Chapter 6

Tissue Specific Transcript Annotation and Expression Profiling with Complimentary Next-generation Sequencing Technologies

Matthew S. Hestand\textsuperscript{1,2}, Andreas Klingenhoff\textsuperscript{3}, Matthias Scherf\textsuperscript{3}, Yavuz Ariyurek\textsuperscript{2}, Yolande Ramos\textsuperscript{4}, Wilbert van Workum\textsuperscript{5}, Makoto Suzuki\textsuperscript{6}, Thomas Werner\textsuperscript{3}, Gert-Jan B. van Ommen\textsuperscript{1}, Johan T. den Dunnen\textsuperscript{1,2}, Matthias Harbers\textsuperscript{6}, Peter A.C. ’t Hoen\textsuperscript{1}

\textsuperscript{1}The Center for Human and Clinical Genetics, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
\textsuperscript{2}Leiden Genome Technology Center, Leiden University Medical Center, Postzone S4-0P, PO Box 9600, 2300 RC Leiden, The Netherlands.
\textsuperscript{3}Genomatix Software GmbH, Munich, Germany.
\textsuperscript{4}Department of Molecular Cell Biology, Leiden University Medical Centre, 2300 RC Leiden, The Netherlands.
\textsuperscript{5}ServiceXS B.V., Plesmanlaan 1D, 2333 BZ Leiden, The Netherlands
\textsuperscript{6}DNAFORM Inc., Leading Venture Plaza-2, 75-1, Ono-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0046, Japan.

\textit{manuscript submitted}
6.1 Abstract

Next-generation sequencing is excellently suited to evaluate the abundance of mRNAs to study gene expression. Here we compare two alternative technologies, cap analysis of gene expression (CAGE) and serial analysis of gene expression (SAGE), for the same RNA samples. Along with quantifying gene expression levels, CAGE can be used to identify tissue-specific transcription start sites, while SAGE monitors 3’ end usage. We used both methods to get more insight into the transcriptional control of myogenesis studying differential gene expression in differentiated and proliferating C2C12 myoblast cells with statistical evaluation of reproducibility and differential gene expression. Both CAGE and SAGE provided highly reproducible data (Pearson correlations > 0.92 between biological triplicates). With both methods we found around 10,000 genes expressed at levels > 2 transcripts per million (~0.3 copies per cell), with an overlap of 86%. We identified 4,304 and 3,846 genes differentially expressed between proliferating and differentiated C2C12 cells by CAGE and SAGE respectively, with an overlap of 2,144. We identified 196 novel regulatory regions with preferential use in proliferating or differentiated cells. Next-generation sequencing of CAGE and SAGE libraries provides consistent expression levels and can enrich current genome annotations with tissue-specific promoters and alternative 3’ UTR usage.
6.2 Introduction

Next-generation sequencing (NGS) platforms have provided us with the technology needed to expand genomic methods to a new scale. Depending on the technology, these machines can produce gigabases of sequences per day. Due to its superior resolution and sensitivity, NGS is increasingly used to replace array technologies, in particular the genome-wide evaluation of chromatin immunoprecipitation (ChIP-seq) and gene expression profiling experiments. Sequence-based expression analysis can be performed using several approaches. The traditional SAGE (serial analysis of gene expression) method (28), starts with capturing RNA poly-A tails with oligo(dT) beads. Double-stranded cDNA synthesis is performed and a digestion with a restriction enzyme, commonly NlaIII (32), is performed. With the fragments resulting from the digestion only the most 3’ fragment is retained. An additional restriction digest is then performed with MmeI (cuts ∼20 base pairs downstream) to create a fragment of acceptable length for sequencing. In the original method short cDNA fragments, each representing the 3’ most NlaIII digestion site of a specific transcript, were concatenated and cloned, followed by traditional sequencing. However, now the concatenation and cloning steps can be omitted. Instead SAGE library sequences are directly equipped with appropriate sequencing linkers and analyzed in next-generation sequencers (30).

An alternative method is CAGE (cap analysis of gene expression) (29), specifically designed to study gene expression at transcription initiation sites by capturing 5’ ends of mRNAs. After trapping the 5’ cap-structures of mRNAs, sequences are converted to double-stranded cDNA and equipped with a linker containing a restriction site for the enzyme MmeI (or EcoP15I) that cuts ∼20 (or 25-27) base pairs downstream to create a fragment of appropriate length for sequencing and for mapping. Thus where SAGE captures the 3’ most NlaIII digestion site of mRNA and is thus 3’ end biased, CAGE tags represent the ultimate 5’ end of the transcript and indicate the genomic transcription start site (TSS). In both SAGE and CAGE, one transcript is only represented by a single read and (next-generation) sequencing of SAGE and CAGE libraries is therefore referred to as Digital Gene Expression profiling or DeepSAGE and DeepCAGE (30; 31). For simplicity we refer to these in this manuscript simply as SAGE and CAGE. In RNASeq (144), which starts with random fragmentation of the RNA or cDNA, the entire transcript is sequenced. Consequently, a transcript is commonly represented by multiple reads and the amount of reads is dependent on the transcript length. RNASeq gives more detailed information about the structure of the transcripts and alternative splicing, in particular when combined with paired end sequencing, while CAGE is more suitable for analysis of alternative transcription start sites and SAGE for analysis of alternative polyadenylation sites.

Myogenesis is an essential process for muscle development and regeneration, with defects resulting in diseases such as muscular dystrophies. To support our studies towards treatment of muscle-related diseases, we have performed extensive analysis of muscle-derived gene expression profiles (145; 146; 147). This included the analysis of muscle differentiation using a well-established model, the mouse myoblast cell line (C2C12) (148). Two primary transcription factors (TFs) regulating this process are MyoD and Myogenin, but many other regulatory elements have been identified (reviewed in Pownall et al. 2002 (24) and Sartolli and Caretti 2005 (149)). For a better
understanding of how expression profiles change during adaptation to different biological situations, it is important to consider promoter activities and their regulation. Several bioinformatic approaches have been designed for this, including CORE_TF (76) and oPOSSUM (60), searching for shared TF binding sites (TFBSs) in the promoter region. However, these approaches critically depend on correct genome annotations regarding TSSs, which can vary by tissue type. Unfortunately, most studies performed thus far use methods directed at the 3′ end of RNA transcripts (including the well known oligo dT primed cDNA synthesis). Consequently gene annotation is weakest at the 5′ end. CAGE is therefore excellently suitable for the identification of alternative TSSs and putative regulatory regions upstream of those TSSs. We applied both CAGE and SAGE to study muscle differentiation to assess their concordance in estimation of gene expression levels and complementarity in gene annotation.

6.3 Materials and Methods:

6.3.1 Cells, RNA Isolation, and Differentiation Markers

Proliferating C2C12 mouse myoblasts were grown out on collagen coated plates in Dulbecco’s modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS). To induce fusion into myotubes cells were serum deprived by changing to a medium of DMEM supplemented with 2% FBS for nine days (referred to as differentiated cells).

For CAGE and SAGE, RNA was isolated from proliferating and differentiated cells. RNA was isolated from three independent cultures (biological triplicates). Cells grown in (175 cm²) flasks where harvested by trypsinization and centrifugation before RNA extraction with a Nucleospin RNA L kit from Macherey-Nagel. RNA quality was high, as determined with Agilent’s Lab-on-chip total RNA nano assay (RNA integrity number >9). Myogenic properties of the cells were confirmed in RT-PCR/qPCR experiments using primer sets (Additional Table 6.1) specific for Myod1, Myogenin, GAPDH, and HPRT. RT-PCR experiments were performed using oligo dT priming for cDNA synthesis and qPCR was carried out using a Roche Lightcycler 480.

6.3.2 Library Preparation and Next-Generation Sequencing

Separate CAGE libraries were prepared as described previously (31) for each individual RNA sample. The following modifications to the protocol were made: we used modified adapters in the 5′ and 3′ end ligation steps that have linker sequences (proliferating - CCGACAGGTTCAGAGTTCTACAGAGACAGCAG and differentiated - CCGACAGGTTCAGAGTTCTACAGCTTCAGCAG) for Illumina Genome Analyzer II sequencing and have a recognition site for EcoP15I used instead of MmeI.

SAGE libraries were prepared for each individual RNA sample with a FC-102-1005 DGE-Tag Profiling NlaIII SamplePrepKit from Illumina.

Each CAGE and SAGE library was then sequenced on an individual lane on an Illumina Genome Analyzer II for 36 cycles. One CAGE sample from each time point was also sequenced a second time with 32 cycles.
6.3 Materials and Methods:

6.3.3 Initial Sequence Analysis

All sequenced lanes were run through the initial Illumina Genome Analyzer Pipeline (Firecrest ⇒ Bustard ⇒ Gerald) for image analysis and quality control, yielding one scarf file per sample (lane). For reads from SAGE samples, the NlaIII recognition sequence “CATG” was introduced at the 5-end with Linux commands. Scarf files were then run through the open source GAPSS_R pipeline developed in house (www.lgtc.nl/GAPSS). In general, this pipeline takes sequences and has the options to: remove first bases (often of lower quality than other 5′ nucleotides (102)), edits for linkers (present in the sequence reads when sequencing more cycles than the fragment length), aligns to a reference genome with Rmap (40), and reports data as region files (reporting tags in a region, a region defined as a stretch of adjacent nucleotides with aligned reads), and creating UCSC genome browser (103) (http://genome.ucsc.edu/) viewable wiggle tracks.

We ran GAPSS_R with the parameters discussed in the following text. The first base (lower quality) was removed in CAGE samples. CAGE and SAGE samples were edited for 3′ linker sequences (TCGTATGCGTCTTCTGCTTG for CAGE and TCGTATGCGTCTTCTGCTTTGAAAAAAAAAAAAAAA for SAGE), permitting 1 mismatch in the linker (to account for sequencing errors, which occur more towards the 3′ end (102) where linkers were edited from). After linker editing, the majority of CAGE reads were 26 bases in length, whereas SAGE reads were 21 or 22 bases in length (including the “CATG”). Alignment was performed against the mouse repeat masked reference genome build 37 with Rmap v0.41, an alignment tool that reports only unique alignments. Default settings were used during alignment, except to use fasta input and permitting 2 mismatches with CAGE reads and 1 mismatch with SAGE reads. The choice of mismatches permitted is because longer sequences (CAGE) are more likely to contain a sequencing error because the number of errors increases at later sequencing cycles. Region files were created and for CAGE regions we combined adjacent regions, permitting gaps of maximal 100 bases to cluster TSSs and make sure that newly identified TSSs were well separated from annotated TSSs. We kept all data separated by strand since both methods preserve information on the transcribed strand. Wiggle files for visualization in the UCSC genome browser were also separated by strand.

Custom Perl scripts were run on all CAGE and SAGE region files to create reference region files (strand separated) composed of the overlapping regions from all samples. For CAGE region files we again permitted gaps of a maximum by 100 bases. Another custom Perl script was used to link all individual region files to their reference region file, reporting the estimated number of tags in each individual region of the reference region file.

6.3.4 Statistical and Biological Processes Analysis

The statistical language R was then used for analysis of differential expression for CAGE and SAGE data. A threshold of two tags per million aligned reads (average across all samples) was applied to remove transcription events that do not pass the lower limit for consistent detection given our read depth. In addition, for CAGE data, we excluded regions of length 33 or lower. These are likely sample preparation artifacts, since these were usually caused by exactly identical reads of 33 nt, which did
not contain the linker sequence. Even sharply defined TSSs demonstrate variability in start position, resulting in regions that cover >33 nt. Each region was tested separately with a Bayesian algorithm that takes into account library size (150; 36). A Bayesian error rate lower than 0.05 was considered significant. For gene level tests, all tags overlapping a gene (including 1000 bases upstream and downstream of the gene) were summarized before statistical testing. For the calculation of expression ratios between differentiated and proliferating cells, data was first scaled to the average total number of aligned reads. For analysis of reproducibility, data was square root transformed to stabilize variance between samples, after which the Pearson correlation coefficient was calculated.

To compare differentially expressed genes to previously published microarray data we took results from Tomczak et al. 2004 (148), performed VSN normalization (151), and analyzed data from differentiated versus proliferating cells with limma (67; 68) in R. Multiple testing was done according to Benjamini and Hochberg (70). Probes were annotated with NetAffx from the Affymetrix website (www.affymetrix.com) and linked to the CAGE and SAGE top 30 genes based on gene symbols.

To annotate the biological processes we took the top 30 differentially regulated genes from CAGE and SAGE (with a Bayesian Error rate $< 1 \times 10^{-50}$ and sorted for differentiated cells on a ratio of differentiated to proliferating cells), as well as the microarray data (sorted on adjusted p-value), and ran these against 7,689 GO (91; 92) Biological Processes in Anni 2.1 (90).

### 6.3.5 Sequence Annotation

All CAGE and SAGE regions were annotated based on the ElDorado genome annotation (Genomatix, Version 07-2008) for being located in exons, introns, or intergenic regions. Regions that covered an exon and neighboring intron or intergenic region were categorized as partial. In addition a region was categorized as a promoter if it was located in the ElDorado defined promoter region of a transcript. The distance to the nearest TSS (upstream or downstream) was also calculated. CAGE regions were correlated with CAGE data available in ElDorado (originating from the FANTOM3 project (8)).

### 6.3.6 CAGE Region Confirmation

To confirm that our CAGE regions represented newly discovered 5'-ends of transcripts, we designed primers within CAGE regions upstream of 8 genes (Bpag, Cpeb1, Junb, Myl1, Pik3ca, Ppt2, Sertad4x, and Usp34, primers in Additional Table 6.1). RT-PCR experiments were performed using random hexamer priming for cDNA synthesis and qPCR performed on a Roche Lightcycler 480. To provide additional validity to these CAGE regions we inspected multiple UCSC tracks (UCSC genes, Ensembl (3) genes, Vega genes, Other RefSeq, AceView Genes, N-SCAN, and Transcriptome).

To validate that our novel CAGE regions were indicative of myogenic promoters we took all differentially expressed CAGE regions (see Results), expanded or contracted them to a length of 2000 bp, retrieved sequences with Ensembl Perl API scripts, and ran them through CORE_TF (76), a program that identifies over-represented TFBSs. For a background sequence we used 2000 mouse promoters defined as 1000bp before
and 1000bp after the annotated TSS. A Match (55; 51) setting to minimize the sum of false positives and false negatives was used.

We looked into more detail at the upstream CAGE regions of *Myl1*, a myogenic gene that was confirmed to have differential expression in the differentiation analysis. To this we performed standard PCR for a primer set that spans the novel CAGE region into the first UCSC exon (F: TCAGCCAAGTTGA, R: CCTCCAGAAGACCTGTCAGA). We also checked this CAGE region, plus 500 bases upstream sequence, for functional evidence. This was done by taking the mouse sequence, searching for orthologous sequences, and identifying conserved patterns of TFBSs, as has been previously described (152; 153).

### 6.4 Results:

#### 6.4.1 The Biological Model and Experimental Set-up

To study gene expression levels during myogenic differentiation we used C2C12 mouse myoblasts, a common cell model for myogenesis, combined with NGS technology. RNA was isolated from three independent cultures, both of proliferating and differentiated cells. At the latter condition, cells had differentiated into fused and multinucleated myotubes. To confirm successful differentiation, qPCR was performed to determine the expression levels of the genes encoding the late myogenic TF Myogenin and the master myogenic regulator MyoD. Both of these should be expressed at higher levels in differentiated than proliferating cells. qPCR confirmed that cells had started to express Myogenin in differentiated cells and had higher expression of MyoD in differentiated cells (Additional Figure 6.1). CAGE and SAGE libraries were then prepared from all six RNA samples (three independent cell cultures for both proliferating and differentiated cells) and used to determine expression levels based on measurements in the 5′ and 3′ region of the transcripts, respectively. We used both methods to evaluate how well transcript level measurements compare and to improve transcript structure annotation. The latter is essential to facilitate bioinformatic approaches to analyze overall transcription regulation based on shared TFBS promoter profiles.

#### 6.4.2 General Sequencing Data and Alignments

Each CAGE and SAGE library was sequenced on a single lane of the Illumina Genome Analyzer II. To investigate technical reproducibility, two CAGE samples (one from proliferating and one from differentiated cells) were sequenced in duplicate. After running the Illumina Genome Analyzer Pipeline for image and sequence quality analysis, we obtained on average 4.5 and 6.9 million reads from the CAGE and SAGE libraries, respectively (Table 6.1). The scarf files, converted to FASTQ format, containing the reads are available at GEO (154) under the accession number GSE21580. We aligned these reads to the repeat masked mouse reference genome and were able to uniquely map (reporting alignments that are unique to one position in the genome), on average, 1.9 million (42%) and 4.1 million (59%) tags for CAGE and SAGE, respectively (Table 6.1).
| CAGE sample | # reads sequenced | # reads aligned | percent aligned |
|-------------|------------------|----------------|----------------|
| Prolif-1    | 4886341          | 2086233        | 42.7%          |
| Prolif-1 duplo | 3933233      | 1770247        | 45.0%          |
| Prolif-2    | 5003964          | 2421443        | 48.4%          |
| Prolif-3    | 4734605          | 2062081        | 43.6%          |
| Diff-1      | 4525321          | 1679081        | 37.1%          |
| Diff-1 duplo | 3101153        | 1252451        | 40.4%          |
| Diff-2      | 5060041          | 2195263        | 43.4%          |
| Diff-3      | 4830194          | 1578087        | 32.7%          |

| SAGE sample | # reads sequenced | # reads aligned | percent aligned |
|-------------|------------------|----------------|----------------|
| Prolif-1    | 5941753          | 3351426        | 56.4%          |
| Prolif-2    | 7768787          | 4464057        | 57.5%          |
| Prolif-3    | 6723476          | 3878953        | 57.7%          |
| Diff-1      | 9467926          | 5811947        | 61.4%          |
| Diff-2      | 7269002          | 4618715        | 63.5%          |
| Diff-3      | 4392416          | 2494618        | 56.8%          |

Indicators for CAGE and SAGE samples: Prolif for proliferating cells and Diff for differentiating cells, followed by a number representing the biological triplicates. For CAGE there are sequencing duplicates indicated by "duplo." The table contains the number of reads, the number of reads that align uniquely to the repeat masked genome, and the percent aligned.

For visual analysis we constructed UCSC genome browser wiggle files. The wiggle files are available at GEO under accession number GSE21580 and at http://www.lgtc.nl/publications/Hestand_2010_CAGE_SAGE_wig/. To retain information on the direction of transcription, there is one file for each strand. In Figure 6.1 we show an example wiggle track for the Myod1 gene. We clearly see the sharp SAGE peak starting at the most 3’-CATG site followed by 18 additional nucleotides. The CAGE peak at the 5’-end of the transcript is wider, reflecting the variability in the transcription start position. As observed before (8), and observed for many other genes in the current study, CAGE also detects transcription starts in the 3’-region of the gene. This phenomenon is further discussed in the Annotation and Discussion sections. As expected, CAGE and SAGE consistently detect higher expression of Myod1 in differentiated compared to proliferating cells.

We identified 742,355 CAGE regions, consisting of adjacent nucleotides with aligned reads (after concatenating reads permitting gaps of maximally 100 nucleotides to resolve gaps in alignments due to non-unique genomic sequences). 361,655 SAGE regions (not concatenated, since SAGE tags always start at a fixed position) were identified. After applying a threshold of two tags-per-million, a threshold for very low abundant expression (~0.3 copies per cell (36)), 41,862 CAGE and 43,512 SAGE regions remained. The CAGE regions have median lengths of 314 nucleotides, and usually represent clusters of TSSs (plus ~26 nucleotides of downstream sequence). The SAGE tags were 21 or 22 nucleotides long (including the 4 CATG nucleotides representing the NlaIII restriction site).
6.4 Results:

Figure 6.1: CAGE and SAGE wiggle tracks for proliferating (Prolif) and differentiated (Diff) cells in the UCSC Genome Browser for the myogenic marker *Myod1*. We only display reads aligning to the forward strand, the coding direction for *Myod1*. Chromosomal positions are indicated at the top. For each track the Y-axis scale corresponds to the number of tags aligned at that genomic position. Scales use a maximum from each relevant technique in this viewing window (129 for CAGE and 3912 for SAGE). There is 5′ and 3′ concordance for CAGE and SAGE samples, respectively. CAGE provides broader peaks, reflecting TSSs plus ∼26 nucleotides of downstream sequence, whereas SAGE provides discrete peaks. A higher number of tags are in differentiated compared to proliferating samples.

6.4.3 Technical Reproducibility and Biological Overlap

A high correlation was found between the technical CAGE replicates (median Pearson correlation of 0.981) as well as the biological triplicates (median Pearson correlation of 0.963) (Figure 6.2A/B, Additional Table 6.2). As expected, correlation between proliferating and differentiated cells was lower (median Pearson correlation of 0.771) (Figure 6.2C, Additional Table 6.2). Similarly, we observed a high reproducibility for the SAGE experiments (median Pearson correlation of 0.930) between biological triplicates (Figure 6.2D and Additional Table 6.2). Again, the correlation between proliferating and differentiated cells (median Pearson correlation of 0.839) was lower than between cells from the same condition (Figure 6.2E and Additional Table 6.2).
6 CAGE/SAGE: muscle gene structure

Figure 6.2: High reproducibility was found in CAGE regions between sequencing duplicates (A) and biological replicates (B). Panel C shows correlation between CAGE samples from proliferating and differentiated cells. High reproducibility can also be found between SAGE biological replicates (D). Panel E shows the correlation between CAGE samples from proliferating and differentiated cells. The plotted values represent the square root of the number of tags per region.

6.4.4 Annotation of Regions

We annotated the 41,862 CAGE regions using Eldorado’s mouse genome annotation: 9,957 regions map to an annotated exon, 27,190 partially overlap an exon and intron/intergenic region, 2,368 map to an intron, and 2,347 regions are purely intergenic. The median number of tags in the exonic and partial regions (63 tags and 90 tags respectively) were higher than in the intronic and intergenic regions (45 tags and 54 tags, respectively). These data clearly show that our CAGE experiments identifies many (lower abundant) TSSs / transcribed regions that have not yet been identified and/or annotated as such in current genome databases.

Based on ElDorado annotation of our 41,862 CAGE regions, 13,541 of the CAGE regions (32%) contained an annotated TSS, 6,331 CAGE regions (15%) were annotated as promoters (i.e. a genomic region surrounding a TSS containing functional elements like TFBSs that are responsible for the regulation of the expression of the transcript), and 8,028 (19%) CAGE regions contained an annotated transcript 3′-end. 3′-end alignments are consistent with the previously observed (8) significant amount of (shorter) transcripts originating from the 3′-ends of genes. We compared our CAGE results to previous CAGE studies (FANTOM3) contained in ElDorado and identified 31,680 regions (76%) overlapping with at least one on the FANTOM3 CAGE tags. Only 6,119 (15%) and 5,635 (13%) of these regions were observed in FANTOM3 muscle and heart CAGE libraries, respectively. This is explained by the
small size of these muscle and heart libraries (8), together representing only 1% of all available CAGE tags in FANTOM3.

6.4.5 Comparison of CAGE, SAGE, and Microarray Expression Data

To compare overall expression level measurements we assigned CAGE and SAGE regions to genes (including 1000 bases upstream and downstream of the gene). Expression above a threshold of 2 transcripts per million (~0.3 copies per cell) (155) was observed for 10,409 and 10,987 genes respectively. Expression profiles for both methods showed a high correlation (Figure 6.3A-C), with 9,240 genes being expressed in both methods above 2 transcripts per million (Figure 6.3D). Additional Figure 6.2 shows that the relative overlap is even bigger when higher detection thresholds are applied, obviously at the expense of many more genes not reaching the detection threshold. 4,304 genes were differentially expressed between proliferating and differentiated cells (Bayesian error rate < 0.05) according to the CAGE data and 3,846 according to the SAGE data with 2,144 genes present in both lists of significant genes (Figure 6.3E). Most others were just borderline significant according to one of both methods.

We compared the top 30 most differentially expressed genes for both methods (Table 6.2A) to results from a similar microarray dataset on myogenic differentiation in the same cell line (148). In general, the genes identified by CAGE and SAGE also demonstrated very significant changes on the microarrays. However, in the top 30, 13 genes identified by CAGE and 10 identified by SAGE were not represented on the array, demonstrating the comprehensive nature of the CAGE and SAGE-
based gene expression profiling techniques. The biological processes controlled by the top 30 CAGE, SAGE, and microarray genes, were annotated with the Anni2.1 text-mining tool (Table 6.2B). All CAGE and SAGE-derived GO terms can readily be related to muscle development, whereas 3/10 GO terms associated with the microarray-derived gene list can not ("cyclin-dependent protein kinase inhibitor activity," "6-phosphofructokinase," and "tumor suppressor activity").

6.4.6 Differential TSS Use and Validation

In our CAGE data, we identified 111 regions upstream of the start of a known gene and 85 CAGE regions downstream of an annotated gene containing significantly different numbers of tags in proliferating and differentiated cells (Additional Table 6.3). The differential expression of transcripts originating from 7 out of 8 of these regions (upstream from genes \textit{Bpag}, \textit{Cpeb1}, \textit{Junb}, \textit{Myl1}, \textit{Pik3ca}, \textit{Ppt2}, \textit{Sertad4x}, and \textit{Usp34}) were confirmed by RT-PCR/qPCR (Figure 6.4B and Additional Figure 6.3). To evaluate if these novel exons were contained in a transcript of the gene of interest we inspected the following tracks in the UCSC genome browser: UCSC genes, Ensembl genes, Vega genes, Other RefSeq, AceView Genes, N-SCAN, and Transcriptome (Figure 6.4A and Additional Figure 6.4). In all but \textit{Junb} we found the CAGE regions overlapping at least one exon from an additional track connected to the gene of interest (Figure 6.4A and Additional Figure 6.4). This indicates that these CAGE regions usually represent alternative transcripts that are not yet properly annotated in all resources, including the mainstream UCSC and Ensembl annotations. This suggests that the mainstream genome annotation are far from complete and that additional evidence, including our CAGE data, is required to more precisely define transcript structure.

To support that differential transcription in the 196 CAGE regions is regulated by myogenic TFs, we searched for over-represented TFBSs and found the binding sites for the master regulators MyoD (p-value $6.49 \times 10^{-03}$ from CORE_TF’s binomial test) and Myogenin (p-value: $3.87 \times 10^{-02}$ ) and the Ebox motif (p-value $6.02 \times 10^{-03}$) (frequently found in muscle promoters (156; 157)) to be significantly over-represented in 2,000 bp of sequence composed of the CAGE and surrounding regions (Additional Table 6.4).

For one of these novel CAGE regions, \textit{Myl1}, we confirmed by standard RT-PCR that there is a transcript extending from the novel CAGE region into the UCSC defined exon 1 (Figure 6.4C). The CAGE sequencing, RT-PCR/qPCR within the region, and the standard PCR into exon 1 all confirmed that this transcript is only present in differentiated cells, explaining why it is missing in standard genome annotations. For functional evidence that this region is used as a promoter, we also looked for conserved TFBSs in and upstream of this region. Within the Genomatix Suite we identified orthologous sequence regions from human and horse corresponding to the CAGE region and 5' upstream (promoter) sequence. In this area we identified conserved TFBSs for NKX, GATA, and SRF (Figure 6.4D), all of which are known to be involved in the regulation of muscle genes (158). This makes it likely that the region directly upstream of the novel exon 1 is used as an alternative promoter.
### 6.4 Results:

**Table 6.2: Differential Gene Expression**

| A. CAGE Gene | Ratio | Microarray p-val | SAGE Gene | Ratio | Microarray p-val |
|--------------|-------|------------------|-----------|-------|------------------|
| Hfe2         | 4073  | NA               | RP23-36P22.5 | 576   | NA               |
| Myom3        | 1624  | NA               | Neb       | 525   | NA               |
| Lmod2        | 1305  | NA               | Mylpf     | 504.1 | \(1.7 \times 10^{-15}\) |
| Myh7         | 1124  | \(5.98 \times 10^{-3}\) | Ttn       | 380.1 | NA               |
| Mb           | 908   | \(1.07 \times 10^{-14}\) | Myh3      | 368.1 | \(2.4 \times 10^{-14}\) |
| RP23-36P22.5 | 735   | NA               | Xirp1     | 306.1 | \(2.24 \times 10^{-13}\) |
| Pygm         | 717   | \(4.82 \times 10^{-17}\) | 1110002H13Rik263 | NA   |                  |
| Myl4         | 614   | \(8.86 \times 10^{-20}\) | Tnmc1     | 232.1 | \(1.24 \times 10^{-11}\) |
| Synpo2l      | 595   | NA               | Cav3      | 150.1 | \(3.58 \times 10^{-22}\) |
| Myh1         | 561   | \(3.64 \times 10^{-15}\) | Chfa2t3   | 133.1 | \(2.89 \times 10^{-10}\) |
| Tnnl1        | 529   | \(2.24 \times 10^{-9}\) | Chrng     | 115.1 | \(4.63 \times 10^{-9}\) |
| Tnnl2        | 442   | \(3.20 \times 10^{-11}\) | Myom2     | 105.1 | \(6.66 \times 10^{-16}\) |
| Mpa2l        | 410   | NA               | Tnmt1     | 100.1 | \(1.15 \times 10^{-10}\) |
| Ctrlb1       | 406   | \(7.55 \times 10^{-7}\) | Ryr1      | 92.1  | \(7.03 \times 10^{-14}\) |
| Ttn          | 402   | NA               | Apobec2   | 84.1  | \(2.95 \times 10^{-15}\) |
| Neb          | 374   | NA               | Cox6a2    | 72.1  | \(2.45 \times 10^{-16}\) |
| Kcnq4        | 365   | NA               | Dio2      | 64.1  | \(2.14 \times 10^{-10}\) |
| Mylpf        | 341   | \(1.70 \times 10^{-15}\) | C1qtnf5   | 52.1  | \(4.36 \times 10^{-5}\) |
| 1110002H13Rik | 341   | NA               | Htr2b     | 43.1  | \(3.76 \times 10^{-6}\) |
| Inpnp4       | 328   | NA               | Sgcg      | 42.1  | \(1.15 \times 10^{-12}\) |
| Xirp1        | 307   | \(2.24 \times 10^{-13}\) | Fndc5     | 39.1  | NA               |
| Atp2a1       | 304   | \(2.06 \times 10^{-14}\) | Jsrp1     | 36.1  | NA               |
| Casq2        | 297   | \(4.74 \times 10^{-6}\) | Ankrd23   | 36.1  | NA               |
| Cacn1s       | 296   | \(5.20 \times 10^{-19}\) | AK031267  | 29.1  | NA               |
| Ces2         | 245   | NA               | Sema6a    | 26.1  | \(3.08 \times 10^{-3}\) |
| Cox6a2       | 241   | \(2.45 \times 10^{-16}\) | Lgr5      | 23.1  | \(9.33 \times 10^{-1}\) |
| Myo2         | 238   | \(2.36 \times 10^{-6}\) | Pdlim3    | 22.1  | \(3.18 \times 10^{-6}\) |
| Myh3         | 234   | \(2.40 \times 10^{-14}\) | Klhl31    | 22.1  | NA               |
| Tmem182      | 216   | NA               | ORF63     | 21.1  | NA               |
| Tncl1        | 215   | \(1.24 \times 10^{-11}\) | Gfra2     | 19.1  | \(2.98 \times 10^{-2}\) |

B. CAGE GO SAGE GO microarray GO

1) regulation of striated muscle contraction
2) cardiac muscle contraction
3) Myogenesis
4) regulation of myoblast differentiation
5) skeletal muscle development
6) Muscle Development
7) striated muscle contraction
8) myoblast differentiation
9) muscle cell differentiation
10) sarcomere organization

Top 30 genes from SAGE and CAGE expression data (A). All genes with a Bayesian error rate < \(1 \times 10^{-50}\) were sorted on the ratio (normalized tags from differentiated / proliferating cells) and the highest ratios for differentiated cells displayed. The microarray p-values are adjusted p-values for differential gene expression from a similar experiment (proliferating and differentiated C2C12 cells (148)). NA = no probe annotation for the gene. The top 10 GO biological processes (B) associated with the top 30 genes for CAGE, SAGE, and microarray experiments indicate clear muscle relations, with the exception of 3 (in italics) processes in the microarray data.
Figure 6.4: The UCSC display of (A) UCSC/Ensembl defined first exon and an upstream CAGE region for Myl1 (reverse strand reads only, on which the gene lies) for samples Prolif-1 and Diff-1. The Y-axis indicates the number of tags aligned at each position in the genome. We also display additional track information (UCSC genes, Ensembl genes, Vega genes, Other RefSeq, AceView Genes, N-SCAN, and Transcriptome), several of which confirm the presence of the upstream CAGE region. (B) qPCR with primers within the CAGE region for Prolif, Prolif-C (reverse transcriptase control), Diff, and Diff-C (reverse transcriptase control). The qPCR results are plotted as threshold cycle (Cp) values (lower = higher expression), with bars indicating a range of one standard deviation between technical duplicates. (C) standard PCR on agarose gel with forward primer in the novel CAGE region and reverse primer in the conventional exon 1. Comparison with the genomic control verifies the presence of an intron of 200 bases. A 100 bp ladder is included. Panels A-C are all consistent with higher expression in differentiated than proliferating cells. (D) Cross-species conserved muscle specific TFBSs around and upstream of the Myl1 CAGE region support its role as a promoter for this region.

6.5 Discussion

Using CAGE and SAGE methods with NGS we have measured gene expression levels during myogenic differentiation and identified muscle specific TSSs. By elucidating promoter regions and regulation in these myogenic cells we hope to better understand the process of muscle development and regeneration, providing clues to cure muscle related illnesses. Since biologists and clinicians often study (first) exons and 5′ promoter regions it is crucial to know the positions of TSSs in the genome. Our data will help them identify potentially pathogenic mutations in transcripts and promoters used during myogenic differentiation, which might have been over looked with current genome annotations. On a technical level, this is the first time CAGE and SAGE have been evaluated using the same RNA samples.

We found both the technically demanding CAGE method and the slightly less
Discussion

laborious SAGE method to be extremely robust. Biological triplicates with independent sample preparations and sequencing runs were found to have high correlations (Figure 6.2, Additional Table 6.2). This is in line with previous findings in the FANTOM4 CAGE study (159) and our previous (36) finding with SAGE. Higher technical reproducibility also enhances the ability to verify low expressed genes, which was an obstacle in microarray analysis. The high quality of the data implies that more investments should be made in biological than technical replicates, as demonstrated for CAGE for the first time in the current paper.

This study also highlights other advantages over microarrays. For a third of the top 30 genes (13/31 CAGE genes and 10/30 SAGE genes, Table 6.2A) there was no probe on the microarray. Finding many more significant genes not interrogated by the microarrays stresses the more comprehensive transcript profiling by NGS based methods. We also found more muscle related biological processes associated with the top 30 CAGE and SAGE genes compared to the microarray top 30 genes (Table 6.2B) indicating the higher relevance of the top hits for the process under study.

The data provided by these methods have greatly expanded our knowledge of muscle specific transcription. 56% of the analyzed CAGE regions contained an annotated TSS, indicating discovery of many novel TSSs. 76% of CAGE regions matched known FANTOM3 CAGE tags, but less than 20% of those matched known muscle related CAGE tags. This is likely due to the lower sequencing depth in the previous FANTOM3 CAGE studies. High overlap with the previous FANTOM CAGE regions indicates these reflect true TSSs, but there is a lack of information on the definition of TSS usage in relation to tissue. To exemplify this point, we identified 196 intergenic regions significantly different between proliferating and differentiated cells, indicating muscle-specific alternative promoter and first exon usage. Several of these were verified by PCR and additional UCSC track evidence. We also identified over-represented muscle specific TFBSs in the 196 CAGE regions and additional conserved muscle specific TFBSs upstream of a novel first exon of the Myl1 gene, coding for one of the light chains of the myosin protein complex involved in muscle contraction. These muscle specific TFBSs indicate that the identified regions potentially serve as promoters.

This is the first study to compare NGS of CAGE and SAGE libraries from the same RNA samples. Gene expression measurements by CAGE and SAGE are generally consistent. The high correlation between methods (Figures 6.3A-C), large overlap between genes detected (Figure 6.3D) and differential gene lists (Figure 6.3E), and gene involvement in similar biological pathways (Table 6.2B) indicates these methods are interchangeable for expression analysis. Only when transcript structure (5′ or 3′) is important is one method preferential over another. Correct 5′ usage is crucial for promoter based regulation studies, whereas proper 3′ usage is needed for studies concerning micro-RNA regulation.

Of the 4,304 and 3,846 genes differentially expressed between proliferating and differentiated cells with CAGE and SAGE, respectively, over half (2,144) of the genes are identical. More changes in CAGE than SAGE levels could indicate alternative promoter usage is more common than that of alternative 3′ ends. The detection of genes by one technique, but not the other, is mostly inherent to the use of thresholds. In addition, a minority of transcripts may be missed entirely by one of the methods due to the absence of a CATG site in the transcript (SAGE) or the sequence around...
the TSS not being unique in the genome (CAGE). Rmap does not report a read when it aligns with equal mismatches to multiple regions in the genome. Therefore non-unique TSS sequences will not be reported and included in our analysis. For both techniques, we frequently detected multiple regions in the same gene. 75% of the genes had multiple SAGE tags with abundance above the threshold of 2 transcripts per million. In our previous paper (36), we discussed that this is probably not a technical artifact but most likely due to different 3′ ends and usage of multiple polyadenylation sites.

Similar to previous studies (8), we found a large number of CAGE tags aligning to the 3′ end of known transcripts. This phenomenon has been previously validated by the RACE method and explained as potential 3′ derived regulatory noncoding RNAs (8). With additional analysis this could serve as a method for identification of noncoding RNAs. In addition, these should be recognized as a source of false expression levels identified by the 3′ based SAGE method and microarrays based on 3′ probes.

67% of the genes contained multiple CAGE regions. This phenomenon was previously referred to as "exon painting" (160). Examples of genes where nearly all exons are covered by CAGE tags are Colla1 and Colla2 (Additional Figure 6.5A/B, respectively). This is unexpected since the RNA integrity was high in all samples, the CAGE technique only captures capped transcripts, and even when some non-capped transcripts may be included, the method will only create tags from the ultimate 5′ end. Together with the observation of genes with a highly abundant peak at the 5′ end without any exon painting (Additional Figure 6.5C/D) and the fact that the exon painting patterns are highly reproducible in independent CAGE sample preparations, this suggests that there is a biological explanation for the exon painting phenomenon. The observation of exon painting is consistent with the finding of many short transcripts from exonic regions in a tiling array study (160). It is not clear whether these short transcripts are degradation products from larger transcripts, true de novo transcriptional events, or a combination of both. From our study, it is highly likely that many of these shorter transcripts contain a cap structure. The process of recapping of transcript fragments has been documented before (160). Fejes-Toth et al. propose long RNAs are spliced into mature and translatable RNAs, but that these mature RNAs can also be further processed (160). This further processing involves cleavage into smaller RNA fragments and possible modification by additional 5′ capping (160). The presence of exon painting complicates the identification of novel TSSs and is the reason why we focused on the discovery of novel TSSs in intergenic regions and did not report alternative TSSs within annotated genes. A positive consequence of the exon painting phenomenon is that the CAGE technique gives additional information on the exon structure of many genes.

The large data yield and reproducibility should serve as an example of the advantages of applying NGS to CAGE and SAGE techniques. These methodologies should be expanded to other tissues and processes in the future to enrich our knowledge of the genome of many organisms. This work has provided a substantial increase in our knowledge of myogenic TSSs and expression. This has also demonstrated the technical advantages of CAGE and SAGE in conjunction with NGS.
6.6 Acknowledgments

We wish to thank Michel Villerius, Michiel van Galen, and Ivo Fokkema for their computational assistance. We also wish to thank Rolf Vossen for advice with the Lightcycler 480. We would also like to thank Henk P.J. Buermans and Emile J. de Meijer for wet-lab assistance and providing primers for expression analysis.

6.7 Funding

This project was supported by grants from the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) and the Center for Biomedical Genetics (in the Netherlands).

6.8 Conflict of interest

Andreas Klingenhoff, Matthias Scherf, Wilbert van Workum, Makoto Suzuki, Thomas Werner, and Matthias Harbers declare that they have competing financial interests.
6.9 Additional Files

Additional Figure 6.1

Myogenic confirmation of RNA. Expression levels are relative to the control genes \textit{GAPDH} or \textit{HPRT}.
Overlap between genes detected by SAGE and CAGE. We compared the genes detected by SAGE at different threshold values (expression levels in transcripts per million, x-axis, 10log scale) with the genes detected in CAGE with a fixed threshold of 2 transcripts per million. The number of genes also detected by CAGE is expressed as a percentage of the total number of genes detected by SAGE at different thresholds is plotted in open symbols. The closed symbols represent the percentage of genes remaining after thresholding of the SAGE data.

PCR validation of CAGE tags in novel upstream exons. -B in the legend indicates a no enzyme reverse transcriptase control. The Y-axis is the threshold cycle (Cp) value (lower = higher expression), with plotted bars indicating +/- one standard deviation.
Additional Figure 6.4

UCSC/Ensembl defined first exon and an upstream CAGE region for 8 example genes (only reads on the strand which the genes lie on are displayed) for samples proliferating-1 and differentiated-1. The Y-axis indicates the number of tags aligned at each position in the genome. We also display additional track information (UCSC genes, Ensembl genes, Vega genes, Other RefSeq, AceView Genes, N-SCAN, and Transcriptome), several of which confirm the presence of the upstream CAGE regions. A larger color figure is available upon request.
Additional Figure 6.5

Examples of exon painting conservation in 8 different CAGE datasets divided over each strand (F or R) across the genes Col1a1 (A) and Col1a2 (B). The same datasets are also plotted against the genes Fgf11 (C) and Got2 (D), which do not show exon painting. To ensure regions with a low number of tags are visible and not over-shadowed by regions with a high number of tags the maximum display is set to 25 tags per position.
### Additional Table 6.1

| Expression | Primer Name | Sequence                                       |
|------------|-------------|------------------------------------------------|
| MyoD       | ex3-F       | CCCAATGCGATTTATCAGGT                           |
| MyoD       | ex3-R       | TCTGCTCTTCCTCCCTCCCTCT                       |
| MyoD       | ex1-F       | GACAGGGAGGAGGAGGAGGAGGAGG                      |
| MyoD       | ex1-R       | AAGTCTATAGTCCCGGAGTGG                        |
| MyoG-R     |             | TGGGAGTTGCAATCACAAGTGG                        |
| MyoG-F     |             | CCTTGCTCAGCTCCCTCA                            |
| HPRT-F     |             | TCCCTGGTTAAGCAGTACCAGCC                      |
| HPRT-R     |             | CGAGAGGTCTCTTTTCACCAGC                       |
| GAPDH-F    |             | TCCATGACAACTTTTGCGATG                        |
| GAPDH-R    |             | TCACGCCACAGCTTTCCA                            |

| CAGE region | Primer Name | Sequence                                       |
|-------------|-------------|------------------------------------------------|
| Myl1.CF     |             | TCAGCCAAAATTCCAAGTTGA                          |
| Myl1.CR     |             | CCACTTCCTAAGAAGCTTTACCG                       |
| Usp34.CF    |             | CGGACGGAAGGAGGAAGAAG                          |
| Usp34.CR    |             | GCCTCCTCCTCCGACACAC                          |
| Ppt2.CF     |             | CACTGGCAGGGTTTGTGTC                           |
| Ppt2.CR     |             | GACAAACTGCTCTCACAGATCC                       |
| Bpag.CF     |             | GTGCTGAGTCATGCGAGAC                         |
| Bpag.CR     |             | CCGGAACGACTGATGGAG                          |
| Pik3.CF     |             | GTGGGGAAGAGGTTTCGTTGGTT                      |
| Pik3.CR     |             | GTTCTGCTTTTTCGCTACAT                        |
| Cpeb.CF     |             | GTCTGTCCAGCCCTAGC                            |
| Cpeb.CR     |             | GAAGCTGTTGTTCGCCGAGAG                       |
| Sert.CF     |             | GCTCAGTCACGCTACTGCAC                        |
| Sert.CR     |             | CCTCCCTCTGTACAGCACAC                        |
| Junb.CF     |             | GGAAGAGGACTTAAGGTTCA                        |
| Junb.CR     |             | GTAGGGGCATTGGAGAAGAAG                       |

PCR primers used in study.
### Additional Table 6.2

|          | Prolif-1<sup>d</sup> | Prolif-1 | Prolif-2 | Prolif-3 | Diff-1<sup>d</sup> | Diff-1 | Diff-2 | Diff-3 |
|----------|----------------------|----------|----------|----------|--------------------|--------|--------|--------|
| CAGE     | 1.000                | 0.983    | 0.960    | 0.959    | 0.754              | 0.754  | 0.787  | 0.770  |
| Prolif-1 | 0.983                | 1.000    | 0.961    | 0.959    | 0.751              | 0.755  | 0.790  | 0.771  |
| Prolif-2 | 0.960                | 0.961    | 1.000    | 0.982    | 0.771              | 0.774  | 0.810  | 0.783  |
| Prolif-3 | 0.959                | 0.959    | 0.982    | 1.000    | 0.765              | 0.767  | 0.804  | 0.777  |
| Diff-1<sup>d</sup> | 0.754            | 0.751    | 0.771    | 0.765    | 1.000              | 0.978  | 0.963  | 0.962  |
| Diff-1   | 0.754                | 0.755    | 0.774    | 0.767    | 0.978              | 1.000  | 0.970  | 0.967  |
| Diff-2   | 0.787                | 0.790    | 0.810    | 0.804    | 0.963              | 0.970  | 1.000  | 0.966  |
| Diff-3   | 0.770                | 0.771    | 0.783    | 0.777    | 0.962              | 0.967  | 0.966  | 1.000  |

|          | Prolif-1 | Prolif-2 | Prolif-3 | Diff-1 | Diff-2 | Diff-3 |
|----------|----------|----------|----------|--------|--------|--------|
| SAGE     | 1.000    | 0.968    | 0.920    | 0.806  | 0.879  | 0.839  |
| Prolif-1 | 0.968    | 1.000    | 0.940    | 0.802  | 0.885  | 0.851  |
| Prolif-2 | 0.920    | 0.940    | 1.000    | 0.721  | 0.852  | 0.838  |
| Prolif-3 | 0.806    | 0.802    | 0.721    | 1.000  | 0.887  | 0.824  |
| Diff-1   | 0.879    | 0.885    | 0.852    | 0.887  | 1.000  | 0.941  |
| Diff-2   | 0.839    | 0.851    | 0.838    | 0.824  | 0.941  | 1.000  |

Reproducibility of CAGE and SAGE methods: Pearson correlations. <sup>d</sup> = sequencing duplicate.

### Additional Table 6.3

*available upon request*

Regions upstream and downstream of a known genes that were significantly different between proliferating and differentiated cells. Distance is distance from the CAGE region to the TSS. Bay.error is the Bayesian error rate.
Additional Table 6.4

| TFBS                          | p-value            |
|-------------------------------|--------------------|
| V$SREBP1_Q6                  | $6.9010 \times 10^{-04}$ |
| V$BACH2_01                   | $3.6910 \times 10^{-03}$ |
| V$CREB_01                    | $5.7810 \times 10^{-03}$ |
| V$EBOX_Q6.01                 | $6.0210 \times 10^{-03}$ |
| V$BACH1_01                   | $6.2110 \times 10^{-03}$ |
| V$MYOD_01                    | $6.4910 \times 10^{-03}$ |
| V$PTF1BETA_Q6                | $8.0810 \times 10^{-03}$ |
| V$PR_01                      | $1.0110 \times 10^{-02}$ |
| V$NRSE.B                     | $1.0610 \times 10^{-02}$ |
| V$TFE.Q6                     | $1.1210 \times 10^{-02}$ |
| V$USF.C                      | $1.1710 \times 10^{-02}$ |
| V$E12_Q6                     | $1.2110 \times 10^{-02}$ |
| V$SZF11_01                   | $1.4010 \times 10^{-02}$ |
| V$AML1.01                    | $1.4510 \times 10^{-02}$ |
| V$CACBINDINGPROTEIN.Q6       | $1.4610 \times 10^{-02}$ |
| V$SREBP.Q3                   | $1.6310 \times 10^{-02}$ |
| V$TGIF.01                    | $1.6910 \times 10^{-02}$ |
| V$FXR_IR1.Q6                 | $1.7110 \times 10^{-02}$ |
| V$AP4.01                     | $1.7210 \times 10^{-02}$ |
| V$PADS.C                     | $1.8410 \times 10^{-02}$ |
| V$STRA13.01                  | $2.1910 \times 10^{-02}$ |
| V$AREB6.03                   | $2.9110 \times 10^{-02}$ |
| V$RREB1.01                   | $3.0710 \times 10^{-02}$ |
| V$NFY.C                      | $3.3510 \times 10^{-02}$ |
| V$ETS1_B                     | $3.3910 \times 10^{-02}$ |
| V$HTF.01                     | $3.6310 \times 10^{-02}$ |
| V$NANOG.01                   | $3.7510 \times 10^{-02}$ |
| V$MYOGENIN_Q6                | $3.8710 \times 10^{-02}$ |
| V$ZTA.Q2                     | $4.5110 \times 10^{-02}$ |
| V$COUP_DR1.Q6                | $4.6510 \times 10^{-02}$ |

Top 30 (sorted on decreasing p-value significance) CORE_TF over-represented TF-BSs (represented as a TRANSFAC position weight matrix) in the 196 differentially expressed intergenic CAGE regions.