Cloning and Characterization of Buffalo *NANOG* Gene: Alternative Transcription Start Sites, Splicing, and Polyadenylation in Embryonic Stem Cell-Like Cells

Natwar Singh,^{1,2} Ruchi Sharma,² Aman George,² Suresh K. Singla,² Prabhat Palta,²
Radhaysham Manik,² Manmohan S. Chauhan,² and Dheer Singh¹

NANOG is a critical homeodomain transcription factor responsible for maintaining embryonic stem cell (ESC) self-renewal and pluripotency. In the present study, we isolated, sequenced, and characterized the *NANOG* gene in buffalo ESC-like cells. Here, we demonstrated that *NANOG* mRNA is expressed as multiple isoforms and uses four alternative transcriptional start sites (TSSs) and five different polyadenylation sites. The TSSs identified by 5'-RNA ligase-mediated rapid amplification of cDNA ends (RLM-5'-RACE) were positioned at 182, 95, 35, and 17 nucleotides upstream relative to the translation initiation codon. 3'-RACE experiment revealed the presence of tandem polyadenylation signals, which leads to the expression of at least five different 3'-untranslated regions (269, 314, 560, 566, and 829 nucleotides). Expression analysis showed that these alternatively polyadenylated transcripts expressed differentially. Sequence analysis showed that the open reading frame of buffalo *NANOG* codes for a 300-amino-acid-long protein. Further, results showed that alternative splicing leads to the expression of two types of transcript variants encoded by four and five exons. *In silico* analysis of cloned 5'-flanking region (3366 nucleotides upstream of translation start codon) identified several putative transcription factors binding sites in addition to a TATA box and CAAT box at -30 and -139 bp (upstream to the distal most TSS), respectively, in the buffalo *NANOG* promoter.

Introduction

NANOG IS A HOMEODOMAIN containing transcription factor belonging to the homeobox family of proteins. *NANOG* expression is restricted to pluripotent lineages in the embryo, including blastocyst inner cell mass (ICM), epiblast before gastrulation, and primordial germ cells, and has been shown to regulate pluripotency in embryonic stem cells (ESCs) (Chambers *et al.*, 2003; Mitsui *et al.*, 2003). *NANOG*, in combination with OCT4 and SOX2, serves as both an activator and repressor of multiple target genes responsible for self-renewal and differentiation (Pan and Thomson, 2007). *NANOG* expression is rapidly downregulated in ESCs during differentiation. Even in the presence of leukaemia inhibitory factor (LIF), *NANOG* downregulation occurs at the outer layer of ESC colonies' aggregates together with primitive endoderm formation (Hamazaki *et al.*, 2004), which can be rescued by *NANOG* overexpression even in the absence of LIF (Mitsui *et al.*, 2003; Hamazaki *et al.*, 2004).

Transcriptional initiation (selection of transcriptional start sites [TSSs]) and mRNA polyadenylation are the integral parts of gene expression and crucial steps in gene regulation.

The core promoter is the minimal region of DNA required for RNA polymerase II (Pol II) to assemble with the general transcription factors and form the preinitiation complex for initiation of activator-independent (basal) transcription (Gross and Oelgeschlager, 2006). At the center of the core promoter is the initiator (INR) sequence that contains the TSS, which is defined as the most 5'-nucleotide of mRNA transcribed by Pol II (Gross and Oelgeschlager, 2006; Sandelin *et al.*, 2007). During the maturation of most eukaryotic mRNA, a polyadenine [poly(A)] tail is added to the cleaved 3'-end of a precursor mRNA (pre-mRNA) post-transcriptionally. Such a modification of mRNA has been shown to affect its stability, translation competence, and nuclear-to-cytoplasmic export (Danckwardt *et al.*, 2008). The post-transcriptional processing of mRNA is an event that has also been found tightly coupled with splicing and transcription termination (Proudfoot *et al.*, 2002; Proudfoot, 2004). Thus, it is an essential and critical processing event as an integral part of gene expression. The polyadenylation process requires two major components: the *cis*-elements or poly(A) signals of the pre-mRNA, and the *trans*-acting factors that carry out the cleavage and addition of the

¹Molecular Endocrinology Laboratory, Animal Biochemistry Division and ²Embryo Biotechnology Laboratory, Animal Biotechnology Centre, National Dairy Research Institute (NDRI), Karnal, Haryana, India.

poly(A) tail at the 3'-end. These *trans*-acting factors are a complex of about 25–30 proteins involved in signal recognition, cleavage, and polyadenylation (Proudfoot, 2004).

Despite the advances in the knowledge of the functional role of pluripotency-maintaining factors, very little is known about the transcriptional regulation of *NANOG* in ESCs. In this study, we sought to characterize and isolate the *NANOG* gene from ESC-like cells of buffalo, a domestic farm animal. To understand the mechanisms involved in the transcriptional regulation of *NANOG*, we first characterized the TSSs in buffalo ESC-like cells. Correct identification of the TSS in buffalo ESC-like cells will lead to the location of the *NANOG* core promoter, including core and *cis*-acting elements, and provide insights into the molecular mechanisms involved in expression. In addition, *NANOG* mRNA transcripts resultant of alternative polyadenylation has also been elucidated.

Material and Methods

All the chemicals and media were purchased from Sigma Chemical Co. and the disposable plastic wares were from Nunc unless otherwise indicated.

ESC derivation, maintenance, and characterization

Buffalo ESC-like cells were isolated and characterized as described earlier (Verma *et al.*, 2007; George *et al.*, 2011). Briefly, blastocyst-derived ESC-like cells were cultured onto mitomycin C ($10 \mu\text{g mL}^{-1}$)-treated buffalo fetal fibroblast feeder layers in ESC medium, which comprised of Knockout DMEM (Invitrogen Corporation) + 15% Knockout serum replacer (Invitrogen) + 2 mM L-glutamine + 1% MEM nonessential amino acids + 1000 U mL^{-1} LIF + 5 ng mL^{-1} bFGF-2 + 50 $\mu\text{g mL}^{-1}$ gentamicin sulfate. Colonies were mechanically passaged using a microblade, every 4–5 days, and the medium was changed every 24 h. The cell line is being maintained in our laboratory for more than 1 year (>100 passages). Expression of intracellular and surface markers, specific for ESCs, was studied using either reverse transcription–polymerase chain reaction (RT-PCR) and/or immunofluorescence (IF) staining.

RNA isolation and RT-PCR

Total RNA was prepared from cultured buffalo ESC-like cells. Isolation of total RNA was performed with TRI reagent (Ambion) according to the manufacturer's instruction. The cDNA was synthesized using SuperScript III First-Strand Synthesis System (Invitrogen). RT-PCR was performed and amplified PCR products were cloned in pGEM-T Easy vector (Promega) and verified by DNA sequencing.

Isolation and cloning of open reading frame

The cDNA prepared using ESC-like cells was used for PCR amplification of complete open reading frame (ORF) of *NANOG*. Primers used for amplification were 5' TCACA CCCGGAGATCTTCACCTT 3' (sense primer) and 5' TTGT ACTTTTGCCCCCTGTGCTT 3' (antisense primer) based on bovine (GenBank accession No. NM_001025344.1). The PCR products were cloned into the pGEM-T vector and multiple clones were sequenced. Genomic fragments of the coding region were amplified and sequenced to verify the exon–intron boundaries of *NANOG* in buffalo.

The 5'- and 3'-rapid amplification of cDNA ends

The 5'- and 3'-RNA ligase-mediated rapid amplification of cDNA ends (5'- and 3'-RLM-RACE) was performed with FirstChoice® RLM-RACE Kit (Ambion) according to the manufacturer's protocol. Briefly, 1 μg of total RNA from buffalo ESC-like cells was treated with calf intestinal phosphatase to remove the 5'-phosphates from any degraded or noncapped mRNA, followed by treatment with tobacco acid pyrophosphatase (TAP) to remove the 5'-cap structure from full-length mRNA, leaving a 5'-monophosphate. A 5'-RACE RNA adapter oligonucleotide was ligated to the TAP-treated mRNA using T4 RNA ligase. After adapter ligation, mRNA was reverse-transcribed using moloney murine Leukemia virus (M-MLV) reverse transcriptase and random decamers. The regions corresponding to the legitimate 5'-ends of the capped mRNA species were amplified by two consecutive PCR amplifications. The first round of PCR was performed using a sense FirstChoice 5'-RLM RACE outer primer (5' GCTGATGGCGATGAAT GAACACTG 3') and an antisense *NANOG*-specific primer (5' GGGACCGTCTCTTCCTTCTC 3'). A nested PCR amplification was carried out using 2 μL outer PCR products as a template DNA with a sense FirstChoice 5'-RLM RACE inner primer (5' CGCGGATCCGAACACTGCGTTTGCTGGCT TTGATG 3') and a nested antisense *NANOG*-specific primer (5'-GGA GGA GGG AAG AGG AGA GA-3').

To obtain 3'-ends of *NANOG*, 2 μg total RNA was reverse-transcribed using 3'-RACE adapter primers. After cDNA synthesis, the 3'-end of RNA was obtained by two rounds of PCR amplification. The first round of PCR was performed using an antisense FirstChoice 3'-RACE outer primer (5' GCGAGCACAGAATTAATACGACT 3') and a *NANOG* gene-specific outer sense primer (5'GTTTTGAGGCTTTG CAGCTC 3'). A nested PCR amplification was performed using 1 μL PCR product of first-round amplification, using FirstChoice 3'-RACE inner antisense primer (5' CGCGGA TCCGAATTAATACGACTCACTATAGG 3') and *NANOG*-specific inner sense primer (5' CACTGATTTATTTCCCAA CTAC 3'). The resulting PCR products of 5'- and 3'-RACE were then fractionated and gel extracted (1.5% agarose gel). The purified PCR products were cloned into the pGEM-T Easy vector (Promega), and ligated products were transformed into One-Shot chemically competent cells (Invitrogen). Recombinant plasmid DNAs were isolated and purified using QIAprep® miniprep system (Qiagen) prior to sequencing.

Relative expression analysis of alternative polyadenylation transcripts

Relative expression of transcripts with alternative polyadenylation was performed by real-time PCR. Briefly, around 50-passage ESC-like cells were washed in PBS and total RNA was isolated by TRI reagent (Ambion). After first-strand cDNA synthesis, RNase H treatment was performed. Primers were designed using poly(A) tail region in antisense primers to avoid overlapped amplification. The primer pairs used to amplify specific transcript (amplified PCR product size in brackets) were *BbuNANOG1* (GenBank accession No. JN231312; sense, 5' GTGTCAATTTGAGGGAAGGG 3'; antisense, 5' TTTTTTTTGCCCCCTGTGCT 3' [177 bp]), *BbuNANOG2* (GenBank accession No. JN231313; sense, 5' GGG AGGTCAACATGGAAATG 3'; antisense, 5' TTTTTTTT ACTCACTTCTAGTC 3' [91 bp]), *BbuNANOG3* (GenBank

accession No. JN231316; sense, 5' GGGAGGTCAACATG GAAATG 3'; antisense, 5' TTTTTTTTACAATGGCTATTT 3' [59 bp]), *BbuNANOG4* (GenBank accession No. JN231315; sense, 5' GGGAGGTCAACATGGAAATG 3'; antisense, 5' TTTTTTTTAAATGTAAAATGG 3' [59 bp]), *BbuNANOG5* (GenBank accession No. JN231314; sense, 5 GGGAG GTCAA CATGGAAATG 3'; antisense, 5' TTTTTTTTAAAA TGGCT ATTTT 3' [59 bp]), and *GAPDH* gene as an internal control (sense, 5 TTTGTGATGGGCGTGAACC 3'; antisense, 5' ACA GTCTTCTGGGTGGCAGT3' [173 bp]). The annealing temperature was 60°C for all PCRs. The cDNA sample was amplified in triplicate using SYBR Green Master Mix (Bio-Rad). PCR was run using MJ minithermal cycler (Real-time PCR; Bio-Rad). PCR program was 95°C for 10 min followed by 40 cycles of 10 s at 95°C, 10 s at 60°C, and 30 s at 72°C. A melt curve verified individual PCR amplicons. Results are presented as means $\pm$ standard error of the mean ($n=3$). The CPs determined for the different polyadenylation transcripts were normalized with the housekeeping *GAPDH* gene. The lowest expression was set as 1, and differences of other are expressed by the x -fold difference.

Gene expression analysis

RNA was extracted from fetal tissues (liver and heart), adult tissues (liver, brain, and heart), two-cell, eight-cell, morula, and blastocyst stages, and 10th and 60th passages of ESC-like cells. Briefly, 1 μ g of total RNA was reverse-transcribed using SuperScript III First-Strand Synthesis System (Invitrogen). PCR was run using MJ minithermal cycler (Bio-Rad) and the cycling parameters were 95°C for 5 min followed by 25 cycles of 30 s at 95°C, 30 s at 60°C, and 30 s at 72°C. Five microliters of PCR products were analyzed on 2% agarose gel.

Isolation and cloning of 5'-flanking region of buffalo NANOG

A genomic region of the buffalo *NANOG* promoter was cloned using PCR-based strategy. The overlapping primers

were designed based on closely related species bovine; genomic assembly was accessed from Ensemble database (www.ensembl.org). The DNA was isolated using DNAzol reagent (MRC). Overlapped fragments were amplified by PCR and cloned into the pGEM-T vector. The 3.4-kb upstream region of buffalo *NANOG* was cloned and sequenced and submitted to NCBI (GenBank accession No. AY512658). Primers used for amplification of 5'-flanking region were pair 1 (sense, 5' TGTGGCTGTGGTGTAAAGAGC 3'; antisense primer, 5' CAAGGCCATTGTAAATGCAAA 3'), pair 2 (sense, 5' CAGAGGCTCTATACCTGAAACACA 3'; antisense primer, 5' TTGTTGCAAGAAACAAGGAAAGTA 3'), pair 3 (sense, 5' CCCAACAAACCCACAGTCTC 3'; antisense primer, 5' GGACTTCCCCAGCAGTCTAGT 3'), and pair 4 (sense, 5' ATCCCTTATCCTTGACAAT 3'; antisense primer, 5' GCCACCAGGTACATTGGTTT 3'). The schematic position of primers is shown in Figure 1. The positive clones were sequenced in both sense and antisense orientations. Nucleotide position of *NANOG* promoter sequence has been designated by setting +1 as the translation initiation site.

In silico analysis: comparative genomics

Nucleotide sequence data reported here are available in GenBank database under the following accession numbers: HM585138, HM585139, HM585140, HM585144, HM585145, HM585146, DQ126153, JN231312, JN231313, JN231314, JN231315, JN231316, JN231317, JN231318, JN231319, JN231320, JN231321, JN231322, JN231323, JN231324, JN231325, JN231326, and JN231327. The comparative analysis of buffalo sequences and other mammals were carried out by alignment using ClustalW2 program. Pairwise comparison of nucleotide sequences was performed on EMBOSS Pairwise Alignment Algorithms program. The GC-rich regions were identified by using Gene Runner (Version 3.0) software (Hastings Software, Hudson). The 5'-flanking region of buffalo *NANOG* (promoter) was analyzed using TRANSFAC software (www.gene-regulation.com), TFSEARCH (www.cbrc.jp/research/db/TFSEARCH.html), and Genomatix software (www.genomatix.de).

FIG. 1. Schematic representation of the strategy used to determine the structural organization of the buffalo *NANOG* gene. The 5'-flanking region of *NANOG* (3427 nt) was isolated and cloned using overlapping PCR fragments as illustrated by the closed and open boxes for the sense and antisense primers, respectively. The coding region was amplified using cDNA by RT-PCR. The genomic DNA is depicted with exons as rectangular boxes and the introns as a solid line; 5'- and 3'-UTRs were determined by RLM 5'-RACE and 3'-RACE; primer positions for RACE are shown by arrows. The exon-intron boundaries were determined by PCR amplification and sequencing. RT-PCR, reverse transcription-polymerase chain reaction; UTRs, untranslated regions; 5'- and 3'-RLM-RACE, 5'- and 3'-RNA ligase-mediated rapid amplification of cDNA ends.

Results

Cloning and 5'- and 3'-RACE

The overall strategy used to clone and characterize the buffalo *NANOG* gene is shown in Figure 1. The 5'-flanking region was amplified from genomic DNA using overlapping PCR and the coding region was amplified from cDNA using RT-PCR followed by cloning. TSSs (5'-untranslated region [5'-UTR]) and polyadenylation sites (3'-UTR) were determined by RLM-RACE. The RLM 5'-RACE analysis identified the presence of multiple amplicons on gel (Fig. 2). The sequence analysis of independent clones of RLM-RACE products identified four transcripts with different TSS. The different 3'-UTRs of *NANOG* are shown in Figure 3. The location of TSSs are represented in Figure 4. For the verification of different TSS, 22 clones were sequenced, and each TSS was found to be supported by at least three sequenced clones.

Analysis of 3'-UTR/polyadenylation

In addition to TSS selection, alternative polyadenylation contributes to transcript miscellany. As the structure of 3'-UTR of *NANOG* gene have not been characterized in any species, we decided to investigate the formation of 3'-ends of this gene or whether alternative polyadenylation occurs. To investigate this, we performed RLM 3'-RACE. The results of 3'-RACE are shown in Figure 3. Figure 3A shows alternate splicing resulting in transcript variants having different poly(A) signals present in *NANOG* 3'-UTR. Sequencing results of 3'-RACE experiment identified five different *NANOG* transcripts with different 3'-UTR lengths of 269, 314, 560, 566, and 829 bp (Fig. 3B). The gel analysis of RLM 3'-RACE showed multiple bands (Fig. 3C). However, one transcript was not shown to have a canonical poly(A) signal (AAU

AAA) in upstream to cleavage/polyadenylation site but an AU-rich region present in upstream. The AU-rich region can act as a polyadenylation signal (Fig. 3D). The polyadenylation for *NANOG* gene could be an AAUAAA-dependent or -independent process in ESC-like cells.

Transcript variants and organization of buffalo *NANOG* gene

Sequence analysis based on the coding region, 5'-RACE, and 3'-RACE showed the presence of a splice variant of *NANOG* (Figs. 3 and 4). However, most of *NANOG* transcripts span four exons. The results of the present study showed that, in addition, an alternatively spliced variant coexists with these transcripts, consisting of five exons (GenBank accession No. JN_231312). Sequence alignment of this transcript variant with other transcripts revealed that a 290 bp long intron was further spliced out from exon 4. Splice junctions between the first four exons were in agreement with the GT-AG splicing rule, but exon 5 did not follow this rule as shown in Figure 4. The overall organization of buffalo *NANOG* gene spans ~6 kb. The first splice variant is separated by three intronic DNAs. The first exon comprised of the variable 5'-UTR followed by 136 bp that encode part of the protein. The second and third exons contain protein-coding sequence of 263 and 87 bp, respectively. The fourth exon contains the stop codon and the variable 3'-UTR (Fig. 4). The locations of the intron was confirmed by sequence analysis. As shown, the exon-intron boundaries conform to classical splice donor and acceptor consensus sequences. The exon sequence agreed with that determined for the cDNA, indicating that the obtained cDNA was free of PCR artifacts. The buffalo *NANOG* transcripts differ in size because of alternative usages of TSSs and alternative polyadenylation sites. At least five different transcripts occur because of usage of alternative polyadenylation, which contains 269, 314, 560, 566, and 829-bp-long 3'-UTR. The longest transcript of buffalo *NANOG* spans 1914 bp and includes a 930-bp ORF that encodes for 300 amino acids. The coding region of buffalo *NANOG* showed 96% identity to the bovine at nucleotide and amino acid levels.

Expression of *NANOG* transcripts in ESC-like cells

To investigate whether the 3'-UTR variants of the buffalo *NANOG* express differentially, we did transcript specific RT-PCR. The analysis of expression of transcripts with different poly(A) signals showed that the transcript variant using pA1 site for polyadenylation was expressed at a higher level. This transcript represents a splicing variant containing five exons. In addition, another transcript that uses pA4 site for polyadenylation was also expressed at a higher level in buffalo ESC-like cells when compared with other variants (Fig. 5). The longest transcript was found to be expressed at the lowest level (Fig. 5). Similarly, overall expression of *NANOG* in buffalo ESC-like cells and other tissues was checked. Results showed that *NANOG*, a pluripotency gene, was found to be expressed in ESC-like cells and two-cell, eight-cell, morula, and blastocyst stages but not in other fetal and adult tissues studied (Fig. 6). The expression of *NANOG* was found to be higher in ESC-like cells when compared with two-cell, eight-cell, morula, and blastocyst stages and remained stable in ESC-like cells after several passages (Fig. 6).

FIG. 2. RNA ligase-mediated 5'-RACE analysis of *NANOG*. Lanes 1 and 2 are 1-kb DNA marker and products of the inner PCR for 5'-RACE, respectively. RLM 5'-RACE revealed that the buffalo *NANOG* uses at least four TSSs. TSSs, transcription start sites.

FIG. 3. Structure of the buffalo *NANOG* 3'-UTR. The 3' portions of *NANOG* gene leading to alternative poly(A) site selection are shown. Five alternative poly(A) sites present in buffalo *NANOG* are diagrammed here for comparison. **(A)** Organization of different transcripts of buffalo *NANOG* gene encoded by two splice variants. **(B)** Diagrammatic representation of transcripts due to utilization of alternative polyadenylation. I-V, showed different transcripts of *NANOG* containing 269, 314, 560, 566 and 829 nt long 3' UTRs (NCBI acc. nos. JN_231312, JN_231313, JN_231316, JN_231315 and JN_231314) respectively. **(C)** Gel analysis of 3'-RACE showing the presence of three different amplicons: M: DNA marker; lane 1: inner 3'-RACE PCR; lane 2: outer 3'-RACE PCR. **(D)** Alignment of polyadenylation regions: four out of five transcripts utilized canonical poly(A) signal (AATAAA) for polyadenylation (bold). One transcript utilized UA-rich region (underlined) as a noncanonical poly(A) signal for polyadenylation. The 3'-noncoding region of *NANOG* contains three poly(A) sites within the 3'-UTR in the exon 4 box, arranged one behind the other in tandem. These three poly(A) signal sequences responsible for genesis of five different transcripts due to utilization of five different cleavage sites. The site of the termination codon of the protein in the fourth exon is indicated by the dotted vertical bar and the asterisk. Poly(A), polyadenine.

Structure of *NANOG* 3'-UTR and in silico analysis

Analysis of the 3'-UTR of *NANOG* showed that there were two splice variants and the overall structure of *NANOG* 3'-UTRs is shown in Figure 7. Splice variant type 1, encoded by four exons, contains four cleavage sites. The 3'-UTR of longest transcript was further analyzed for putative regulatory regions (1914 bp; GenBank accession No. JN_231314) *in silico* (Fig. 7A). The putative downstream polyadenylation signal was composed of a hexanucleotide sequence (AAUAAA) present at 525 and 769 bp downstream of the coding region (for polyadenylation at pA1, pA3, pA4, and pA5). Sequence analysis revealed that a *NANOG* transcript (GenBank accession No. JN_231313) lacks the canonical hexanucleotide sequences, A(A/U)UAAA, as a polyadenylation signal sequence, but an AU-rich sequence present in upstream to cleavage/polyadenylation site pA2 could be a putative

polyadenylation signal element for polyadenylation. Splice variant type 2, encoded by five exons, contains a single cleavage site (pA1). The canonical hexanucleotide (AAUAAA) signal for polyadenylation was found to be positioned 239 bp downstream of the coding region (GenBank accession No. JN_231312). The sequence analysis of *NANOG* mRNA 3'-UTR revealed the existence of several conserved motifs such as U-rich elements (UREs), AU-rich elements (AREs), and GU-rich elements (GREs) and cytoplasmic polyadenylation elements (CPEs).

NANOG 5'-flanking region

The 5'-flanking region (−3366/+61) of the buffalo *NANOG* gene was cloned by employing PCR-based amplification and cloning method as shown in the schematic representation (Fig. 1). As shown in Figure 8, the cloned

and SOX2 binding centered at position -208, -329, -1046, and -1095; one more site was present for Octamer binding centered at -2439; four more sites were present for Sox binding at -929, -1898, -3166, and -3350. Similarly, consensus sites for binding of many other transcription factors were identified as highlighted in Figure 8.

NANOG is a critical transcription factor in the regulation of cell fate of the pluripotent ICM during embryonic development, maintaining the pluripotent epiblast and preventing differentiation to primitive endoderm (Mitsui *et al.*, 2003;

FIG. 5. Differential expression of the buffalo *NANOG* variants. Relative expression of three *NANOG* variants in buffalo ESC-like cells was determined by real-time PCR. The experiments were performed in triplicate and the results are expressed as mean $\pm$ standard deviation. RNA samples representing 60th-passage buffalo ESC-like cells. *GAPDH* mRNA expression was used to normalize their expression. The expression of different transcripts varies significantly. A splice variant containing five exons is expressed more abundantly (GenBank accession No. JN231312). In the other splice variant, a transcript with a medium-length 3'-UTR (1651 bp; GenBank accession No. JN231315) was found to be expressed at a relatively high level when compared with other transcripts. ESC, embryonic stem cell.

Hamazaki *et al.*, 2004). To understand the transcriptional regulation of the *NANOG* gene in buffalo ESC-like cells, in this present study, we isolated and characterized *NANOG* gene and its 5'-flanking region.

5'-RACE analysis identified four transcription initiation sites (TSSs), including three novel initiation sites located in exon 1 of the gene expressed by buffalo ESC-like cells. A distal most TSS was located at position -182 nucleotides upstream to translation start site. The presence of multiple transcriptional initiation sites is a typical feature of genes having TATA-less promoters (Frith *et al.*, 2008). In mice, previous study showed that the *NANOG* gene promoter lacks a TATA box and CAAT box (Wu and Yao, 2005). In contrast, in buffalo, the 5'-flanking region of *NANOG* contained both a TATA box and CAAT box upstream of the distal most TSS. Therefore, in buffalo, it could be interesting to study whether these TATA and CAAT boxes are func-

tional. However, because of experimental limitations with buffalo ESC-like cells, we could not perform deletion analysis in the present study. Many genes have multiple TSSs located in close proximity to each other and the rules for start site selection are fundamentally different for different promoters. A TSS is defined as a unique nucleotide that is the first to be transcribed, whereas the core promoter is defined as a genomic region that spans this and close TSSs (Frith *et al.*, 2008). We identified the four TSSs (-17, -35, -95, and -182) that are responsible for transcription initiation of the *NANOG* gene. Interestingly, our observation for *NANOG* gene expression supports that two different strategies are used by Pol II for transcription initiation. The first strategy is "focused initiation" in which a single TSS and the core promoter contain a TATA box and other core promoter elements (INR, downstream promoter element [DPE], and transcription factor IIB (TFIIB) recognition element [BRE]) (Juven-Gershon *et al.*, 2006). On the other hand, in the second strategy, multiple weak TSSs are dispersed over DNA regions of ~50–150 bp, and it is thereby dubbed "dispersed initiation" (Juven-Gershon *et al.*, 2006). *NANOG* appeared to use the second strategy for its expression in ESC-like cells. The mechanisms of dispersed initiation are not clear but probably involve selective usage of numerous upstream and downstream recognition and promoter elements.

Polyadenylation is an integrated step in the maturation of all eukaryotic cellular mRNAs with the exclusion of histone mRNAs. Regulation can also occur after transcription using sequences in the 3'-UTR of the mRNA to affect mRNA stability and/or translation efficacy (de Moor *et al.*, 2005). Alternative polyadenylation generates mRNAs with 3'-untranscribed regions of different lengths, often affecting transcript stability. Because the precise regulation of *NANOG* expression level is essential for ESC-like cells' pluripotency, we characterized 3'-UTR using 3'-RACE. 3'-RACE experiments showed the presence of tandem polyadenylation signals, which leads to the expression of at least five different 3'-UTRs (269, 314, 560, 566, and 829 nt) (Fig. 3). There are other several examples in which the 3'-UTRs of mRNAs play significant roles in regulating gene expression (Chen and Shyu, 1995; Edwalds-Gilbert *et al.*, 1997; Zhao *et al.*, 1999; Pesole *et al.*, 2001; Zhang *et al.*, 2002). The alternative polyadenylation could also play a role in gene silencing. Messenger RNA (mRNA) 3'-end processing defines the end of the transcript through endonucleolytic cleavage of the precursor transcript, provides a protective polyadenylate tail, and enables subsequent termination of transcription by RNA Pol II. Just as alternative splicing allows enormous diversity of mRNA products from a limited number of genes,

FIG. 6. Expression of the buffalo *NANOG* gene. *NANOG* expression was assessed by RT-PCR. Upper panel shows expression of *NANOG* in ESC-like cells and two-cell, eight-cell, morula, and blastocyst stages, but no

detectable expression was seen in the fetal and adult tissues. RNAs were from FL, fetal liver; FH, fetal heart; FF, fetal fibroblast cells; 2-C, two-cells embryo; 8-C, eight-cells embryo; M, morula; B, blastocysts; E10, ESC-like cells of 10th passage; E60, ESC-like cells of 60th passage; AL, adult liver; AB, adult brain; AH, adult heart. M, DNA marker; NTC, no template control. *GAPDH* was used as an internal control and expression is shown in the lower panel.

FIG. 7. Schematic representation of *NANOG* mRNA 3'-UTR. *NANOG* mRNA 3'-UTR sequences were from two splice variants. **(A)** In splice variant transcript 1, four tandem cleavage sites are present (pA1–4). The two polyadenylation signals are positioned at 525 and 769 bp downstream of the coding region, and a third element forming putative polyadenylation signal at an AU-rich sequence is highlighted in gray box and correlates with a *NANOG* transcript (GenBank accession No. JN231313), which lacks the canonical hexanucleotide sequence, A(A/U)UAAA, as a polyadenylation signal. The 5'-terminus defined by 5'-RACE beginning 182 nucleotides upstream of the translation start is set as the start of transcription. Nucleotide numbering is based on *NANOG* transcript sequences (GenBank accession No. JN231314). **(B)** In splice variant transcript 2 (GenBank accession No. JN231312), a polyadenylation signal is positioned at 239 bp downstream of the coding region. The cleavage site is shown as pA5, and AUUUA motifs (dark gray) and U/GU-rich sequences (gray) are highlighted or boxed. Binding sites for PCBP (UCCCCU), PTB (UCUUU), AUBF (AUUUA), and other factors are underlined. AUBF, adenosine-uridine binding factor. PCBP, poly(c) binding protein; PTB, polypyrimidine tract binding protein; PAS, polyadenylation site; DSE, downstream sequence element.

A *NANOG* splicing variant 1

B *NANOG* splicing variant 2

in animals and plants it is estimated that >50% of genes have alternative polyadenylation sites. The most common mechanism is one in which alternative polyadenylation is a consequence of tandem arrays of poly(A) signals within a single 3'-UTR. The differential expression of *NANOG* gene that undergoes alternative poly(A) site choice or polyadenylation/splicing competition could be regulated at the level of amounts and activities of either generic or tissue-specific polyadenylation factors and/or splicing factors. Relative expression analysis of preferential uses of poly(A) signal revealed that the transcript variant using the pA1 site for polyadenylation was expressed at a higher level. In addition, another transcript that uses pA4 site for polyadenylation was also expressed at a higher level in buffalo ESC-like cells when compared with other variants. The longest transcript was

found to be expressed at the lowest level (Fig. 5). The 3'-UTR is often the site for RNA elements that regulate various steps in the mRNA-processing pathway, such as mRNA transport, transcript stability, and translation efficiency. The best-studied 3'-UTR element is the group of AU-rich RNA instability elements, often containing multiple copies of AUUUA motifs (Zhang *et al.*, 2002).

Sequence analyses of the *NANOG* 3'-UTR revealed the existence of several conserved motifs such as UREs, AREs, and GREs and CPEs in 3'-UTR. These types of regulatory sites are functional regulators involved in post-transcriptional regulation of gene expression (Halees *et al.*, 2011). CPEs are usually located 20–30 nucleotides upstream of the AAUAAA element, although several examples of mRNAs with much longer CPE-to-AAUAAA distance have been described.

FIG. 8. Nucleotide sequence of the 5'-flanking region of the buffalo *NANOG* gene. The nucleotide sequence contains 3366 nt 5' to the translation start site. The translational start site was designated as +1. The TSSs are underlined and indicated by arrows (positions -182, -95, -35, and -17). The sequence was analyzed for regulatory elements that share homology to known transcription factor binding sites, using the TFSEARCH program. Putative transcription and regulatory elements are highlighted in gray boxes. Putative TATA box and CAAT box are double underlined. Four Oct-Sox composite sites are highlighted in gray and dark gray boxes.

Cytoplasmic polyadenylation is a key mechanism affecting genes that are involved in controlling mRNA translation during early development (de Moor *et al.*, 2005). AREs are well-described sequence elements to which a range of different proteins can bind, for example, AUF1, HuR, and KSRP. AREs and GREs are largely present in the 3'-UTRs of many unstable mammalian mRNAs, responsible for destabilization/degradation of mRNAs. Genome-wide analysis showed that at least 5% of human genes contain GREs in their 3'UTRs, with functional over-representation in genes involved in transcription, nucleic acid metabolism, developmental processes, and neurogenesis. RNA-binding protein CUGBP1 elicits mRNA decay through binding to GREs (Vlasova *et al.*, 2008). *NANOG* expression is limited to few stages of development or in pluripotent cells; in this way, transitory expression of *NANOG* mRNA may be regulated post-transcriptionally by these de-

stabilization elements present in the 3'-UTR. *NANOG* contains multiple regulatory elements or binding sites for many proteins, such as PCBP, poly(c) binding protein (UCCCCU); PTB, polypyrimidine tract-binding protein (UCUUU) and AUBF, adenosine-uridine binding factor (AUUUA), in the 3'-UTR. Conserved AREs have been reported to be mainly present in the 3'-UTR of a variety of mRNAs susceptible to rapid degradation (Winstall *et al.*, 1995; Sully *et al.*, 2004). The cytoplasmic trans-factors interact at or near the AUUUA, and other destabilizing motifs could modulate the degradation of *NANOG* message accounting for variable message stability.

The third important mechanism that generates huge diversity of transcripts is alternative splicing. 3'-RACE revealed that there were two sorts of splicing for premature *NANOG* mRNA. Two different splice variants were found, one splice variant contains four exons and other has five

exons (Fig. 3). Splicing variation occurred in the buffalo *NANOG* gene because of an additional splicing event in exon 4, giving rise to a distinct variant of buffalo *NANOG*. Generally, the fourth exon partially codes for *NANOG* protein and the rest codes for the 3'-UTR. Our results show that a strong conservation of gene structure and sequence exists between buffalo *NANOG* and its bovine, human, and mouse counterparts and that *NANOG* is expressed in the buffalo ESC-like cells and in embryonic developmental stages, such as two-cell, four-cell, morula, and blastocyst stages, but not in other somatic tissues of adult or fetus (Fig. 6).

Further, the 5'-flanking sequence of the buffalo *NANOG* gene contains a consensus TATA box and CAAT box. The 5'-RACE analysis identified multiple transcription initiation sites in exon 1 of the *NANOG* gene expressed by buffalo ESC-like cells. The presence of multiple transcriptional initiation sites is a typical feature of TATA-less promoter regions. However, the present finding in buffalo reveals the presence of both features, that is, TATA box and CAAT box in upstream of the distal most TSS and multiple TSSs. So it could be interesting to study whether the TATA box and CAAT box are functional or not or whether both strategies can be used for transcription. It is now well documented that TATA-driven transcription preinitiation complex assembly is the exception rather than the rule in eukaryotic transcription, as only 10%–20% of mammalian promoters contain a functional TATA box (Hochheimer and Tjian, 2003). Taken together, in the initial characterization of the promoter of the buffalo *NANOG* gene, four TSSs were detected and two promoter regions were established. The promoter region and 5'UTR contain multiple potential transcription factor binding sites, including Sp1 (GC-box), a ubiquitously expressed transcription factor belonging to the Sp1-like/KLF family of transcription factors. The role of these two putative Sp1 binding sites located in the *NANOG* gene 5'-flanking region may participate in the regulation of *NANOG* gene transcription (Wu and Yao, 2006). The TFSEARCH revealed the presence of multiple binding sites for OCT4 and SOX2 within 3.4 kb of the *NANOG* 5'-flanking region. *NANOG* has been reported to be regulated by OCT4 and SOX2 through a co-operative interaction (Rodda *et al.*, 2005). Consensus sites for binding of many other transcription factors including Ap1, myeloid zinc finger (MZF1), CCAAT/enhancer binding protein (C/EBP), ADR1, Bcd, CdxA, GATA1, GATA-2, c-Myb, lymphoid transcription factor (Lyl-1), MyoD, IRF-1, STRE, and NF-Kap were identified (Fig 8).

Conclusion

In conclusion, we cloned, sequenced, and analyzed the buffalo ortholog of the *NANOG* gene. The present finding demonstrated that in buffalo ESC-like cells the *NANOG* mRNA was expressed as different isoforms that include alternative transcriptional start, alternative splicing, and alternative polyadenylation sites. Specifically, four TSSs were identified, one corresponding to that previously described in bovine and three that were novel for *NANOG*. Beside the variants generated by the 5'-UTR, *NANOG* has five different variants because of the selection of different polyadenylation signals in the 3'-UTR. The combination of alternative 5'- and 3'-UTRs increases the convolution of *NANOG* expression

and the possible roles of RNA isoforms in ESC physiology. Further, analysis of the 5'-flanking region revealed the presence of a TATA box and CAAT box sequence elements in the proximal promoter region; all the TSSs were found to be located within these motifs. The expression analysis of three variants that use alternative polyadenylation signals showed that these transcripts were differentially expressed. Therefore, results showed that alternative *NANOG* transcripts generated with different 5'- and 3'-UTRs could account for differential expression. However, the functional relevance of different transcripts with alternative transcript start sites and polyadenylation is to be yet studied for this key regulator of ESC pluripotency regulator transcription factor.

Acknowledgments

This work was funded by the National Agriculture Innovative Project (NAIP) grant to M.S.C. (C-2067 and 075), ICAR, New Delhi, India.

Disclosure Statement

No competing financial interests exist.

References

- Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113**, 643–655.
- Chen, C.Y., and Shyu, A.B. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci* **20**, 465–470.
- Danckwardt, S., Hentze, M.W., and Kulozik, A.E. (2008). 3' end mRNA processing: molecular mechanisms and implications for health and disease. *The EMBO J* **27**, 482–498.
- de Moor, C.H., Meijer, H., and Lissenden, S. (2005). Mechanisms of translational control by the 3' UTR in development and differentiation. *Semin Cell Dev Biol* **16**, 49–58.
- Edwards-Gilbert, G., Veraldi, K.L., and Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res* **25**, 2547–2561.
- Frith, M.C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., and Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. *Genome Res* **18**, 1–12.
- George, A., Sharma, R., Singh, K.P., Panda, S.K., Singla, S.K., Palta, P., Manik, R., and Chauhan, M.S. (2011). Production of cloned and transgenic embryos using buffalo (*bubalus bubalis*) embryonic stem cell-like cells isolated from *in vitro* fertilized and cloned blastocysts. *Cell Reprogram* **13**, 263–272.
- Gross, P., and Oelgeschlager, T. (2006). Core promoter-selective RNA polymerase II transcription. *Biochem Soc Symp* **73**, 225–236.
- Halees, A.S., Hitti, E., Al-Saif, M., Mahmoud, L., Vlasova-St. Louis, I.A., Beisang, D.J., Bohjanen P.R., and Khabar K.S.A. (2011). Global assessment of GU-rich regulatory content and function in the human transcriptome. *RNA Biol* **8**, 681–691.
- Hamazaki, T., Oka, M., Yamanaka, Y., *et al.* (2004). Aggregation of embryonic stem cells induces *NANOG* repression and primitive endoderm differentiation. *J Cell Sci* **117**, 5681–5686.
- Hochheimer, A., and Tjian, R. (2003). Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes Dev* **17**, 1309–1320.

- Juven-Gershon, T., Hsu, J.Y., and Kadonaga, J.T. (2006). Perspectives on the RNA polymerase II core promoter. *Biochem Soc Trans* **34**, 1051–1054.
- Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein NANOG is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631–642.
- Pan, G., and Thomson, J.A. (2007). NANOG and transcriptional networks in embryonic stem cell pluripotency. *Cell Res* **17**, 42–49.
- Pesole, G., Mignonea, F., Gissia, C., Grillo, G., Licciulli, F., and Liunib, S. (2001). Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **276**, 73–81.
- Proudfoot, N. (2004). New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol* **16**, 272–278.
- Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. *Cell* **108**, 501–512.
- Rodda, D.J., Chew, J.L., Lim, L.H., Loh, Y.H., Wang, B., Ng, H.H., and Robson, P. (2005). Transcriptional regulation of NANOG by OCT4 and SOX2. *The J Biol Chem* **280**, 24731–24737.
- Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**, 424–436.
- Sully, G., Dean, J.L.E., Wait, R., Rawlinson, L., Santalucia, T., Saklatvala, J., and Clark A.R. (2004). Structural and functional dissection of a conserved destabilizing element of cyclo-oxygenase-2 mRNA: evidence against the involvement of AUF-1 [AU-rich element/poly(U)-binding/degradation factor-1], AUF-2, tristetraprolin, HuR (Hu antigen R) or FBP1 (far-upstream-sequenceelement-binding protein 1). *Biochem J* **377**, 629–639.
- Verma, V., Gautam, S.K., Singh, B., Manik, R.S., Palta, P., Singla, S.K., Goswami, S.L., and Chauhan, M.S. (2007). Isolation and characterization of embryonic stem cell-like cells from *in vitro*-produced buffalo (*Bubalus bubalis*) embryos. *Mol Reprod Dev* **74**, 520–529.
- Vlasova, I.A., Tahoe, N.M., Fan, D., Larsson, O., Rattenbacher, B., SternJohn, J.R., Vasdevani, J., Karypis G., Reilly, C.S., Biterman, P.B., and Bohjanen, P.R. (2008). Conserved GU-rich elements mediate mRNA decay by binding to CUG-binding protein 1. *Mol Cell* **29**, 263–270.
- Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476.
- Winstall, E., Gamache, M., and Raymond V. (1995). Rapid mRNA degradation mediated by the *c-fos* 3' AU-rich element and that mediated by the granulocyte-macrophage colony-stimulating factor 3' AU-rich element occur through similar polysome-associated mechanisms. *Mol Cell Biol* **15**, 3796–3804.
- Wu, D.Y., and Yao, Z. (2005). Isolation and characterization of the murine NANOG gene promoter. *Cell Res* **15**, 317–324.
- Wu, D.Y., and Yao, Z. (2006). Functional analysis of two Sp1/Sp3 binding sites in murine NANOG gene promoter. *Cell Res* **16**, 319–322.
- Zhang, T., Kruijs, V., Huez, G., and Gueydan, C. (2002). AU-rich element-mediated translational control: complexity and multiple activities of trans-activating factors. *Biochem Soc Trans* **30**, 952–958.
- Zhao, J., Hyman, L., and Moore, C., (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**, 405–445.

Author correspondence to:
Dheer Singh, Ph.D.

Molecular Endocrinology Laboratory
Animal Biochemistry Division
National Dairy Research Institute (NDRI)
Karnal 132001, Haryana
India

E-mail: drdheer.singh@gmail.com

Received for publication August 8, 2011; received in revised form September 2, 2011; accepted September 2, 2011.