

SSR repeat dynamics in mitochondrial genomes of five domestic animal species

Sushil Kumar Shakyawar^{1,2}, Balwinder Kumar Joshi², Dinesh Kumar^{2*}

¹Department of Biotechnology, Indian Institute of Technology, Guwahati-781 039, Assam, INDIA; ²Genes & Genetic Resources Molecular Analysis Lab, National Bureau of Animal Genetic Resources, Karnal-132 1001, Haryana, INDIA; Dinesh Kumar – Email: dinesh@iastate.edu; *Corresponding author

Received June 13, 2009; Revised August 03; Accepted September 11, 2009; Published October 15, 2009

Abstract:

SSR (simple sequence repeats) are ubiquitously abundant in genomes. In organellar mitochondrial genome of animals, its distribution, size dynamics and effectiveness for phylogenetic relationship have not been understood. Present investigation reveals organisation of SSR in genic and intergenic region, its length and repeat motif dynamics, extent of conservation of flanking regions, appropriateness of these SSR data in establishing phylogenetic relationship. Contrary to eukaryotic nuclear abundance of SSR in non-coding region, we found abundance in coding region. Like nuclear SSR, most hyper mutable repeats were found in non coding region having di nucleotide motifs of mitochondrial genome but contrary to human having high mutable tetra repeats in case of mitochondrial genomes this was found to be with tri-motif repeats. SSR of mitochondrial genomes also show cyclical expansion and shrinkage in pattern of SHM (simple harmonic motion) with respect to time its non-linear thus not appropriate for phylogenetic analysis though the flanking regions of these SSR also conserved like nuclear SSR.

Key words: Cyclical expansion, Distribution, Dynamics, Phylogenetic relationship, SHM, SSRs.

Background:

The near-absence of genetic recombination and high mutation rate with some selectivity in mitochondrial DNA makes it a useful source for analyzing microsatellite or STR (simple tandem repeat) or simple sequence repeat (SSR) dynamics on the archaeological time scale. Difference between the mean repeat sizes of two lineages is a linear function of the time since they diverged [1]. Since some factors prevent allele from becoming staying large (since they cease to behave like microsatellite [2], so there is a maximum threshold value of SSR allele size beyond which alleles starts shrinking in size because of background mutation and repaired by DNA polymerase I up to minimum threshold allele size and again it starts increasing in its size as a function of time. Thus it's cyclical in nature in terms of elongation and shrinkage of repeats over evolutionary passage of time. The flanking region constancy has been found to be extensively conserved across distant taxa. For example, presence of homologous loci in each test of marine species within two families (Cheloniidae and Dermochelyidae), as well as in a freshwater species (Emydidae and Trachemys scripta) is the indication for this constancy approximately over 300 million years of divergent evolution [3]. The flanking region is conserved and perpetuates down in evolution but the SSR loci reflects cyclical (shrinkage and expansion) variation in size over time which has been well documented in case of nuclear SSR [4].

Though earlier attempt has been made to relate the major cereal crops viz. rice, wheat, maize and sorghum phylogenetically using organellar genomes (mitochondria and chloroplast) SSR [5] but no such attempt has been made in case of animals. Unlike these crops, the relative abundance of type of repeat motif and dynamics across species in case of domestic animals are not documented with comparative study. In case of advanced eukaryotes these repeats are almost all in intronic (non-coding) regions but does similar situation exists in organellar SSR or not? It is yet to be known. Do these genomes have universal flanking regions bracketing SSR loci is not known. Like the case of crop studies [5] can we have a phylogenetic tree using SSR data of organellar genome are not known. The cyclical dynamism of repeat length over the archaeological time line has also not been studied in organellar genome.

The present *in silico* work in which five domestic animal species viz. buffalo, cattle, goat, sheep and yak has been taken as model, to investigate relative abundance of type of repeat motif, their distribution in coding and non coding regions, evaluation of mitochondrial SSR data for appropriateness of phylogenetic tree,

dynamism of length and repeat motif and extent of conservation of flanking regions across loci with respect to time.

Methodology:

Mining of mitochondrial genomes:

The mitochondrial genomes of five species viz. cattle (NC_005971), Sheep (NC_001941), Yak (NC_006380.3), Buffalo (NC_006295), and Goat (NC_005044) were downloaded from GenBank (<http://www.ncbi.nlm.nih.gov>).

Mining of STR from mitochondrial genomes:

For mining of microsatellite data especially type of repeat and motif viz di-, tri-, tetra, penta- and hexa- online programme SSRIT (Simple Sequence Repeat Identification Tools) were used [6]. An off line software FastPCR ([7], (<http://www.biocenter.helsinki.fi/bi/programs/fastpcr.htm>)) was used further to generate data on mono motif with iterative repeats >4. The existing annotations in GenBank were used while generating data on coding and non-coding regions of respective mitochondrial genomes (Table 1 in supplementary material). For a particular repeat motif in each specific genome, corresponding alleles in other genomes were identified by the presence of same flanking sequence.

Designing of universal primers:

SSR regions were identified in all the five mitochondrial genomes using FastPCR. A set of universal primers were designed (using online software Primer 3 ([8], <http://fokker.wi.mit.edu/primer3/input.htm>)) across flanking regions of these identified SSR (Table 2 in supplementary material). While designing the universal primers across species the null alleles (mutations in 3' end region of primer) encountered were bursted by locking one primer and increasing the expected PCR product length to get its counterpart compatible primer. Before using the designed primers for ePCR, *in silico* evaluation of primer quality was done on three parameters viz. self dimer, cross dimer and self hairpin loop of each of the primer pair. The delta G values of these evaluations are shown in (Table 2 in supplementary material).

ePCR on mitochondrial genomes:

All the five mitochondrial genomes subjected to analysis were at par in terms to become focal species for cross species electronic PCR amplification. For primer designing of two sets each from buffalo and yak (alphabetically ascending and descending in order) were treated as focal species to evaluate and further design the universal set of primer in remaining four 'heterologous' species. While designing the compatible universal primer in heterologous species,

the estimated PCR products size(s) were treated as species specific allelic data for length polymorphism and microsatellite dynamics.

Microsatellite dynamics analysis across species:

The generated data of both SSR loci (Table 2 in supplementary material) in five species were used to establish allele size dynamics (shrinkage/expansion) as a function of time (Figure 1). The established relationship of organelle SSR is compared with dynamism of nuclear SSR on time line which has been reported [4]. Comparative analysis of extent of mitochondrial SSR constancy across five species was done to observe magnitude of allelic size dynamism with respect to time.

Phylogenetic tree construction:

Phylogenetic tree was constructed using algorithms of both structural polymorphism as well as length polymorphism. Both rooted (Figure 2(A) and Figure 2(D)) structural tree was constructed by TreeTop (http://www.genebee.msu.su/services/phree_reduced.html) and unrooted (Figure 2(B) and Figure 2(E)) structural tree was constructed by Phylo dendron (http://iubio.bio.indiana.edu/treeapp/treeprint-sample1.html). The relative allelic length difference is calculated from the generated data and is further used to make rooted tree (Figure 2(C) and Figure 2(F)) on the basis of length polymorphism by using UPGMA method.

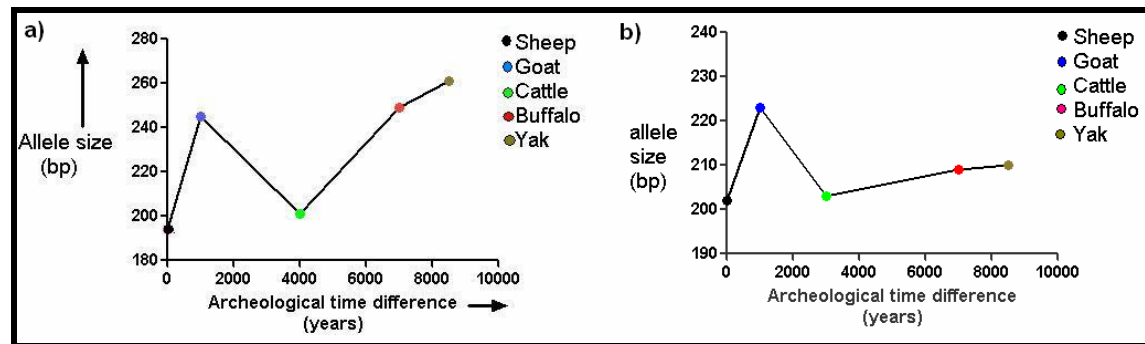


Figure 1: Allelic size variation on archaeological time line. (a) primer set I; (b) primer set II.

Discussion:

The *in silico* mining of SSR in mitochondrial genomes of five species reveals that they are more abundant in coding region unlike the nuclear genomes where SSRs are usually present in intronic/non coding regions [9]. The percent of SSR in coding regions of buffalo, cattle, sheep, goat, and yak mitochondrial genomes are 65.28 %, 68.59 %, 65.01 %, 64.40 % and 65.58 %, respectively. An average of 29.66 % of whole genome in all the 5 species (28.95 %, 29.59 %, 29.45 %, 30.31 % and 30.01 % in buffalo, cattle, sheep, goat and yak, respectively). Contrary to nuclear genome SSR interestingly, the organellar SSR shows more abundance in coding region than non coding region in all the mitochondrial genomes. Such SSR abundance in coding region has been reported in lower eukaryotes like [10]. This might be because of prokaryotic origin of mitochondria (endosymbiont hypothesis) or evolutionary legacy in lower eukaryotes.

The microsatellites mined from whole genomes were classified into two classes viz. class I containing only mono- repeat motif and class II containing di- to hexa- motif repeats. The density of class II microsatellites was found to be 108 bp-113 bp per kbp in exonic region and 118 bp-129 bp per kbp in intronic region. Di-motif repeats are abundant in each case (Table 1 in supplementary material). Maximum number (754) of di-motif repeats were found in goat and minimum is 704 in buffalo in coding region. In non-coding region it was relatively less abundant (maximum 396 in goat and min 350 in cattle). Least frequent repeats were of penta and hexa motifs. Goat mitochondrial genomes were with maximum number of penta- (50) and hexa- (24) motif repeats. Maximum numbers (155) of tetra- motif repeats were found in sheep and least (126) in cattle. Among the di-motif repeats (CA)_n were abundant where n varies from 148 to 204. Our data shows that hexa- and penta- repeat motif are less abundant which is because of selectivity involved on maintaining microsatellite within certain range. A similar size constraint in repeat number and length over the period in different taxa has been reported [10].

The mutational differences amongst different motifs di- (Figure 4(A) and Figure 4(B)), tri-(Figure 4(C) and Figure 4(D)) and tetra-(Figure 4(E) and Figure 4(F)) between 5 species were compared in

both coding and non coding region of genomes. Among non coding region di motifs were having faster mutation and in coding region tri motifs were with high mutation rate (Figure 3(A) and Figure 3(B)). This is in contrary to human STRs where tetra repeats shows hyper mutation [11]. Interestingly in rare cases in human coding region triplet repeat slippage event gives rise to copy number mutation or dynamic mutation example –Fragile X syndrome [12]. Microsatellite mutation processes have been inferred by direct observations both on artificial constructs in yeast [13] and in human pedigrees [14]. The general conclusion from these studies is that there is an exceptionally high rate of mutation adding or subtracting a small number of perfect repeats. In humans, the average overall mutation rate for 28 di- and tetra- nucleotide microsatellites was estimated at about 0.001, with the tetra nucleotide repeats significantly more mutable than the dinucleotide repeats. The most popular explanation for the high mutation rate is polymerase slippage [15], a hypothesis that received considerable support from an elegant *in vitro* analysis showing that polymerase tends to miscopy repeated tracks of DNA [16]. A sub set of SSRs namely trinucleotide plays important role in eukaryotes because of expansion of these triplet repeats, where the rate of mutation depends on the number of tandem units within the repeat, this is the basis of dynamic mutation [17]. The designed primer to generate SSR allelic data by electronic PCR using two sets of universal primers shows that the flanking regions are well conserved in all the 5 species. Such conservation of flanking regions of SSRs has been reported over a longer evolutionary period of time which is as high as 300 million years [3]. Though the flanking regions are conserved but STR loci shows differential alleles in all the five species investigated.

In all the five species, generated data of allele size dynamics (shrinkage/expansion) as a function of time (Figure 1) shows simple harmonic motion (SHM) pattern. This pattern is similar to established dynamism of nuclear SSR [4]. This is because of distinct mutational processes ([18, 19]), slippage mutation which involves the addition or subtraction of one repeat unit. SHM was shown by SSR but not by flanking region of SSR, the cause for this phenomenon is that, the mutation rate of SSRs which is 10⁻³/cell/generation that is hyper mutation which is 10⁶ times more

(addition of repeats >> deletion of repeats) where as in flanking region mutation (mutation rate is 10^{-9} /cell/generation i.e. background mutation rate) which works on both SSR region (concerning single simple repeats into compound interrupted repeats) as well as in flanking region[20].

Evaluation of mitochondrial SSR data for appropriateness of phylogenetic tree based on length polymorphism and structural polymorphism revealed that the trees are not in conformity with established phylogenetic relationship of these five taxa. The structural polymorphism based rooted (Figure 2(A)) and unrooted tree (Figure 2(B)) and length polymorphism data (allelic length difference) UPGMA based rooted tree (Figure 2(C)) all were not in

conformity. A second set of generated STR data for same set of tree shows same unusual pattern (Figure 2(D), Figure 2(E) and Figure 2(F)). Though the UGMA based tree of both set of data shows differences but again non-conformity with established phylogenetic relationship among five model species studied. SHM was shown by SSR but not by flanking region of SSR, the cause for this phenomenon is that, the mutation rate of SSRs which is 10^{-3} /cell/generation that is hyper mutation which is 10^6 times more (addition of repeats >> deletion of repeats) where as in flanking region mutation (mutation rate is 10^{-9} /cell/generation i.e. background mutation rate) which works on both SSR region (concerning single simple repeats into compound interrupted repeats) as well as in flanking region [20].

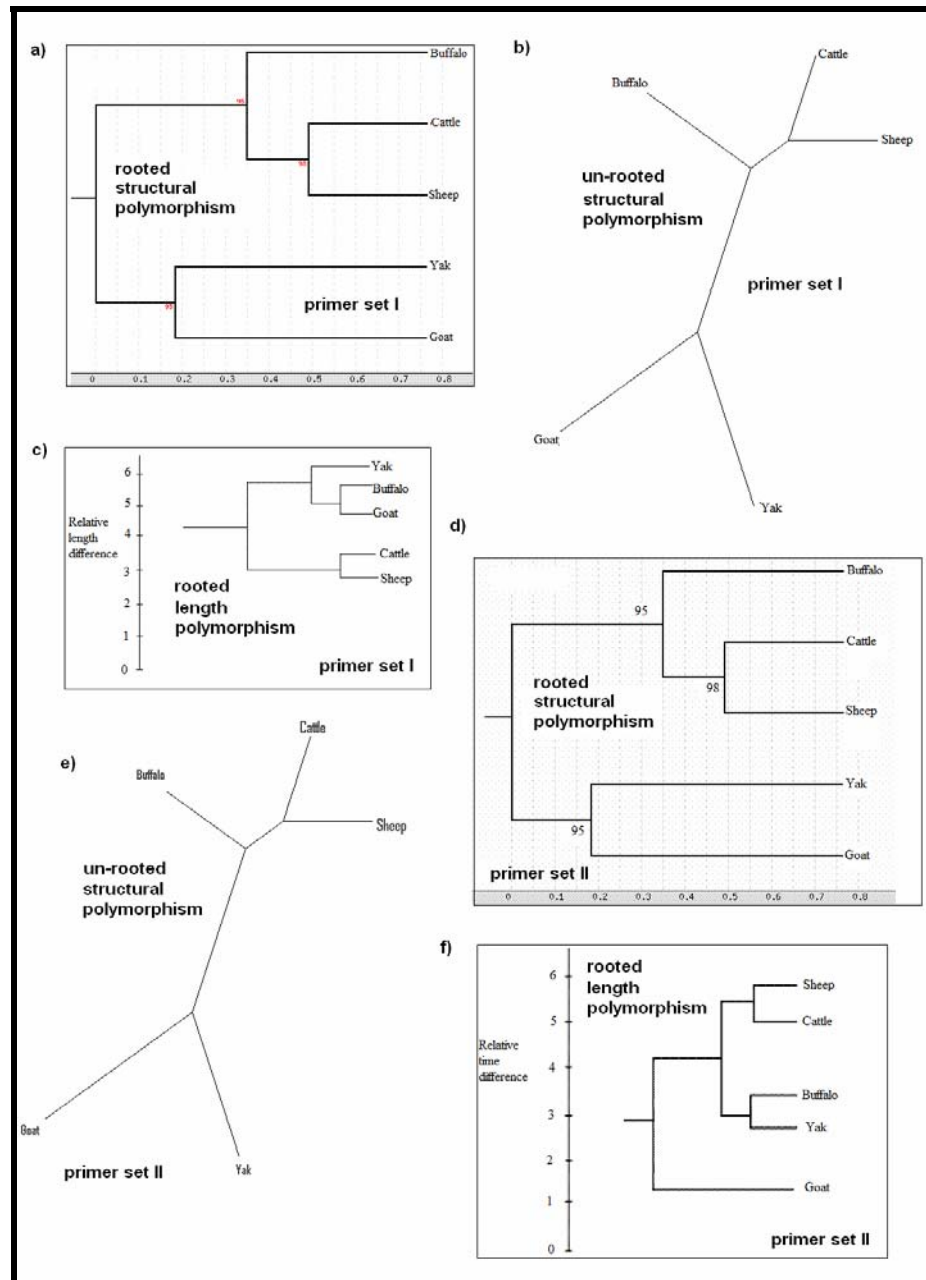


Figure 2: Rooted and unrooted trees based on structural and length polymorphism for primer set I and II.

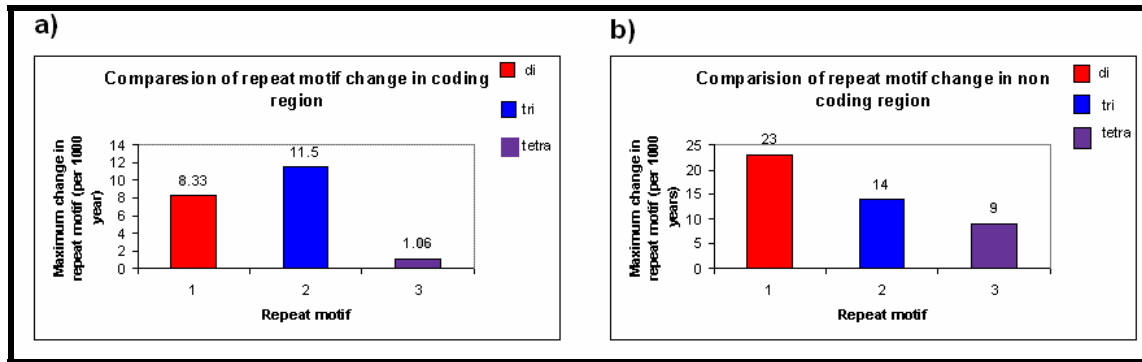


Figure 3: Repeat motif change in (a) coding region; (b) non coding region

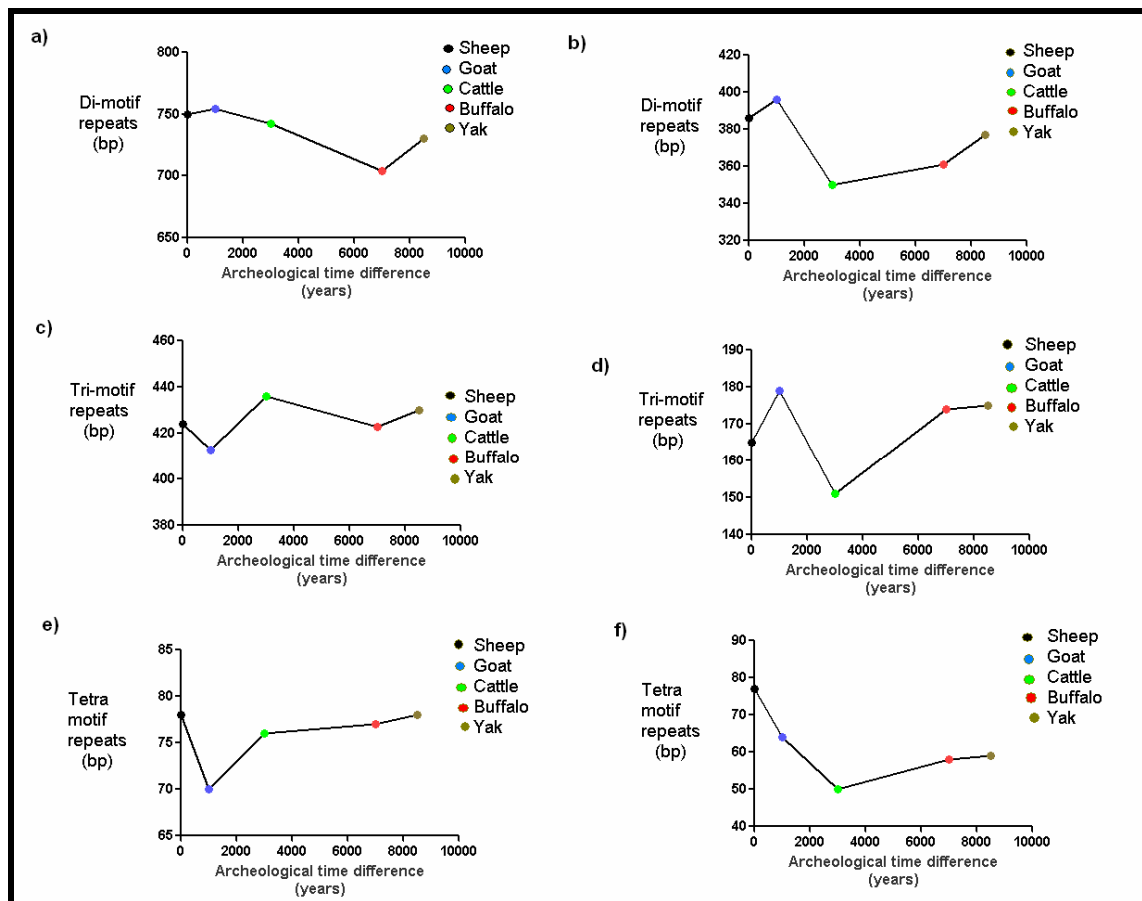


Figure 4: Di-motif repeat variation on archaeological time line (a: coding region b: non-coding region); Tri-motif repeat variation on archaeological time line (a: coding region b: non-coding region); Tetra-motif repeat variation on archaeological time line (a: coding region b: non-coding region);

Conclusion:

Present study revealed the relative abundance and motifs of SSR in mitochondrial genomes of five domestic species. Contrary to eukaryotic nuclear abundance of SSR in non-coding region, we found abundance in coding region. Like nuclear SSR, most hyper mutable repeats were found in non coding region having di-nucleotide motifs of mitochondrial genome but contrary to human having high mutable tetra- repeats in case of mitochondrial genomes this was found to be with tri-motif repeats. SSRs of mitochondrial genomes also show cyclical expansion and shrinkage in pattern of SHM with respect to long evolutionary time which is non-linear. Because of such SSR data are not appropriate for phylogenetic

analysis though the flanking regions of these SSRs also conserved like nuclear SSR. Lengths of SSRs obtained are much useful in predicting the influence of transcriptional activity in promoter region. The present work does not take account of intra-species variation in each species which can be further verified with reasonable sample size in wet lab work to revalidate the work further, hitherto not done.

Acknowledgment:

Authors are thankful to Director, NBAGR for kind permission and facilities for this work. The help and encouragement of Dr Utpal Bora, and Dr Pranab Goswami, Chairman, Department of

Biotechnology, Indian Institute of Technology Guwahati, Assam, India is thankfully acknowledged. The financial support of Ministry of Agriculture, Government of India is thankfully acknowledged.

References

- [1] DB Goldstein *et al.*, *Natl Acad Sci U.S.A.* **92**: 6723 (1995) [PMID:7624310]
- [2] O Rose, D Falush, *Mol Biol Evol* **15**: 613 (1998) [PMID: 9580993]
- [3] NN Fitzsimmons *et al.*, *Mol Biol Evol* **12**: 432 (1995) [PMID: 7739385]
- [4] D Falush, Y Iwasa, *Mol Biol Evol* **16**: 960 (1999)
- [5] P Rajendrakumar *et al.*, *In Silico Biology* **8**: 87 (2008) [PMID:18928198]
- [6] S Temnykh *et al.*, *Genome Res.* **11**: 1441 (2001) [PMID: 11483586]
- [7] JA Castro *et al.*, *In Microbial.* **1**: 327 (1998) [PMID: 10943382]
- [8] S Rozen, HJ Skaletsky, *Methods in Molecular Biology* (2000) 365.
- [9] JM Hancock *et al.*, *J Mol Evol.* **41**: 1038 (1995) [PMID: 8587102]
- [10] E Levdansky *et al.*, *Eukaryotic Cell* **6**: 1380 (2007) [PMID: 17557878]
- [11] DB Goldstein, DD Pollock, *Journal of heredity* **88**: 335 (1997) [PMID: 9378905]
- [12] GR Sutherland, RI Richards, *Proc Natl Acad Sci USA.* **92**: 3636 (1995) [PMCID: PMC42017]
- [13] ST Henderson, TD Petes, *Mol Cell Biol.* **12**: 2749 (1992) [PMID: 1588966]
- [14] JL Weber, C Wong, *Hum Mol Genet.* **2**: 1123 (1993) [PMID: 8401493]
- [15] G Levinson, G Gutman, *Mol Biol Evol.* **4**: 203 (1987) [PMID: 3328815]
- [16] C Schlotterer, D Tautz, *Nucl Acid Res* **20**: 211 (1992) [PMID: 1741246]
- [17] RI Richards, GR Sutherland, *Cell* **70**: 709 (1992) [PMID: 1516128]
- [18] JB Walsh *et al.*, *Genetics* **115**: 553 (1987) [PMID: 3569882]
- [19] H Tachida, M Iizuka, *Genetics* **131**: 471 (1992) [PMID: 1644281]
- [20] LW James, W Carmen, *Human Molecular Genetics* **2**: 1123 (1993) [PMID: 8401493]

Edited by P. Kanguane

Citation: Shakyawar *et al.*, *Bioinformatics* 4(4): 158-163 (2009)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Supplementary material:

Table 1: Repeat motif data in coding and non-coding region

Repeat	Sheep		Goat		Cattle		Buffalo		Yak	
	C	N	C	N	C	N	C	N	C	N
Mono	0	19	12	27	0	14	5	12	5	7
Di	750	386	754	396	742	350	704	361	730	377
Tri	424	165	413	179	436	151	423	174	430	175
Tetra	78	77	70	64	76	50	77	58	78	59
Penta	10	14	30	20	30	16	18	24	22	16
Hexa	16	8	10	14	12	12	2	6	6	14

Table 2: Specific primers for finding conserved region

Species	Forward primer (5' to 3')	Reverse primer (5' to 3')	AS (bp)	T ^A
Buffalo	tgggattatcgtagtgcgtgat	tgcttagggcttgaaggctcttg	250	60.17 (First set of primer)
	ggcattcggatcggatgctta	gggaagatattaggtgggatcg	209	60.32
Cattle	tgggattatcgtagtgcgtgat	tgcttagggcttgaaggctcttg	201	60.17
	ggcattcggatcggatgctta	gggaagatattaggtgggatcg	203	60.32
Goat	tgggattatcgtagtgcgtgat	tgcttagggcttgaaggctcttg	245	60.17
	ggcattcggatcggatgctta	gggaagatattaggtgggatcg	223	60.32
Sheep	tgggattatcgtagtgcgtgat	tgcttagggcttgaaggctcttg	194	60.17
	ggcattcggatcggatgctta	gggaagatattaggtgggatcg	202	60.32
Yak	tgggattatcgtagtgcgtgat	tgcttagggcttgaaggctcttg	261	60.17
	ggcattcggatcggatgctta	gggaagatattaggtgggatcg	210	60.32 (Second set of primer)
First set of primer	Self dimer -3.6 (-42.5)	Cross dimer -1.6 (-42.5)		Hairpin ±0.5
	-3.5 (-47.6)	-1.6 (-42.5)		±0.7
Second set of primer	-6.6 (-43.5)	-4.6 (-43.5)		±1.4
	-4.6 (-43.1)	-4.6 (-43.5)		±1.3

T^A = Annealing temperature; AS = Allele size; Delta G value = calculated value (Worst case) for forward primer calculated value (Worst case) for reverse primer

Table 3: Date and place of origin of five domestic species

Species	Date	Place of origin
Buffalo	4000 BC	India, China
Cattle	8000 BC [17][18]	India, middle east, subsahara Africa
Goat	10000 BC [16]	Iran
Sheep	9000-11000 BC [19][20]	South west Asia
Yak	2500 BC	Tibet