INTRODUCTION
1. INTRODUCTION AND OBJECTIVES

The eukaryotic genome contains a predominant portion (~55%) of
different repetitive sequences and a small (2-3%) portion of mature
transcripts (Ugarkovic, 1995; Bennett, 2000; Jasinska et al., 2004). Repetitive sequences are dynamic components of the genome
encompassing major satellites and simple sequence repeats (SSRs)
comprising minisatellites and microsatellites (Charlesworth et al., 1994;
Jeffereys et al., 1998). The highly polymorphic and multi-allelic SSRs
(Tautz, 1989) containing short tandem iterations, are potentially involved in
genome evolution by creating and maintaining genetic variability (Bennett,
2000; Toth et al., 2000; Verstrepen et al., 2005). Most of these SSRs are
found in non-coding genome whereas a small fraction is retained in the
transcriptome (Bennett, 2000; Jasinska et al., 2004) which participate in
gene regulation through transcription, translation, slipped strand mispairing
or gene silencing (Rocha et al., 2002; Li et al., 2004). Repeat sequences
are known to shrink and expand fuelling the process of copy number
alteration (John and Ali, 1997; Nakamura et al., 1987) and have been
associated with tumorigenesis and several other genetic anomalies
(Epplen, 1988; Kizawa et al., 2005; Ross et al., 2005). The expansion and
contraction of the SSRs within the protein-coding sequences is proposed
to modulate disease risks such as Huntington’s disease, Myotonic
dystrophy and fragile X Syndrome (Sutherland and Richards, 1995;
Richards, 2001; Borstnik and Pumpernik, 2002; Di Prospero and
Fischbeck, 2005; Dushlaine et al., 2005). Majority of these SSRs are
evolutionarily conserved (Robles et al., 2004; Tautz, 1989) whereas others
remain unique to a given genome (Ali et al., 1999). Their evolutionary
conservation and polymorphism within and across the
tissues/sex/stage/species substantiates their vital regulatory roles in higher
eukaryotes (Tautz, 1989; Ali et al., 1999; Robles et al., 2004). However,
the distribution and significance of SSRs within the non-coding and coding
genomes, even in the best characterized organisms including human,
remains unclear.
To explore the organization and expression of such repeat-tagged genes, we targeted the transcriptome of water buffalo *Bubalus bubalis* as a model system. Buffalo is an important animal in agriculture, dairy and meat industries in the Indian sub-continent as India has about half of the world's buffalo population. Compared to its importance, still no information is available yet on the genetic makeup of this important livestock animal. Novelty also lied in the fact that buffalo genome is unexplored in terms of genes present and its association with the major satellites or SSRs.

Simple consensus repeat of 16 nucleotide long (5' CACCTCTCCACCTGCC 3') of 33.15 repeat loci originating from the human myoglobin gene have been studied in a number of species (Ali and Wallace, 1988; Jeffreys *et al.*, 1985; Weitzel *et al.*, 1988). The repeats of GACA and GATA sequences were identified from the Banded krait minor (*Bkm*) satellite DNA in snakes (ZW) and found to be conserved in a number of species including human, with their highest frequency on the sex chromosomes of various eukaryotes (Epplen *et al.*, 1982; Singh and Jones, 1982; Singh *et al.*, 1980; Hobza *et al.*, 2006). High condensation of these repeats in somatic cells and decondensation in germ cells during early ages of development with sex-/tissue-specific expression in higher eukaryotes supported their crucial role in sex differentiation (Singh and Jones, 1982; Singh *et al.*, 1994; Subramanian *et al.*, 2003). However, the distribution of such important repeats in the non-coding genomes, and their organization within the mRNA transcripts originating from somatic/gonadal tissue and spermatozoa remain largely unclear.

Ejaculate spermatozoa are terminally differentiated cells in which transcription and/or translation of nuclear encoded mRNAs are unlikely. Therefore, until recently, the male genome was considered to be the only cargo carried by the spermatozoa. The discovery of many soluble signaling molecules, transcription factors and structures such as centriole being introduced by spermatozoa into the zygotic cytoplasm upon fertilization has changed this perception (Saunders *et al.*, 2002; Krawetz, 2005; Miller *et al.*, 2005). Despite the transcriptionally dormant state, spermatozoa retain an entourage of transcripts, encoding transcription factors and proteins involved in signal transduction, cell proliferation, DNA
Introduction

condensation, regulation of sperm motility, capacitation and acrosome reaction (Wykes et al., 1997; Miller, 2000; Lambard et al., 2004; Krawetz, 2005; Miller et al., 2005; Ostermeier et al., 2005). The delivery of such spermatozoal transcripts to ooplasm entails their potential significance during fertilization, embryogenesis and morphogenesis.

Owing to the tissue- and sex-specific organization of GACA, GATA and 33.15 repeats and role of spermatozoal RNA in and post syngamy, the transcriptome fraction tagged with these repeats in somatic/gonadal tissues and spermatozoa of water buffalo Bubalus bubalis using Minisatellite/Microsatellite Associated Sequence Amplification (MASA) was studied. The transcripts so uncovered were characterized in detail. Moreover, the detailed isolation and characterization of the candidate full length genes; Secreted modular calcium binding protein-1 (Smoc-1) and Protooncogene c-kit receptor (c-kit) were also performed from buffalo Bubalus bubalis. The characterization included domain organization, copy number status, in silico structural and functional analysis, in-vitro protein expression & purification, tissue & age specific transcription/translation and localization of the same onto the metaphase chromosomes & basement membrane zone. Smoc-1 belongs to the BM-40 family which has been implicated with tissue remodeling, angiogenesis and bone mineralization whereas c-kit is implicated with spermatogenesis, melanogenesis and hematopoeisis. Besides their anticipated roles in such important pathways, Smoc-1 and c-kit has been characterized only in a few mammalian species. We took advantage of their association with the repeats to characterize these genes. However, the brief objectives of present thesis included:

1. In silico analysis to explore the distribution of GACA, GATA and 33.15 repeats within the non-coding and coding genomes across the species.

2. Identification and cloning of the satellite tagged transcripts using different consensus repeat motifs and random primers following MASA with the cDNA from somatic tissues, gonads and spermatozoa in water buffalo Bubalus bubalis.
3. Sequencing and computational analysis of the MASA uncovered sequences to assess their homology status across the species, evolutionary studies and sequence organization in different tissues, if any.

4. Assessment of germline modulation of the MASA uncovered mRNA transcripts.

5. Tissue and stage specific expression for individual MASA uncovered genes/gene fragments using RNA slot blot hybridization, RT-PCR and Real Time PCR analysis.

6. Copy number calculation of all the MASA entrapped genes using SYBR green assays and Real Time PCR, and Chromosomal localization of candidate genes using Fluorescence in Situ Hybridization (FISH).

7. Isolation and detailed characterization of candidate genes and their *in vitro* expression studies.