCGAGAG  CTCTTGAGCG  CTAAT
    Translate   Gln Asn Thr  Glu Glu Ser Arg  Val Pro Val  Ala Lys Glu  Ala Leu Glu Tyr  Lys Val Gly  Arg Ser Arg  Ala Leu Glu Arg  Stp
```

Figure 1: *Cis* regulatory region predicted by CLC Gene Workbench v. 1.0 – on forward strand at position 1094bp to 1344bp; ***Pattern found within this cis regulatory region.***

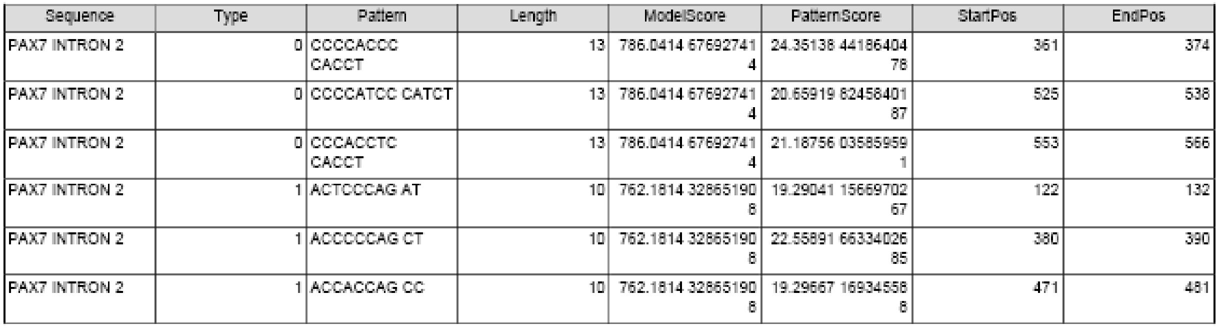
PAX7 INTRON ANALYSIS: INTRON 2

<u>*PAX7* intron 2 patterns found by Softberry Pattern Search Software: Found 5 pattern(s)</u>

1) *Pattern 1*, Length = 10, Power: 1, 422bp - 431bp AAAGGATAAA
2) *Pattern 2*, Length = 10, Power: 1, 423bp - 432bp AAGGATAAAG
3) *Pattern 3*, Length = 10, Power: 1, 421bp - 430bp GAAAGGATAA
4) *Pattern 4*, Length = 9, Power: 1, 95bp - 103bp AGGAAAGTA
5) *Pattern 5*, Length = 9, Power: 1, 421bp - 429bp GAAAGGATA

<u>Proscan: Version 1.7:</u>
Processed Sequence: 572 bp. No *cis* element regions predicted.
<u>Softberry programs:</u>
0 promoter/enhancer(s) predicted
<u>Promoter 2.0 Prediction Server:</u>
No *cis* element predicted
<u>CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 2)</u>

Table 2: CLC Gene Workbench v.1.0.1. Pattern Discovery Search



Sequence	Type	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 2	0	CCCCACCC CACCT	13	786.0414 676927414	24.35138 4418640478	361	374
PAX7 INTRON 2	0	CCCCATCC CATCT	13	786.0414 676927414	20.65919 6245840187	525	538
PAX7 INTRON 2	0	CCCACCTC CACCT	13	786.0414 676927414	21.18756 035859591	553	566
PAX7 INTRON 2	1	ACTCCCAG AT	10	762.1814 328651908	19.29041 1566970267	122	132
PAX7 INTRON 2	1	ACCCCCAG CT	10	762.1814 328651908	22.55891 6633402685	380	390
PAX7 INTRON 2	1	ACCACCAG CC	10	762.1814 328651908	19.29567 169345588	471	481

Figure 2

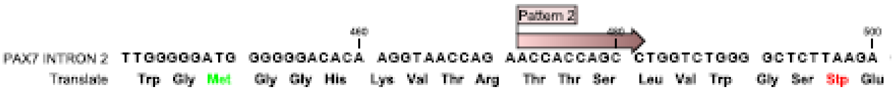

Pattern 2
460
480
500
PAX7 INTRON 2
TTGGGGGATG GGGGGACACA AGGTAACCAG AACCACCAGC CTGGTCTGGG GCTCTTAAGA
Translate
Trp Gly Met Gly Gly His Lys Val Thr Arg Thr Thr Ser Leu Val Trp Gly Ser Stp Glu

Figure 2: No *cis* element predicted, but 1 pattern found within 400-500 bps by CLC Gene Workbench v. 1.0.

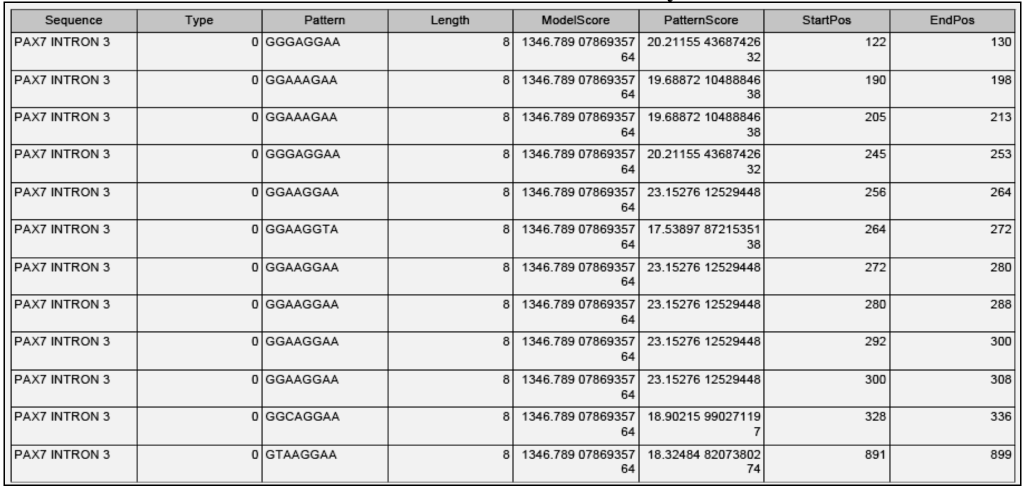
PAX7 INTRON ANALYSIS: INTRON 3

<u>*PAX7* intron 3 patterns found by Softberry Pattern Search Software</u>: *No patterns found*
<u>Proscan: Version 1.7</u> No *cis* element regions predicted.
<u>Softberry TSSG</u>: No *cis* element regions predicted.
<u>Promoter 2.0 Prediction Server</u>:

Position	Score	Likelihood
200	0.557	Marginal prediction

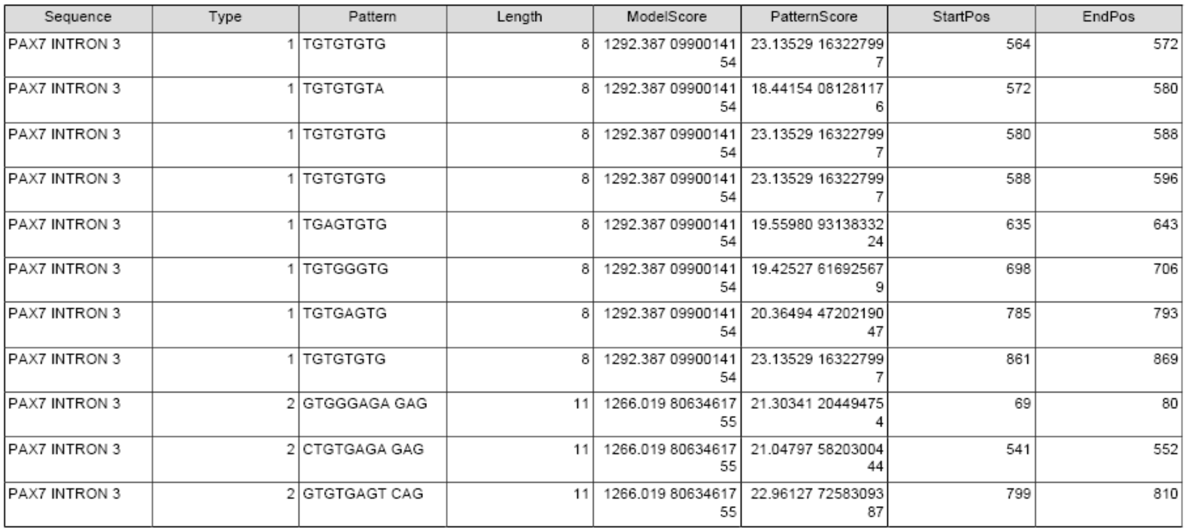
<u>Proscan: Version 1.7:</u>
Processed Sequence: 996 Base Pairs. No *cis* element regions predicted.
<u>Softberry programs:</u>
0 promoter/enhancer(s) predicted.

Table 3a: CLC Gene Workbench v.1.0.1. Pattern Discovery Search



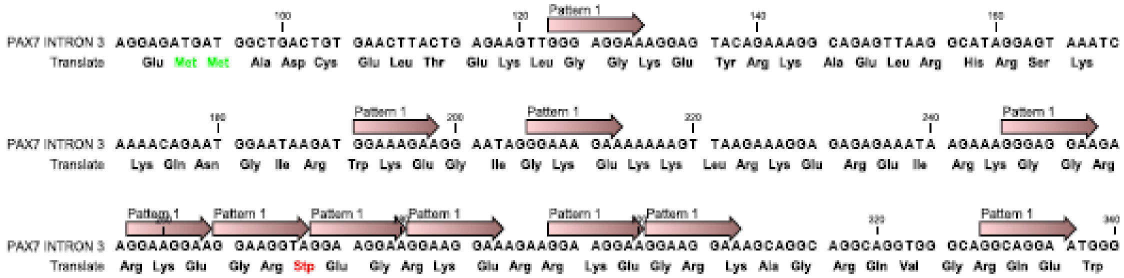
Sequence	Type	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 3	0	GGGAGGAA	8	1346.789 0786935764	20.21155 4368742632	122	130
PAX7 INTRON 3	0	GGAAAGAA	8	1346.789 0786935764	19.68872 1048884638	190	198
PAX7 INTRON 3	0	GGAAAGAA	8	1346.789 0786935764	19.68872 1048884638	205	213
PAX7 INTRON 3	0	GGGAGGAA	8	1346.789 0786935764	20.21155 4368742632	245	253
PAX7 INTRON 3	0	GGAAGGAA	8	1346.789 0786935764	23.15276 12529448	256	264
PAX7 INTRON 3	0	GGAAGGTA	8	1346.789 0786935764	17.53897 8721535138	264	272
PAX7 INTRON 3	0	GGAAGGAA	8	1346.789 0786935764	23.15276 12529448	272	280
PAX7 INTRON 3	0	GGAAGGAA	8	1346.789 0786935764	23.15276 12529448	280	288
PAX7 INTRON 3	0	GGAAGGAA	8	1346.789 0786935764	23.15276 12529448	292	300
PAX7 INTRON 3	0	GGAAGGAA	8	1346.789 0786935764	23.15276 12529448	300	308
PAX7 INTRON 3	0	GGCAGGAA	8	1346.789 0786935764	18.90215 990271197	328	336
PAX7 INTRON 3	0	GTAAGGAA	8	1346.789 0786935764	18.32484 8207380274	891	899

Table 3b: 1 CLC Gene Workbench v.1.0.1. Pattern Discovery Search



Sequence	Type	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 3	1	TGTGTGTG	8	1292.387 0990014154	23.13529 163227997	564	572
PAX7 INTRON 3	1	TGTGTGTA	8	1292.387 0990014154	18.44154 081281176	572	580
PAX7 INTRON 3	1	TGTGTGTG	8	1292.387 0990014154	23.13529 163227997	580	588
PAX7 INTRON 3	1	TGTGTGTG	8	1292.387 0990014154	23.13529 163227997	588	596
PAX7 INTRON 3	1	TGAGTGTG	8	1292.387 0990014154	19.55980 9313833224	635	643
PAX7 INTRON 3	1	TGTGGGTG	8	1292.387 0990014154	19.42527 616925679	698	706
PAX7 INTRON 3	1	TGTGAGTG	8	1292.387 0990014154	20.36494 4720219047	785	793
PAX7 INTRON 3	1	TGTGTGTG	8	1292.387 0990014154	23.13529 163227997	861	869
PAX7 INTRON 3	2	GTGGGAGA GAG	11	1266.019 8063461755	21.30341 204494754	69	80
PAX7 INTRON 3	2	CTGTGAGA GAG	11	1266.019 8063461755	21.04797 5820300444	541	552
PAX7 INTRON 3	2	GTGTGAGT CAG	11	1266.019 8063461755	22.96127 7258309387	799	810

Figure 3



```
PAX7 INTRON 3  AGGAGATGAT  GGCTGACTGT  GAACTTACTG  AGAAGTTGGG  AGGAAAGGAG  TACAGAAAGG  CAGAGTTAAG  GCATAGGAGT  AAATC
    Translate     Glu Met Met  Ala Asp Cys  Glu Leu Thr  Glu Lys Leu Gly  Gly Lys Glu  Tyr Arg Lys  Ala Glu Leu Arg  His Arg Ser  Lys

PAX7 INTRON 3  AAAACAGAAT  GGAATAAGAT  GGAAAGAAGG  AATAGGGAAA  GAAAAAAAGT  TAAGAAAGGA  GAGAGAAATA  AGAAAGGGAG  GAAGA
    Translate     Lys Gln Asn  Gly Ile Arg  Trp Lys Glu Gly  Ile Gly Lys  Glu Lys Lys  Leu Arg Lys Glu  Arg Glu Ile  Arg Lys Gly  Gly Arg

PAX7 INTRON 3  AGGAAGGAAG  GAAGGTAGGA  AGGAAGGAAG  GAAAGAAGGA  AGGAAGGAAG  GAAAGCAGGC  AGGCAGGTGG  GCAGGCAGGA  ATGGG
    Translate    Arg Lys Glu  Gly Arg Stp Glu  Gly Arg Lys  Glu Arg Arg  Lys Glu Gly Arg  Lys Ala Gly  Arg Gln Val  Gly Arg Gln Glu  Trp
```

Figure 3: *Cis* regulatory region predicted by CLC Gene Workbench v. 1.0 on forward strand from 190-300bps; ***Several patterns found within this region.***

PAX7 INTRON ANALYSIS: INTRON 4

PAX7 intron 4 patterns found by Softberry Pattern Search Software: Found 5 pattern(s)

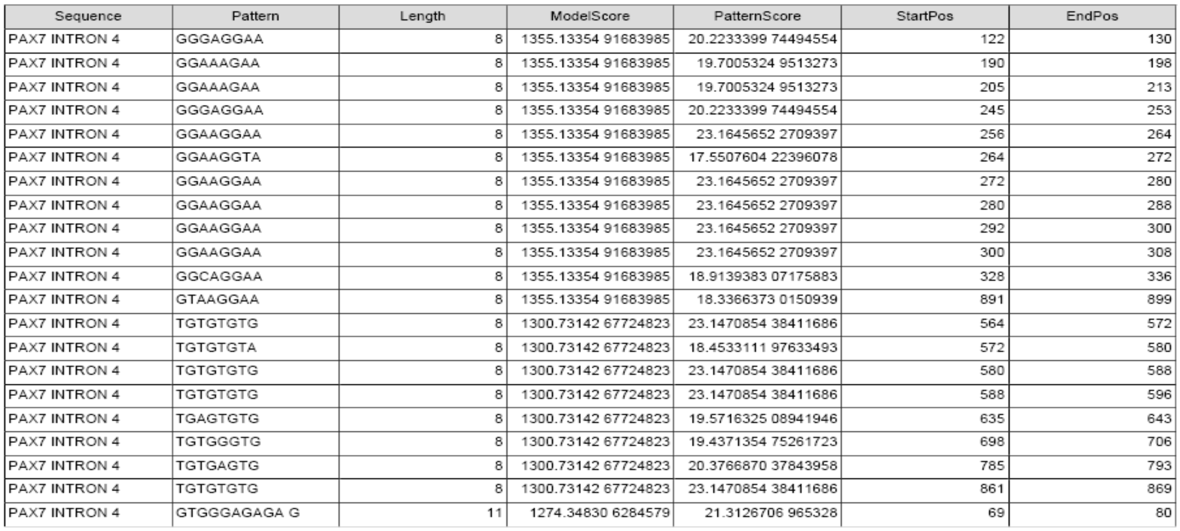
1) *Pattern* *1*, Length = 9, Power: 1, 916bp - 924bp CTTTCTCCC
2) *Pattern* *2*, Length = 10, Power: 1, 948bp - 957bp CCTCTGCTCC
3) *Pattern* *3*, Length = 10, Power: 1, 942bp - 951bp CTCGCTCCTC
4) *Pattern* *4*, Length = 10, Power: 1, 916bp - 925bp CTTTCTCCCA
5) *Pattern* *5*, Length = 10, Power: 1, 946bp - 955bp CTCCTCTGCT

Proscan: Version 1.7:
Processed Sequence: 1002bp. No *Cis* regulatory regions predicted.
Softberry programs:
0 promoter/enhancer(s) predicted
Promoter 2.0 Prediction Server:

Position (bp)	Score	Likelihood
200	0.557	Marginal prediction;

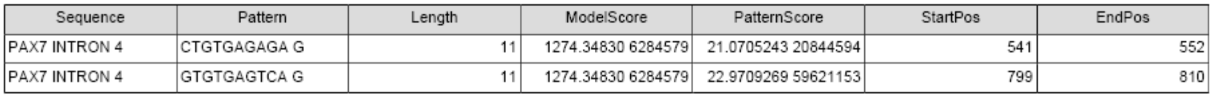
CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 4a & 4b)

Table 4a: CLC Gene Workbench v.1.0.1. Pattern Discovery Search



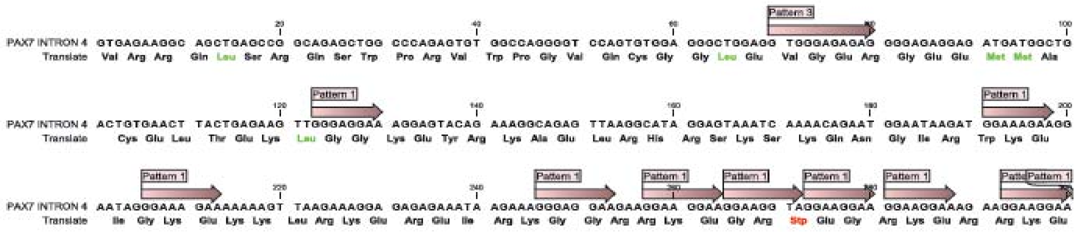
Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 4	GGGAGGAA	8	1355.13354 91683985	20.2233399 74494554	122	130
PAX7 INTRON 4	GGAAAGAA	8	1355.13354 91683985	19.7005324 9513273	190	198
PAX7 INTRON 4	GGAAAGAA	8	1355.13354 91683985	19.7005324 9513273	205	213
PAX7 INTRON 4	GGGAGGAA	8	1355.13354 91683985	20.2233399 74494554	245	253
PAX7 INTRON 4	GGAAGGAA	8	1355.13354 91683985	23.1645652 2709397	256	264
PAX7 INTRON 4	GGAAGGTA	8	1355.13354 91683985	17.5507604 22396078	264	272
PAX7 INTRON 4	GGAAGGAA	8	1355.13354 91683985	23.1645652 2709397	272	280
PAX7 INTRON 4	GGAAGGAA	8	1355.13354 91683985	23.1645652 2709397	280	288
PAX7 INTRON 4	GGAAGGAA	8	1355.13354 91683985	23.1645652 2709397	292	300
PAX7 INTRON 4	GGAAGGAA	8	1355.13354 91683985	23.1645652 2709397	300	308
PAX7 INTRON 4	GGCAGGAA	8	1355.13354 91683985	18.9139383 07175883	328	336
PAX7 INTRON 4	GTAAGGAA	8	1355.13354 91683985	18.3366373 0150939	891	899
PAX7 INTRON 4	TGTGTGTG	8	1300.73142 67724823	23.1470854 38411686	564	572
PAX7 INTRON 4	TGTGTGTA	8	1300.73142 67724823	18.4533111 97633493	572	580
PAX7 INTRON 4	TGTGTGTG	8	1300.73142 67724823	23.1470854 38411686	580	588
PAX7 INTRON 4	TGTGTGTG	8	1300.73142 67724823	23.1470854 38411686	588	596
PAX7 INTRON 4	TGAGTGTG	8	1300.73142 67724823	19.5716325 08941946	635	643
PAX7 INTRON 4	TGTGGGTG	8	1300.73142 67724823	19.4371354 75261723	698	706
PAX7 INTRON 4	TGTGAGTG	8	1300.73142 67724823	20.3766870 37843958	785	793
PAX7 INTRON 4	TGTGTGTG	8	1300.73142 67724823	23.1470854 38411686	861	869
PAX7 INTRON 4	GTGGGAGAGA G	11	1274.34830 6284579	21.3126706 965328	69	80

Table 3b:CLC Gene Workbench v.1.0.1. Pattern Discovery Search



Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 4	CTGTGAGAGA G	11	1274.34830 6284579	21.0705243 20844594	541	552
PAX7 INTRON 4	GTGTGAGTCA G	11	1274.34830 6284579	22.9709269 59621153	799	810

Figure 4



```
PAX7 INTRON 4  GTGAGAAGGC AGCTGAGCCG GCAGAGCTGG CCCAGAGTGT GGCCAGGGGT CCAGTGTGGA GGGCTGGAGG TGGGAGAGAG GGGAGAGGAG ATGATGGCTG
     Translate  Val Arg Arg  Gln Leu Ser Arg  Gln Ser Trp  Pro Arg Val  Trp Pro Gly Val  Gln Cys Gly  Gly Leu Glu  Val Gly Glu Arg  Gly Glu Glu  Met Met Ala

PAX7 INTRON 4  ACTGTGAACT TACTGAGAAG TTGGGAGGAA AGGAGTACAG AAAGGCAGAG TTAAGGCATA GGAGTAAATC AAAACAGAAT GGAATAAGAT GGAAAGAAGG
     Translate  Cys Glu Leu  Thr Glu Lys  Leu Gly Gly  Lys Glu Tyr Arg  Lys Ala Glu  Leu Arg His  Arg Ser Lys Ser  Lys Gln Asn  Gly Ile Arg  Trp Lys Glu

PAX7 INTRON 4  AATAGGGAAA GAAAAAAAGT TAAGAAAGGA GAGAGAAATA AGAAAGGGAG GAAGAAGGAA GGAAGGAAGG TAGGAAGGAA GGAAGGAAAG AAGGAAGGAA
     Translate  Ile Gly Lys  Glu Lys Lys  Leu Arg Lys Glu  Arg Glu Ile  Arg Lys Gly  Gly Arg Arg Lys  Glu Gly Arg  Stp Glu Gly  Arg Lys Glu Arg  Arg Lys Glu
```

Figure 4: *Cis* regulatory region predicted by CLC Gene Workbench v. 1.0 on forward strand from 200-300bps; ***Patterns found within this promoter region.***

PAX7 INTRON ANALYSIS: INTRON 5

PAX7 intron 5 patterns found by Softberry Pattern Search Software: Found 5 pattern(s)

1) *Pattern 1*, Length = 10, Power: 1, 5710bp - 5719bp TTTTTTTTTA
2) *Pattern 2*, Length = 10, Power: 1, 3023bp - 3032bp TATTTTTTTT
3) *Pattern 3*, Length = 10, Power: 1, 3024bp – 3033bp ATTTTTTTTT
4) *Pattern 4*, Length = 10, Power: 1, 50527bp – 50536bp TATTATTTTT
5) *Pattern 5*, Length = 10, Power: 1, 3760bp – 3769bp TATTTATTTT

From the four software programs used, many promoters were predicted (55382 bp).

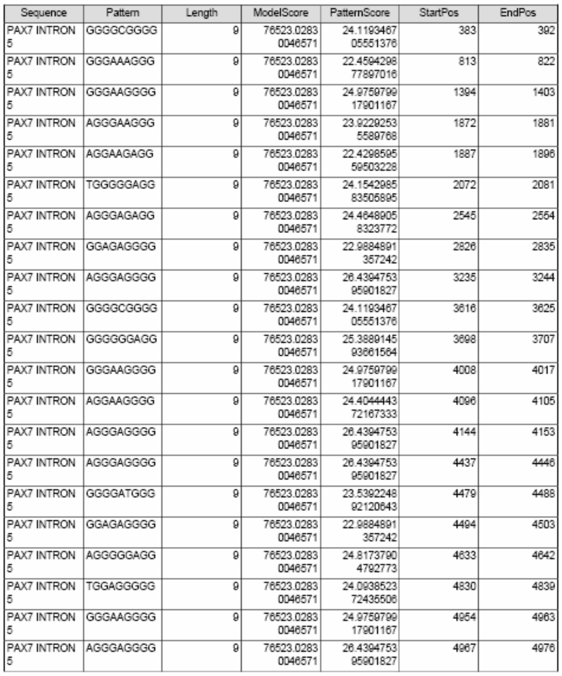
Proscan: Version 1.7 : 100+ promoters predicted

Softberry TSSG: 23 promoter/enhancer(s) predicted

Promoter 2.0 Prediction Server: 45+ promoters predicted

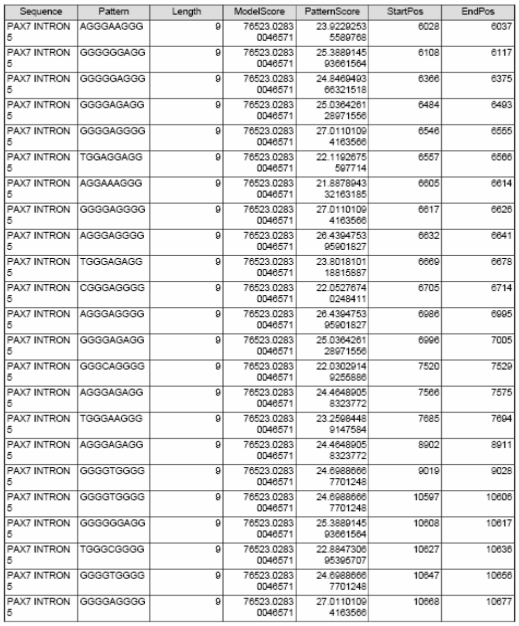
CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 5a & 5b)

Table 5a: CLC Gene Workbench v.1.0.1. Pattern Discovery Search



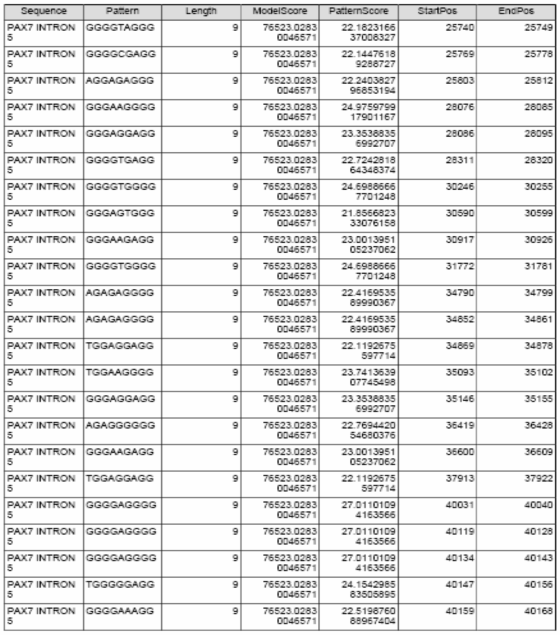
Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 5	GGGGCGGGG	9	76523.02830046571	24.119346705551376	383	392
PAX7 INTRON 5	GGGAAAGGG	9	76523.02830046571	22.459429877897016	813	822
PAX7 INTRON 5	GGGAAGGGG	9	76523.02830046571	24.975979917901167	1394	1403
PAX7 INTRON 5	AGGGAAGGG	9	76523.02830046571	23.92292535589768	1872	1881
PAX7 INTRON 5	AGGAAGAGG	9	76523.02830046571	22.429859559503228	1887	1896
PAX7 INTRON 5	TGGGGGAGG	9	76523.02830046571	24.154298583505895	2072	2081
PAX7 INTRON 5	AGGGAGAGG	9	76523.02830046571	24.46489058323772	2545	2554
PAX7 INTRON 5	GGAGAGGGG	9	76523.02830046571	22.9884891357242	2826	2835
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	3235	3244
PAX7 INTRON 5	GGGGCGGGG	9	76523.02830046571	24.119346705551376	3616	3625
PAX7 INTRON 5	GGGGGGAGG	9	76523.02830046571	25.388914593661564	3698	3707
PAX7 INTRON 5	GGGAAGGGG	9	76523.02830046571	24.975979917901167	4008	4017
PAX7 INTRON 5	AGGAAGGGG	9	76523.02830046571	24.404444372167333	4096	4105
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	4144	4153
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	4437	4446
PAX7 INTRON 5	GGGGATGGG	9	76523.02830046571	23.539224892120643	4479	4488
PAX7 INTRON 5	GGAGAGGGG	9	76523.02830046571	22.9884891357242	4494	4503
PAX7 INTRON 5	AGGGGGAGG	9	76523.02830046571	24.81737904792773	4633	4642
PAX7 INTRON 5	TGGAGGGGG	9	76523.02830046571	24.093852372435506	4830	4839
PAX7 INTRON 5	GGGAAGGGG	9	76523.02830046571	24.975979917901167	4954	4963
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	4967	4976

Table 5b:CLC Gene Workbench v.1.0.1. Pattern Discovery Search



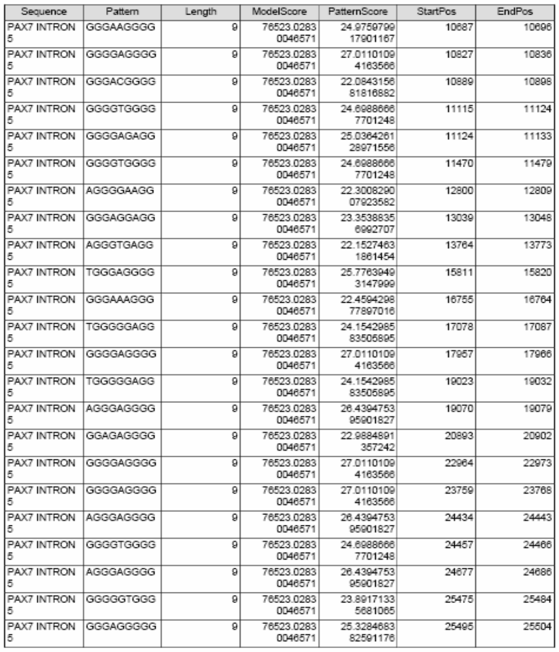
Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 5	AGGGAAGGG	9	76523.02830046571	23.92292535589768	6028	6037
PAX7 INTRON 5	GGGGGGAGG	9	76523.02830046571	25.388914593661564	6108	6117
PAX7 INTRON 5	GGGGGAGGG	9	76523.02830046571	24.846949366321518	6366	6375
PAX7 INTRON 5	GGGGAGAGG	9	76523.02830046571	25.036426128971556	6484	6493
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	6546	6555
PAX7 INTRON 5	TGGAGGAGG	9	76523.02830046571	22.1192675597714	6557	6566
PAX7 INTRON 5	AGGAAAGGG	9	76523.02830046571	21.887894332163185	6605	6614
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	6617	6626
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	6632	6641
PAX7 INTRON 5	TGGGAGAGG	9	76523.02830046571	23.801810118815887	6669	6678
PAX7 INTRON 5	CGGGAGGGG	9	76523.02830046571	22.05276740248411	6705	6714
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	6986	6995
PAX7 INTRON 5	GGGGAGAGG	9	76523.02830046571	25.036426128971556	6996	7005
PAX7 INTRON 5	GGGCAGGGG	9	76523.02830046571	22.03029149255886	7520	7529
PAX7 INTRON 5	AGGGAGAGG	9	76523.02830046571	24.46489058323772	7566	7575
PAX7 INTRON 5	TGGGAAGGG	9	76523.02830046571	23.25984489147584	7685	7694
PAX7 INTRON 5	AGGGAGAGG	9	76523.02830046571	24.46489058323772	8902	8911
PAX7 INTRON 5	GGGGTGGGG	9	76523.02830046571	24.69886667701248	9019	9028
PAX7 INTRON 5	GGGGTGGGG	9	76523.02830046571	24.69886667701248	10597	10606
PAX7 INTRON 5	GGGGGGAGG	9	76523.02830046571	25.388914593661564	10608	10617
PAX7 INTRON 5	TGGGCGGGG	9	76523.02830046571	22.884730695396707	10627	10636
PAX7 INTRON 5	GGGGTGGGG	9	76523.02830046571	24.69886667701248	10647	10656
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	10668	10677

Table 5c: CLC Gene Workbench v.1.0.1. Discovery Search



Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 5	GGGGTAGGG	9	76523.02830046571	22.182316637006327	25740	25749
PAX7 INTRON 5	GGGGCGAGG	9	76523.02830046571	22.14476189286727	25769	25778
PAX7 INTRON 5	AGGAGAGGG	9	76523.02830046571	22.240382796853194	25803	25812
PAX7 INTRON 5	GGGAAGGGG	9	76523.02830046571	24.975979917901167	28076	28085
PAX7 INTRON 5	GGGAGGAGG	9	76523.02830046571	23.35388356992707	28086	28095
PAX7 INTRON 5	GGGGTGAGG	9	76523.02830046571	22.724281864348374	28311	28320
PAX7 INTRON 5	GGGGTGGGG	9	76523.02830046571	24.69886667701248	30246	30255
PAX7 INTRON 5	GGGAGTGGG	9	76523.02830046571	21.856682333076158	30590	30599
PAX7 INTRON 5	GGGAAGAGG	9	76523.02830046571	23.001395105237062	30917	30926
PAX7 INTRON 5	GGGGTGGGG	9	76523.02830046571	24.69886667701248	31772	31781
PAX7 INTRON 5	AGAGAGGGG	9	76523.02830046571	22.416953589990367	34790	34799
PAX7 INTRON 5	AGAGAGGGG	9	76523.02830046571	22.416953589990367	34852	34861
PAX7 INTRON 5	TGGAGGAGG	9	76523.02830046571	22.1192675597714	34869	34878
PAX7 INTRON 5	TGGAAGGGG	9	76523.02830046571	23.741363907745498	35093	35102
PAX7 INTRON 5	GGGAGGAGG	9	76523.02830046571	23.35388356992707	35146	35155
PAX7 INTRON 5	AGAGGGGGG	9	76523.02830046571	22.769442054680376	36419	36428
PAX7 INTRON 5	GGGAAGAGG	9	76523.02830046571	23.001395105237062	36600	36609
PAX7 INTRON 5	TGGAGGAGG	9	76523.02830046571	22.1192675597714	37913	37922
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	40031	40040
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	40119	40128
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	40134	40143
PAX7 INTRON 5	TGGGGGAGG	9	76523.02830046571	24.154298583505895	40147	40156
PAX7 INTRON 5	GGGGAAAGG	9	76523.02830046571	22.519876088967404	40159	40168

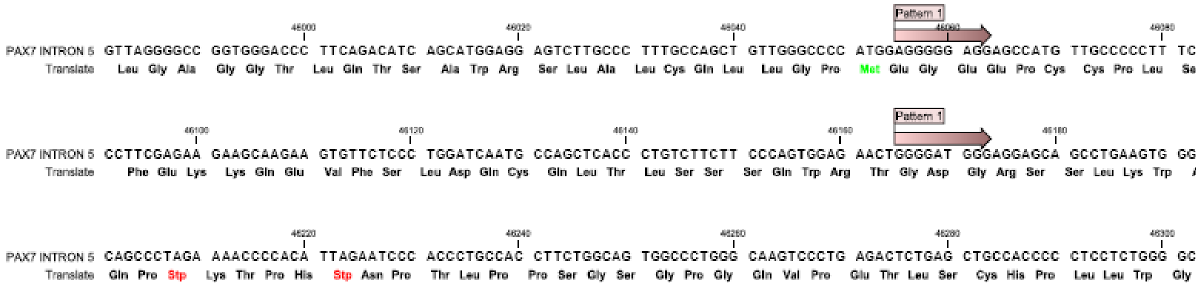
Table 5d: CLC Gene Workbench v.1.0.1. Pattern Pattern Discovery Search



Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 5	GGGAAGGGG	9	76523.02830046571	24.975979917901167	10687	10696
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	10827	10836
PAX7 INTRON 5	GGGACGGGG	9	76523.02830046571	22.084315681816882	10889	10898
PAX7 INTRON 5	GGGGTGGGG	9	76523.02830046571	24.69886667701248	11115	11124
PAX7 INTRON 5	GGGGAGAGG	9	76523.02830046571	25.036426128971556	11124	11133
PAX7 INTRON 5	GGGGTGGGG	9	76523.02830046571	24.69886667701248	11470	11479
PAX7 INTRON 5	AGGGGAAGG	9	76523.02830046571	22.300829007923582	12800	12809
PAX7 INTRON 5	GGGAGGAGG	9	76523.02830046571	23.35388356992707	13039	13048
PAX7 INTRON 5	AGGGTGAGG	9	76523.02830046571	22.15274631861454	13764	13773
PAX7 INTRON 5	TGGGAGGGG	9	76523.02830046571	25.77639493147999	15811	15820
PAX7 INTRON 5	GGGAAAGGG	9	76523.02830046571	22.459429877897016	16755	16764
PAX7 INTRON 5	TGGGGGAGG	9	76523.02830046571	24.154298583505895	17078	17087
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	17957	17966
PAX7 INTRON 5	TGGGGGAGG	9	76523.02830046571	24.154298583505895	19023	19032
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	19070	19079
PAX7 INTRON 5	GGAGAGGGG	9	76523.02830046571	22.9884891357242	20893	20902
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	22964	22973
PAX7 INTRON 5	GGGGAGGGG	9	76523.02830046571	27.01101094163566	23759	23768
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	24434	24443
PAX7 INTRON 5	GGGGTGGGG	9	76523.02830046571	24.69886667701248	24457	24466
PAX7 INTRON 5	AGGGAGGGG	9	76523.02830046571	26.439475395901827	24677	24686
PAX7 INTRON 5	GGGGGTGGG	9	76523.02830046571	23.89171335681065	25475	25484
PAX7 INTRON 5	GGGAGGGGG	9	76523.02830046571	25.328468382591176	25495	25504

CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 5e & 5f)

Figure 5



```
                                                   46000                   46020                   46040                Pattern 1 46060                    46080
PAX7 INTRON 5  GTTAGGGGCC  GGTGGGACCC  TTCAGACATC  AGCATGGAGG  AGTCTTGCCC  TTTGCCAGCT  GTTGGGCCCC  ATGGAGGGGG  AGGAGCCATG  TTGCCCCCTT  TC
    Translate   Leu Gly Ala  Gly Gly Thr  Leu Gln Thr Ser  Ala Trp Arg  Ser Leu Ala  Leu Cys Gln Leu  Leu Gly Pro  Met Glu Gly  Glu Glu Pro Cys  Cys Pro Leu  Se

                                    46100                   46120                   46140                   46160 Pattern 1         46180
PAX7 INTRON 5  CCTTCGAGAA  GAAGCAAGAA  GTGTTCTCCC  TGGATCAATG  CCAGCTCACC  CTGTCTTCTT  CCCAGTGGAG  AACTGGGGAT  GGGAGGAGCA  GCCTGAAGTG  GG
    Translate    Phe Glu Lys  Lys Gln Glu  Val Phe Ser  Leu Asp Gln Cys  Gln Leu Thr  Leu Ser Ser  Ser Gln Trp Arg  Thr Gly Asp  Gly Arg Ser  Ser Leu Lys Trp  /

                                                   46220                   46240                   46260                   46280                   46300
PAX7 INTRON 5  CAGCCCTAGA  AAACCCCACA  TTAGAATCCC  ACCCTGCCAC  CTTCTGGCAG  TGGCCCTGGG  CAAGTCCCTG  AGACTCTGAG  CTGCCACCCC  CTCCTCTGGG  GC
    Translate  Gln Pro Stp  Lys Thr Pro His  Stp Asn Pro  Thr Leu Pro  Pro Ser Gly Ser  Gly Pro Gly  Gln Val Pro  Glu Thr Leu Ser  Cys His Pro  Leu Leu Trp  Gly
```

Figure 5: *Cis* regulatory region predicted by CLC Gene Workbench v. 1.0 on forward strand from 45000-46000bps; ***Patterns found within this cis regulatory region.***

PAX7 INTRON ANALYSIS: INTRON 6

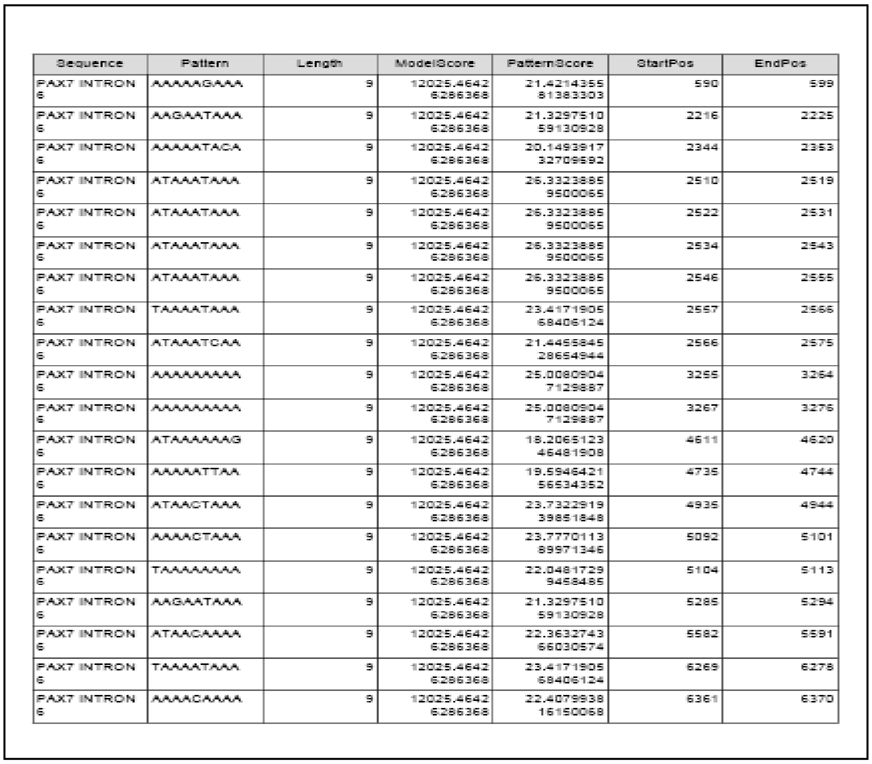
PAX7 intron 6: patterns found by Softberry Pattern Search Software: *No patterns found*

Proscan: Version 1.7
Promoter region predicted on forward strand in 1456 to 1706bps
Promoter region predicted on forward strand in 1791 to 2041bps
Promoter 2.0 Prediction Server:

Position	Score	Likelihood
1000	1.021	Highly likely prediction
3600	1.084	Highly likely prediction
4000	0.584	Marginal prediction
5400	1.262	Highly likely prediction
6800	0.660	Marginal prediction
7400	0.666	Marginal prediction

Softberry TSSG: 2 promoter/enhancer(s) predicted
Promoter Pos: 893 LDF: TATA box at 863bp   AATATATG
Promoter Pos: 5092 LDF: TATA box at 5062bp TATAAATA
Softberry programs:
Promoter Pos: 5092 LDF: TATA box at   5062bp   TATAAATA

CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 6)



Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 6	AAAAAGAAA	9	12025.46426286368	21.421435581383303	590	599
PAX7 INTRON 6	AAGAATAAA	9	12025.46426286368	21.329751059130928	2216	2225
PAX7 INTRON 6	AAAAATACA	9	12025.46426286368	20.149391732709692	2344	2353
PAX7 INTRON 6	ATAAATAAA	9	12025.46426286368	26.33238859500065	2510	2519
PAX7 INTRON 6	ATAAATAAA	9	12025.46426286368	26.33238859500065	2522	2531
PAX7 INTRON 6	ATAAATAAA	9	12025.46426286368	26.33238859500065	2534	2543
PAX7 INTRON 6	ATAAATAAA	9	12025.46426286368	26.33238859500065	2546	2555
PAX7 INTRON 6	TAAAATAAA	9	12025.46426286368	23.417190568406124	2557	2566
PAX7 INTRON 6	ATAAATCAA	9	12025.46426286368	21.445584528654944	2566	2575
PAX7 INTRON 6	AAAAAAAAA	9	12025.46426286368	25.00809047129887	3255	3264
PAX7 INTRON 6	AAAAAAAAA	9	12025.46426286368	25.00809047129887	3267	3276
PAX7 INTRON 6	ATAAAAAAG	9	12025.46426286368	18.206512346481908	4611	4620
PAX7 INTRON 6	AAAAATTAA	9	12025.46426286368	19.594642156534352	4735	4744
PAX7 INTRON 6	ATAACTAAA	9	12025.46426286368	23.732291939851848	4935	4944
PAX7 INTRON 6	AAAACTAAA	9	12025.46426286368	23.777011389971346	5092	5101
PAX7 INTRON 6	TAAAAAAAA	9	12025.46426286368	22.04817299458485	5104	5113
PAX7 INTRON 6	AAGAATAAA	9	12025.46426286368	21.329751059130928	5285	5294
PAX7 INTRON 6	ATAACAAAA	9	12025.46426286368	22.363274366030574	5582	5591
PAX7 INTRON 6	TAAAATAAA	9	12025.46426286368	23.417190568406124	6269	6278
PAX7 INTRON 6	AAAACAAAA	9	12025.46426286368	22.407993816150068	6361	6370

Figure 6

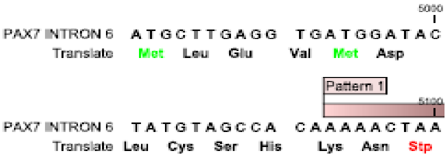

5000
PAX7 INTRON 6 ATGCTTGAGG TGATGGATAC
Translate Met Leu Glu Val Met Asp
Pattern 1
5100
PAX7 INTRON 6 TATGTAGCCA CAAAAACTAA
Translate Leu Cys Ser His Lys Asn Stp

Figure 6: Cis regulatory region predicted by CLC Gene Workbench v. 1.0 on forward strand from 5000-5100bps; ***Pattern found within this cis regulatory region.***

PAX7 INTRON ANALYSIS: INTRON 7

PAX7 intron 7 patterns found by Softberry Pattern Search Software: Found 5 pattern(s)

1) *Pattern* 1, Length = 10, Power: 1, 1134bp – 1143bp CCCTCCCCCT
2) *Pattern* 2, Length = 10, Power: 1, 1133bp – 1142bp TCCCTCCCCC
3) *Pattern* 3, Length = 10, Power: 1, 870bp - 879bp CCCCCCACTC
4) *Pattern* 4, Length = 10, Power: 1, 2157bp – 2166bp CCTTCCCTCC
5) *Pattern* 5, Length = 10, Power: 1, 2156bp – 2165bp CCCTTCCCTC

Proscan: Version 1.7 *No promoter regions predicted.*
Softberry TSSG: 2 promoter/enhancer(s) predicted
Promoter Pos: 1015 LDF: TATA box at 985bp TATAAGAT
Promoter Pos: 333 LDF: TATA box at 304bp TAAAAATC
Promoter 2.0 Prediction Server:

Position	Score	Likelihood
600	0.670	Marginal prediction
1100	0.648	Marginal prediction

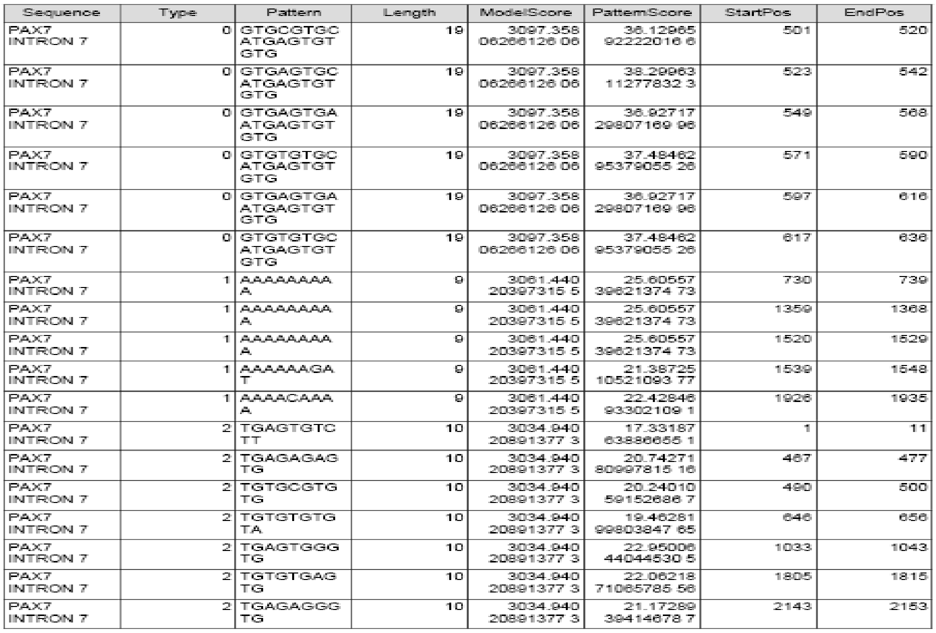
2000	0.661	Marginal prediction

Softberry programs:

Promoter Pos: 333 LDF: TATA box at 304bp TAAAAATC

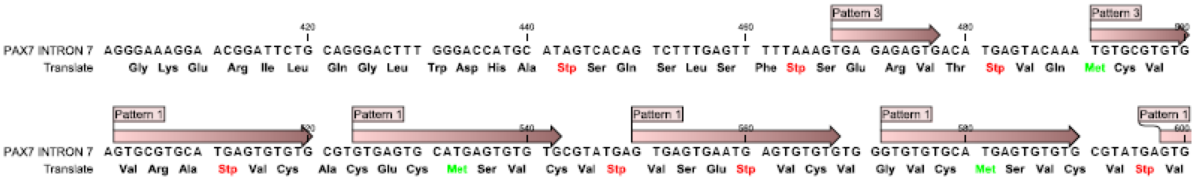
CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 7)

Table 7:CLC Gene Workbench v.1.0.1. Pattern Discovery Search



Sequence	Type	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 7	0	GTGCGTGCATGAGTGTGTG	19	3097.35806266126 06	36.1296592222016 6	501	520
PAX7 INTRON 7	0	GTGAGTGCATGAGTGTGTG	19	3097.35806266126 06	38.2996311277832 3	523	542
PAX7 INTRON 7	0	GTGAGTGAATGAGTGTGTG	19	3097.35806266126 06	36.9271729807169 96	549	568
PAX7 INTRON 7	0	GTGTGTGCATGAGTGTGTG	19	3097.35806266126 06	37.4846295379055 26	571	590
PAX7 INTRON 7	0	GTGAGTGAATGAGTGTGTG	19	3097.35806266126 06	36.9271729807169 96	597	616
PAX7 INTRON 7	0	GTGTGTGCATGAGTGTGTG	19	3097.35806266126 06	37.4846295379055 26	617	636
PAX7 INTRON 7	1	AAAAAAAAA	9	3061.44020397315 5	25.6055739621374 73	730	739
PAX7 INTRON 7	1	AAAAAAAAA	9	3061.44020397315 5	25.6055739621374 73	1359	1368
PAX7 INTRON 7	1	AAAAAAAAA	9	3061.44020397315 5	25.6055739621374 73	1520	1529
PAX7 INTRON 7	1	AAAAAAGAT	9	3061.44020397315 5	21.3872510521093 77	1539	1548
PAX7 INTRON 7	1	AAAACAAAA	9	3061.44020397315 5	22.4284693302109 1	1926	1935
PAX7 INTRON 7	2	TGAGTGTCTT	10	3034.94020891377 3	17.3318763886655 1	1	11
PAX7 INTRON 7	2	TGAGAGAGTG	10	3034.94020891377 3	20.7427180997815 16	467	477
PAX7 INTRON 7	2	TGTGCGTGTG	10	3034.94020891377 3	20.2401059152686 7	490	500
PAX7 INTRON 7	2	TGTGTGTGTA	10	3034.94020891377 3	19.4628199803847 65	646	656
PAX7 INTRON 7	2	TGAGTGGGTG	10	3034.94020891377 3	22.9500644044530 5	1033	1043
PAX7 INTRON 7	2	TGTGTGAGTG	10	3034.94020891377 3	22.0621871065785 56	1805	1815
PAX7 INTRON 7	2	TGAGAGGGTG	10	3034.94020891377 3	21.1728939414678 7	2143	2153

Figure 7

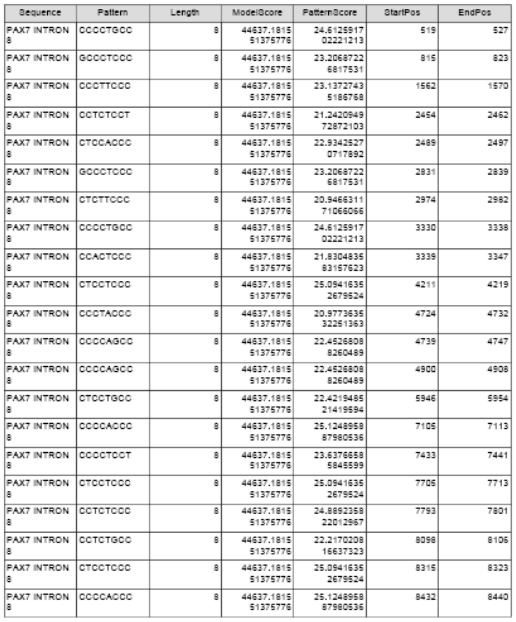


```
                                         420                                   440                                 460                     480                      500
PAX7 INTRON 7  AGGGAAAGGA  ACGGATTCTG  CAGGGACTTT  GGGACCATGC  ATAGTCACAG  TCTTTGAGTT  TTTAAAGTGA  GAGAGTGACA  TGAGTACAAA  TGTGCGTGTG
    Translate     Gly  Lys  Glu   Arg  Ile  Leu   Gln  Gly  Leu   Trp  Asp  His  Ala   Stp  Ser  Gln   Ser  Leu  Ser   Phe  Stp  Ser  Glu   Arg  Val  Thr   Stp  Val  Gln   Met  Cys  Val

                                         520                                   540                                 560                     580                      600
PAX7 INTRON 7  AGTGCGTGCA  TGAGTGTGTG  CGTGTGAGTG  CATGAGTGTG  TGCGTATGAG  TGAGTGAATG  AGTGTGTGTG  GGTGTGTGCA  TGAGTGTGTG  CGTATGAGTG
    Translate     Val  Arg  Ala   Stp  Val  Cys   Ala  Cys  Glu  Cys   Met  Ser  Val   Cys  Val  Stp   Val  Ser  Glu  Stp   Val  Cys  Val   Gly  Val  Cys   Met  Ser  Val  Cys   Val  Stp  Val
```

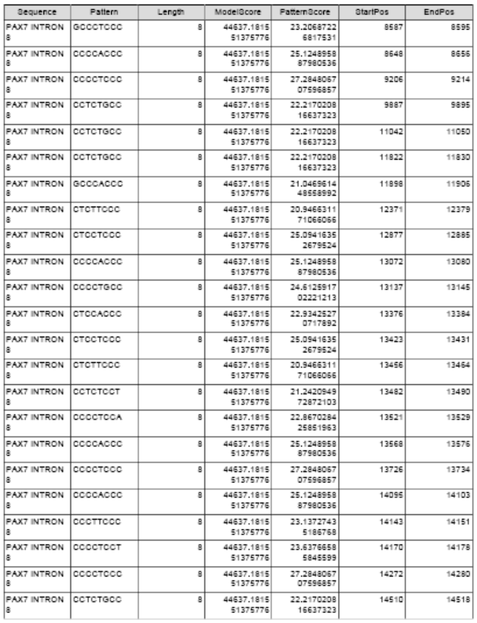
Figure 7: Cis regulatory region predicted by CLC Gene Workbench v. 1.0 on forward strand from 300 -600bps; ***Patterns found within this cis regulatory region.***

Table 8a:CLC Gene Workbench v.1.0.1. Pattern Discovery Search



Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 8	CCCCTGCC	8	44637.181551375776	24.612591702221213	519	527
PAX7 INTRON 8	GCCCTCCC	8	44637.181551375776	23.20687226817531	815	823
PAX7 INTRON 8	CCCTTCCC	8	44637.181551375776	23.13727435186768	1562	1570
PAX7 INTRON 8	CCTCTCCT	8	44637.181551375776	21.242094972872103	2454	2462
PAX7 INTRON 8	CTCCACCC	8	44637.181551375776	22.93425270717892	2489	2497
PAX7 INTRON 8	GCCCTCCC	8	44637.181551375776	23.20687226817531	2831	2839
PAX7 INTRON 8	CTCTTCCC	8	44637.181551375776	20.946631171066066	2974	2982
PAX7 INTRON 8	CCCCTGCC	8	44637.181551375776	24.612591702221213	3330	3338
PAX7 INTRON 8	CCACTCCC	8	44637.181551375776	21.830483583157623	3339	3347
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	4211	4219
PAX7 INTRON 8	CCCTACCC	8	44637.181551375776	20.977363532251363	4724	4732
PAX7 INTRON 8	CCCCAGCC	8	44637.181551375776	22.45268088260489	4739	4747
PAX7 INTRON 8	CCCCAGCC	8	44637.181551375776	22.45268088260489	4900	4908
PAX7 INTRON 8	CTCCTGCC	8	44637.181551375776	22.421948521419594	5946	5954
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	7105	7113
PAX7 INTRON 8	CCCCTCCT	8	44637.181551375776	23.63766585845599	7433	7441
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	7705	7713
PAX7 INTRON 8	CCTCTCCC	8	44637.181551375776	24.889235822012967	7793	7801
PAX7 INTRON 8	CCTCTGCC	8	44637.181551375776	22.217020816637323	8098	8106
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	8315	8323
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	8432	8440

Table 8b: CLC Gene Workbench v.1.0.1. Pattern Discovery Search



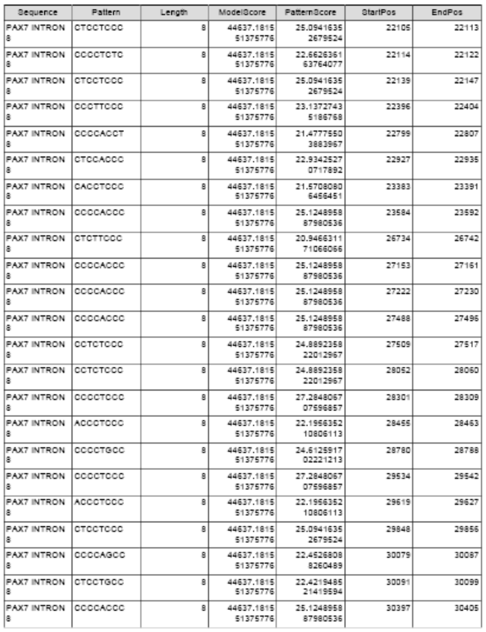
Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 8	GCCCTCCC	8	44637.181551375776	23.20687226817531	8587	8595
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	8648	8656
PAX7 INTRON 8	CCCCTCCC	8	44637.181551375776	27.284806707596857	9206	9214
PAX7 INTRON 8	CCTCTGCC	8	44637.181551375776	22.217020816637323	9887	9895
PAX7 INTRON 8	CCTCTGCC	8	44637.181551375776	22.217020816637323	11042	11050
PAX7 INTRON 8	CCTCTGCC	8	44637.181551375776	22.217020816637323	11822	11830
PAX7 INTRON 8	GCCCACCC	8	44637.181551375776	21.046961448558992	11898	11906
PAX7 INTRON 8	CTCTTCCC	8	44637.181551375776	20.946631171066066	12371	12379
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	12877	12885
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	13072	13080
PAX7 INTRON 8	CCCCTGCC	8	44637.181551375776	24.612591702221213	13137	13145
PAX7 INTRON 8	CTCCACCC	8	44637.181551375776	22.93425270717892	13376	13384
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	13423	13431
PAX7 INTRON 8	CTCTTCCC	8	44637.181551375776	20.946631171066066	13456	13464
PAX7 INTRON 8	CCTCTCCT	8	44637.181551375776	21.242094972872103	13482	13490
PAX7 INTRON 8	CCCCTCCA	8	44637.181551375776	22.867028425851963	13521	13529
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	13568	13576
PAX7 INTRON 8	CCCCTCCC	8	44637.181551375776	27.284806707596857	13726	13734
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	14095	14103
PAX7 INTRON 8	CCCTTCCC	8	44637.181551375776	23.13727435186768	14143	14151
PAX7 INTRON 8	CCCCTCCT	8	44637.181551375776	23.63766585845599	14170	14178
PAX7 INTRON 8	CCCCTCCC	8	44637.181551375776	27.284806707596857	14272	14280
PAX7 INTRON 8	CCTCTGCC	8	44637.181551375776	22.217020816637323	14510	14518

CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 8a & 8b)

PAX7 INTRON ANALYSIS: INTRON 8

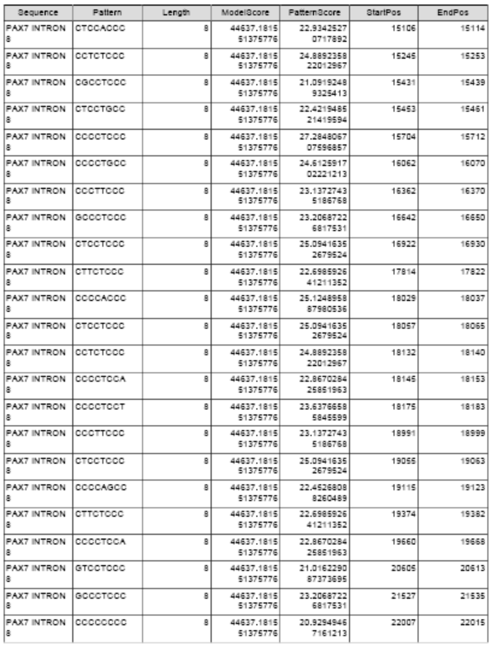
PAX7 intron 8 patterns found by Softberry Pattern Search Software: *No patterns found*
From the other four software programs used, many *cis* elements predicted (32335 bps).
Proscan: Version 1.7: Promoter region predicted on forward strand in 2557 to 2807
TATA found at 2792, Est.TSS = 2822

Table 8c:CLC Gene Workbench v.1.0.1. Pattern Discovery Search



Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	22105	22113
PAX7 INTRON 8	CCCCTCTC	8	44637.181551375776	22.662636163764077	22114	22122
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	22139	22147
PAX7 INTRON 8	CCCTTCCC	8	44637.181551375776	23.13727435186768	22396	22404
PAX7 INTRON 8	CCCCACCT	8	44637.181551375776	21.47775503883967	22799	22807
PAX7 INTRON 8	CTCCACCC	8	44637.181551375776	22.93425270717892	22927	22935
PAX7 INTRON 8	CACCTCCC	8	44637.181551375776	21.57080806456451	23383	23391
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	23584	23592
PAX7 INTRON 8	CTCTTCCC	8	44637.181551375776	20.946631171066066	26734	26742
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	27153	27161
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	27222	27230
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	27488	27496
PAX7 INTRON 8	CCTCTCCC	8	44637.181551375776	24.889235822012967	27509	27517
PAX7 INTRON 8	CCTCTCCC	8	44637.181551375776	24.889235822012967	28052	28060
PAX7 INTRON 8	CCCCTCCC	8	44637.181551375776	27.284806707596857	28301	28309
PAX7 INTRON 8	ACCCTCCC	8	44637.181551375776	22.195635210806113	28455	28463
PAX7 INTRON 8	CCCCTGCC	8	44637.181551375776	24.612591702221213	28780	28788
PAX7 INTRON 8	CCCCTCCC	8	44637.181551375776	27.284806707596857	29534	29542
PAX7 INTRON 8	ACCCTCCC	8	44637.181551375776	22.195635210806113	29619	29627
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	29848	29856
PAX7 INTRON 8	CCCCAGCC	8	44637.181551375776	22.45268088260489	30079	30087
PAX7 INTRON 8	CTCCTGCC	8	44637.181551375776	22.421948521419594	30091	30099
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	30397	30405

Table 8d:CLC Gene Workbench v.1.0.1. Pattern Discovery Search

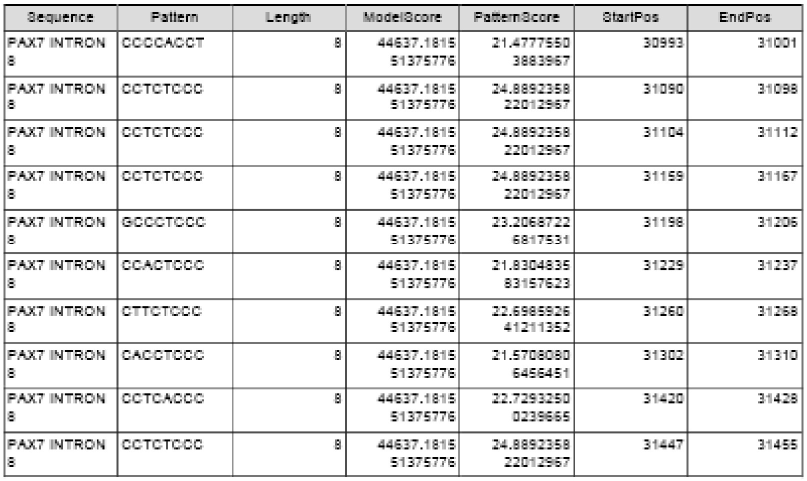


Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 8	CTCCACCC	8	44637.181551375776	22.93425270717892	15106	15114
PAX7 INTRON 8	CCTCTCCC	8	44637.181551375776	24.889235822012967	15245	15253
PAX7 INTRON 8	CGCCTCCC	8	44637.181551375776	21.09192489325413	15431	15439
PAX7 INTRON 8	CTCCTGCC	8	44637.181551375776	22.421948521419594	15453	15461
PAX7 INTRON 8	CCCCTCCC	8	44637.181551375776	27.284806707596857	15704	15712
PAX7 INTRON 8	CCCCTGCC	8	44637.181551375776	24.612591702221213	16062	16070
PAX7 INTRON 8	CCCTTCCC	8	44637.181551375776	23.13727435186768	16362	16370
PAX7 INTRON 8	GCCCTCCC	8	44637.181551375776	23.20687226817531	16642	16650
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	16922	16930
PAX7 INTRON 8	CTTCTCCC	8	44637.181551375776	22.698592641211352	17814	17822
PAX7 INTRON 8	CCCCACCC	8	44637.181551375776	25.124895887980536	18029	18037
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	18057	18065
PAX7 INTRON 8	CCTCTCCC	8	44637.181551375776	24.889235822012967	18132	18140
PAX7 INTRON 8	CCCCTCCA	8	44637.181551375776	22.867028425851963	18145	18153
PAX7 INTRON 8	CCCCTCCT	8	44637.181551375776	23.63766585845599	18175	18183
PAX7 INTRON 8	CCCTTCCC	8	44637.181551375776	23.13727435186768	18991	18999
PAX7 INTRON 8	CTCCTCCC	8	44637.181551375776	25.09416352679524	19055	19063
PAX7 INTRON 8	CCCCAGCC	8	44637.181551375776	22.45268088260489	19115	19123
PAX7 INTRON 8	CTTCTCCC	8	44637.181551375776	22.698592641211352	19374	19382
PAX7 INTRON 8	CCCCTCCA	8	44637.181551375776	22.867028425851963	19660	19668
PAX7 INTRON 8	GTCCTCCC	8	44637.181551375776	21.016229087373695	20605	20613
PAX7 INTRON 8	GCCCTCCC	8	44637.181551375776	23.20687226817531	21527	21535
PAX7 INTRON 8	CCCCCCCC	8	44637.181551375776	20.92949467161213	22007	22015

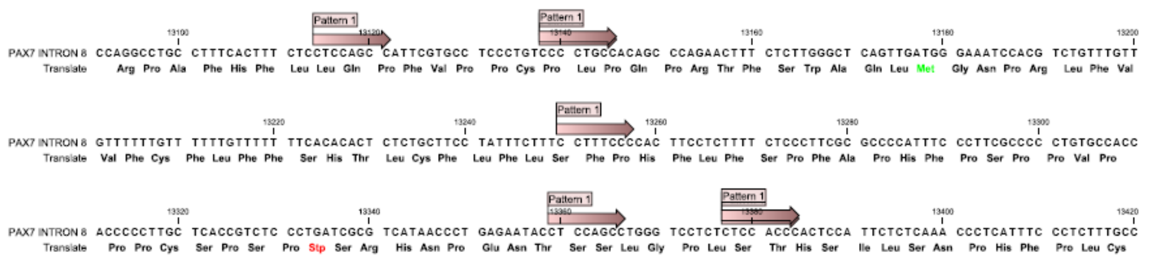
Softberry TSSG: *No promoter regions predicted*
Promoter 2.0 Prediction Server: 31 promoters predicted

CLC Gene Workbench v.1.0.1. Pattern Discovery Search (Table 8e)

Table 8e:CLC Gene Workbench v.1.0.1. Pattern Discovery Search



Sequence	Pattern	Length	ModelScore	PatternScore	StartPos	EndPos
PAX7 INTRON 8	CCCCACCT	8	44637.1815 51375776	21.4777550 3883967	30993	31001
PAX7 INTRON 8	CCTCTCCC	8	44637.1815 51375776	24.8892358 22012967	31090	31098
PAX7 INTRON 8	CCTCTCCC	8	44637.1815 51375776	24.8892358 22012967	31104	31112
PAX7 INTRON 8	CCTCTCCC	8	44637.1815 51375776	24.8892358 22012967	31159	31167
PAX7 INTRON 8	GCCCTCCC	8	44637.1815 51375776	23.2068722 6817531	31198	31206
PAX7 INTRON 8	CCACTCCC	8	44637.1815 51375776	21.8304835 83157623	31229	31237
PAX7 INTRON 8	CTTCTCCC	8	44637.1815 51375776	22.6985926 41211352	31260	31268
PAX7 INTRON 8	CACCTCCC	8	44637.1815 51375776	21.5708080 6456451	31302	31310
PAX7 INTRON 8	CCTCACCC	8	44637.1815 51375776	22.7293250 0239665	31420	31428
PAX7 INTRON 8	CCTCTCCC	8	44637.1815 51375776	24.8892358 22012967	31447	31455



```
PAX7 INTRON 8  CCAGGCCTGC  CTTTCACTTT  CTCCTCCAGC  CATTCGTGCC  TCCCTGTCCC  CTGCCACAGC  CCAGAACTTT  CTCTTGGGCT  CAGTTGATGG  GAAATCCACG  TCTGTTTGTT
    Translate    Arg Pro  Ala  Phe His Phe  Leu Leu Gln  Pro Phe Val Pro  Pro Cys Pro  Leu Pro Gln  Pro Arg Thr Phe  Ser Trp Ala  Gln Leu Met  Gly Asn Pro Arg  Leu Phe Val

PAX7 INTRON 8  GTTTTTTGTT  TTTTGTTTTT  TTCACACACT  CTCTGCTTCC  TATTTCTTTC  CTTTCCCCAC  TTCCTCTTTT  CTCCCTTCGC  GCCCCATTTC  CCTTCGCCCC  CTGTGCCACC
    Translate  Val Phe Cys  Phe Leu Phe Phe  Ser His Thr  Leu Cys Phe  Leu Phe Leu Ser  Phe Pro His  Phe Leu Phe  Ser Pro Phe Ala  Pro His Phe  Pro Ser Pro  Pro Val Pro

PAX7 INTRON 8  ACCCCCTTGC  TCACCGTCTC  CCTGATCGCG  TCATAACCCT  GAGAATACCT  CCAGCCTGGG  TCCTCTCTCC  ACCCACTCCA  TTCTCTCAAA  CCCTCATTTC  CCTCTTTGCC
    Translate  Pro Pro Cys  Ser Pro Ser  Pro Stp Ser Arg  His Asn Pro  Glu Asn Thr  Ser Ser Leu Gly  Pro Leu Ser  Thr His Ser  Ile Leu Ser Asn  Pro His Phe  Pro Leu Cys
```

Figure 8: Cis regulatory region predicted by CLC Gene Workbench v. 1.0 on forward strand in 13487 to 13737bps; ***Patterns found within this cis regulatory region.***

Table 9. SUMMARY OF MOST LIKELY *CIS* REGULATORY SEQUENCES PREDICTED FOR EACH INTRON OF *PAX7*.

Intron number	Pattern (cis element?)	Length(bp)	Start Position in intron	End Position in intron
PAX7 INTRON 1	GAGGAGAG	8	1284	1292
PAX7 INTRON 3	GGAAAGAA	8	190	198
PAX7 INTRON 4	GGAAAGAA	8	205	213
PAX7 INTRON 5	AGGGGGAGG	9	46054	46063
	GGGGATGGG	9	46164	46173
PAX7 INTRON 6	AAAACTAAA	9	5092	5101

PAX7 INTRON 7	GTGAGTGCATGAGTGTGTG	19	523	542
PAX7 INTRON 8	CCCCACCC	8	13072	13080
	CCCCTGCC	8	13137	13145
	CTCCACCC	8	13376	13384
	CTCCTCCC	8	13423	13431

The results shown in table 9 are the predicted *cis*-regulatory elements for *PAX7* and were chosen from the results of computer scans and based on the four criteria listed above. Table 10 displays the transcription factors most likely to bind to these *cis*-elements with the exception of the *cis*-element in intron 5 for which no transcription factor was identified. Transcription factors were identified using the TRANSFAC database. The transcription factors previously identified as being associated with tumourigenesis are indicated in Table 10.

Table 10. SUMMARY OF MOST LIKELY *CIS* REGULATORY SEQUENCES PREDICTED FOR EACH INTRON OF *PAX7* WITH CORRESPONDING TRANSCRIPTION FACTORS

Intron Number	**Cis element**	**Binding Transcription factor from Transfactor** ***
PAX7 INTRON 1	GAGGAGAG	EBNA-1;RAR-gamma; R2; Zmhoxla
PAX7 INTRON 3	GGAAAGAA	NP-TCII; NF-1; GT-IIA
PAX7 INTRON 4	GGAAAGAA	NP-TCII; NF-1; GT-IIA
PAX7 INTRON 5	AGGGGGAGG	Six-3: DR1; CACCC-BF; CAC-BF; Sp1; ADR1
PAX7 INTRON 5	GGGGATGGG	NONE
PAX7 INTRON 6	AAAACTAAA	SRY; PHO2
PAX7 INTRON 7	GTGAGTGCATGAGTGTGTG	Zeste; GCN4; Zeste; MEP-1; MBF-I; Sp1; GHF-1; Pit-1a; RAP1/SBF-E/TUF; USF; TEF;TTF-1
PAX7 INTRON 8	**CCCCACCC	TEF2;MIG1; ACCC-BF; AP-2; CAC-BF; Sp1
PAX7 INTRON 8	*CCCCTGCC	AP-2; CAC-BF; Ttk;LVc
PAX7 INTRON 8	**CTCCACCC	CACCC-BF; CAC-BF; Sp1
PAX7 INTRON 8	*CTCCTCCC	CAC-BF; ADR1; Sp1

**Found in NF1 & PAX3;*
*** Found in PAX3*
****Transcription factors in blue are associated with tumorigenesis.*
http://www.mdcberlin.de/forschung/schwerpunkte/cancer/rosenbauer.htm

2. Conservation of intron 8 region containing novel *cis* regulatory region indicating possible functional significance

To ascertain the possible functional significance of the putative *cis*-elements identified above, sequences surrounding these *cis*-elements were used to search for conservation of the regulatory region in other cancer related human genes, such as in human *PAX3*. Comparisons between DNA sequences of *PAX3* and *PAX7* can be used to determine the relationship between the gene sequences from which functional or regulatory regions can be ascertained which assist with identification of the functions of *PAX7* and its role in tumorigenesis.

Figure 8

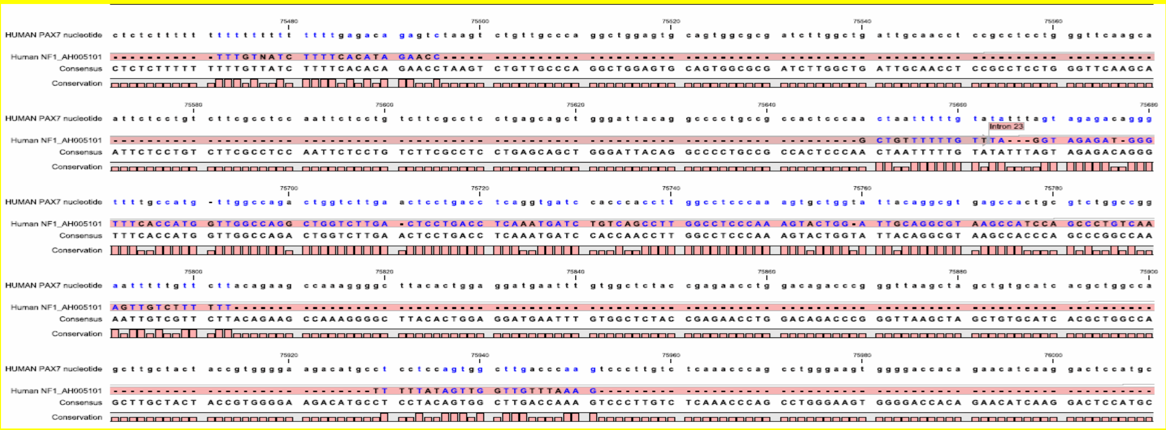


Figure 9: Conserved sequence in intron 8 of *PAX7* also found in intron 23 of the human *NF1* gene (GenBank Accession Number for *NF-1* gene is AH005101). Conserved regions are shown by sequences coloured in blue. The novel *cis*-element identified in intron 8 of *PAX7* is also found in this conserved sequence in intron 10 of *PAX3* at nucleotide 71560 .

One *cis*-element identified in intron 8 was found to be located within a conserved sequence that is also present in alternative intron 10 of human *PAX3* as well as within intron 23 of NF1 (sequence length ~100 bps.). The sequence, approximately 155 nucleotides in length is highly conserved between *PAX7* and *NF-1* (89% conserved) (Figure 9) and between *PAX3* and NF1 (72% conserved) (Figure 10).

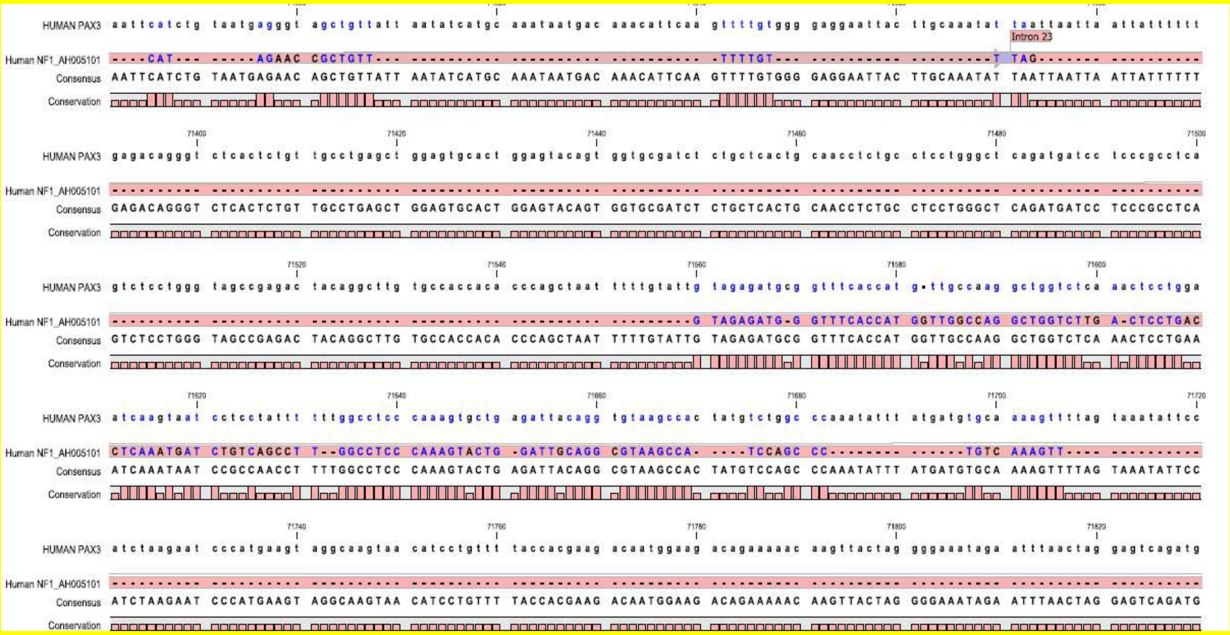


Figure10. Conserved sequence in intron 8 of *PAX7* and intron 23 of *NF-1* also found in alternative intron 10 of *PAX3* (GenBank Accession Number for *PAX3* gene is NM_013942). Conserved regions are shown by sequences coloured in blue.

DISCUSSION

In this paper we have identified novel *cis*-elements in intronic regions of human *PAX7*. We have also identified a conserved intronic region of *PAX7* that is present in introns of *PAX3* and *NF-1*. Moreover, the conserved region contains a newly identified *PAX7 cis*-element and the same *cis* element occurs in the conserved sequence in all three genes. These findings highlight the ability of *in silico* methodologies to uncover putative *cis* regulatory regions. In addition, the sequence alignments performed in this article confirm that patterns of conservation can be useful in identifying regulatory regions.

Cis elements are known to be important in upregulation of genes or in splicing of intronic regions (Pethe et al., 1999; Martin et al., 2004) and therefore crucial in the tumorigenic functions of a gene. The region we have identified in intron 8 of *PAX7*, also found in intron 23 of *NF-1* and alternative intron 10 of *PAX3* may contain regulatory functions common to all three genes and it seems probable that transcription factors and/or spliceosomes would act similarly on all three genes.

Recent experiments identify specific sequences in *NF-1* as being associated with increased tumorigenicity in the childhood cancer, alveolar rhabdomyosarcoma (Dei Tos, et al., 1997). Similarly, *PAX7* and *PAX3* are associated with alveolar rhabdomyosarcoma (Sorensen, et al, 2002). The intronic sequence common to all there genes may be implicated in their tumorigenic properties.

The conserved sequence containing the *cis* regulatory element identified in intron 8 of *PAX7*, intron 10 of *PAX3* and intron 23 of *NF-1* may have arisen by insertion of a regulatory element in all three gene regions or by homologous recombination between chromosome 1 (*PAX7*), chromosome 2 (*PAX3*) and/or chromosome 17 (*NF-1*). The significance of this finding is currently being investigated further by *in vitro* studies**.**

Only in recent literature has there been a spark to delve into the intronic regions of genomic sequences (Oguzkan, et al, 2006). Historically, introns have been viewed as non-coding, nonsense "place holders" between the exons of a given gene (Bernett et al., 2003; de Roos et al, 2005). It was not until the great race for decoding the human genome that researchers realized that introns constitute a large portion of the regulatory regions of the genome (Davies, 2001; Patrinos, 2001). This can only lead one to believe that the once overlooked introns may play a significant role in regulation of cell functions such as cell cycle control, apoptosis, or aberrant cell cycle control as in tumorigenesis. The research performed in this paper represents a cornerstone in *in silico* research of gene sequences as it points the way for future bench work studies so that the findings can be verified and validated.

In conclusion, the results presented here may present significant findings that can be utilised ultimately for the development of therapeutics for the treatment of alveolar rhabdomyosarcoma and other cancers associated with *PAX7*. Furthermore, the methods and findings may have implications for other diseases and other genes. *In silico* biology is currently used by pharmaceutical companies to facilitate and hasten the development of new therapeutics for many diseases.

Acknowledgment

We would like to thank the Memorial Sloan Kettering Institute DNA Sequencing Core Facility Lab manager and technician (Dr. Diane Tabarini & Ms. Donna Wong, respectively) and Memorial Sloan Kettering Cancer Center's Diagnostic Molecular Genetics Laboratory Director (Cindy Yee, PhD) and technologists (Mrs. Nilofer Kreonidis, Mr. Ruben Bacares and Ms. Nadia Shaukat) for their support and assistance for the clinical portions of this project. I would also like to extend my deepest gratitude to Mr. Geoffrey Jackson, of the Centers for Disease Control & Prevention (CDC) who provided the much needed equipment for this project.

Correspondence to:
Maika G. Blackman-Mitchell
Memorial Sloan Kettering Cancer Center
Rockefeller Research Laboratories
430 East 67th Street
4th Floor, Room 453
New York City, New York 10021, USA
Email: blackmam@mskcc.org

REFERENCES

1. Glaser, T., Jepeal, L., Edwards, J. G., Young, S. R., Favor, J., and Maas, R. L. (1994). "PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects." Nat. Genet 7:463-471.
2. Relaix, F., Rocancourt, D. et al. (2004). "Divergent functions of murine Pax3 and Pax7 in limb muscle development." Genes Dev. 18(9): 1088-1105.
3. Chi, N., Epstein J.A. (2002). "Getting your Pax straight: Pax proteins in development and disease." Trends Genet. 18(1): 41-7.
4. Barr, F. G., Fitzgerald, J. C. et al. (1999). "Predominant Expression of Alternative PAX3 and PAX7 Forms in Myogenic and Neural Tumor Cell Lines." Cancer Res. 59(21): 5443-5448.
5. Mercado, G. E., Barr, F.G. (2005). "Looking Downstream of Sarcoma-Associated Chimeric Transcription Factors: When is a Target Really a Target?" Cancer Biol. Ther. 4(4): 456-8.
6. Macina R. A., Barr F. G., Galili N., Riethman H. C. (1995). Genomic organization of the human *PAX3* gene: DNA sequence analysis of the region disrupted in alveolar rhabdomyosarcoma. Genomics, *26:* 1-8.
7. Goulding M. D., Chalepakis G., Deutsch U., Erselius J. R., Gruss P.(1991) Pax-3, a novel murine DNA binding protein expressed during early neurogenesis. EMBO J., *10:* 1135-1147.

8. Bennicelli J. L., Guerry D. I. (1993). Production of multiple cytokines by cultured human melanomas. Exp. Dermatol., *2:* 186-190.
9. Schulte T. W., Toretsky J. A., Ress E., Helman L., Neckers L. M. (1997) Expression of PAX3 in Ewing's sarcoma family of tumors. Biochem. Mol. Med., *60:* 121-126.
10. Hadjistilianou T, Mastrangelo D, Gragnoli A, Capretti MC, De Francesco S, Galluzzi P. (2002) Letter to the editor: neurofibromatosis type 1 (NF 1) associated with embryonal rhabdomyosarcoma of the orbit. Med Pediatr Oncol.. **38**. *6*:449.
11. Dei Tos AP, Dal Cin P. (1997) The role of cytogenetics in the classification of soft tissue tumours. Virchows Arch. 431(2):83-94.
12. Woodruff JM, Christensen WN. Glandular peripheral nerve sheath tumors. *Cancer*. (1993). **72**. *12*:3618-28
13. Oguzkan, et al., (2006) Two neurofibromatosis type 1 cases associated with Rhabdomyosarcoma of bladder, one with a large deletion in the NF1 gene., *Cancer Genetics and Cytogenetics 164*, 159–163
14. Park, Ben Ho and Vogelstein, Bert (2003) Cancer Medicine Volume 6. BC Decker Inc. Hamilton, London
15. Lewin, Benjamin Genes VII**.** **(**2000**)****.** Oxford University Press
16. Robson, Ewan J. D., He, Shu-Jie, **(**2006)A Panorama of PAX Genes in Cancer and Development., Nat Rev Cancer. 6(1):52-62
17. Frith, et al.,(2004) Spouge3 and Zhiping Weng Finding functional sequence elements by multiple local alignment Nucleic Acids Research, , Vol. 32, No. 1 189±200
18. Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. J. Am. Stat. Assoc., 90, 1156±1170.
19. Chen, Jianjun, Sun, Miao, Rowley, Janet D, Hurst, Laurence D**,** (2005)The small introns of antisense genes are better explained by selection for rapid transcription than by 'genomic design', Genetics**,** 171(4):2151-5.
20. Martin, Natalia, Patel, Satyakam, Segre, Julia (2004) A. Long-range comparison of human and mouse Sprr loci to identify conserved noncoding sequences involved in coordinate regulation. Genome Res. 14: 2430-2438
21. Vaiju Pethe, and P. V. Malathy Shekhar (1999). Estrogen Inducibility of c-Ha-ras Transcription in Breast Cancer Cells. Identification of Functional Estrogen-Responsive Transcriptional Regulatory Elements in Exon 1/Intron 1 of the c-Ha-ras Gene. J. Biol. Chem. 274: 30969-30978.
22. Dei Tos AP, Dal Cin P. (1997). The role of cytogenetics in the classification of soft tissue tumours. Virchows Arch. Aug;431(2):83-94. Review
23. Sorensen, Poul H.B., Lynch, James C., Qualman, Stephen J., Tirabosco, Roberto, Lim, Jerian F., Maurer, Harold M., Bridge, Julia A., Crist, William M., Triche, Timothy J., Barr, Frederic G. (2002). PAX3-FKHR and PAX7-FKHR Gene Fusions Are Prognostic Indicators in Alveolar Rhabdomyosarcoma: A Report From the Children's Oncology Group J Clin Oncol . 20: 2672-2679
24. Albert D. G. de Roos. (2005) Origins of introns based on the definition of exon modules and their conserved interfaces Bioinformatics Advance Access, DOI 10.1093/bioinformatics/bth475.Bioinformatics 21: 2-9.
25. Bernett T. K. Lee, Tin Wee Tan, and Shoba Ranganathan (2003) MG AlignIt: a web service for the alignment of mRNA/EST and genomic sequences. Nucleic Acids Res.; 31(13): 3533–3536.
26. Ari Patrinos. (2001). Initial sequencing and analysis of the human genome. The Genome International Sequencing Consortium."*Nature* 409, 860-921.
27. Kevin Davies., (2001) Cracking The Genome: Inside The Race To Unlock Human DNA. Free Press, A division of Simon & Schuster, Inc.