REVIEW OF LITERATURE

3.1 BUBALUS BUBALIS DOMESTICATION

The buffalo (Bubalus bubalis) population in the world is actually about 168 million head, of which 161 million can be found in Asia (95.83%); 3717 million are in Africa and Egypt (2.24%); 3.3 million (1.96%) in South America, 40,000 in Australia (0.02%); 500,000 in Europe (0.30%).

Asian buffalo or Water buffalo is classified under the Genus: Bubalus, Species: bubalis. The Bubalus bubalis belongs to the Class Mammalia, Subclass: Ungulata, Order: Artiodactyla, Suborder: Ruminantia, Family Bovidae, Subfamily Bovinae, Tribe Bovini, which includes the following three groups: Bovina (cattle), Bubalina and Syncerina. Syncerina includes only the species Syncerus caffer (the African buffalo). Bubalina (the Asian buffalo) includes three species: Bubalus depressicornis or Anoa which lives in Indonesia, Bubalus mindorensis which lives in the Philippines and Bubalus bubalis deriving from the domestication of the Bubalus arnee, the Indian wild buffalo. The domestication of this species occurred relatively recently (5000 years ago) compared to the domestication of Bos taurus and Bos indicus (10 000 years ago). Asian buffalo includes two subspecies known as the River and Swamp types, the morphology and purposes of which are different as are the genetics. The River buffalo has 50 chromosomes of which five pairs are sub-metacentric, while 20 are acrocentric: the Swamp buffalo has 48 chromosomes, of which 19 pairs are metacentric. The difference in the diploid number is only apparent. In fact, the large Swamp buffalo chromosome 1 originated from tandem fusion translocation between the River buffalo chromosome 4 (telomeres of p-arm) and 9 (centromere) (Di Berardino and Iannuzzi, 1981). During this phenomenon, the nucleolus organizer regions (NORs) present in the River buffalo chromosome 4p were lost and the centromere of chromosome 9 inactivated (Di Berardino and Iannuzzi, 1981). The two subspecies are inter-fertile and produce progeny with 49 chromosomes. Male crossbred progeny have sometimes displayed fertility problems while female progeny have manifested longer calving intervals only in the case of further backcross. Morphology of the two types differs considerably. Swamp buffaloes are less heavy, the adult male weight ranging between 325 and 450 kg, while the River type weighs between 450 and 1000 kg. Swamp buffaloes are stocky animals with marshy land habitats. They are primarily used for draught power in
paddy fields and haulage but are also used for meat and milk production. They produce a valuable milk yield of up to 600 kg milk per year, Swamp buffaloes are mostly found in South East Asian countries. A few animals can also be found in the northeastern states of India (Sethi, 2003). River buffaloes are generally large in size, with curled horns and are mainly found in India, Pakistan and in some countries of western Asia. They prefer to enter clear water, and are used for milk, meat production and draught purpose. Each subspecies includes many breeds. Buffaloes are known to be better at converting poor-quality roughage into milk and meat. They are reported to have a 5 % higher digestibility of crude fiber than high-yielding cows; and a 4-5 % higher efficiency of utilization of metabolic energy for milk production (Mudgal, 1988).

India has about 97 million animals, which represents 92% of the world buffalo population. India is the first country in the world for number of buffaloes and milk production (about 134 million tons). India is also the first country in Asia for scientific and technological development in buffalo nutrition, production, reproduction, biotechnologies and genetic improvement. Moreover India has implemented national programmes such as the "green revolution" (to increase crop production for animals), the "white revolution" (to increase milk productivity and satisfy human needs for proteins) and finally the "red revolution" (to increase meat production and strengthen the meat industry), particularly with regard to buffalo. India possesses the best River milk breeds in Asia e.g. Murrah, Nili-Ravi, Surti, Jaffarabadi, Mehsana, Kundi, Bhadavari and Nagpuri which originated from the north-western states of India and have a high potential for milk and fat production apart from their use as a work animal and as a supplementary stock for use as meat production (Sethi, 2003). Indian Murrah is the most diffuse breed in the world: from Bulgaria to South America and all over Asia. Trinidad imported several breeds from India between 1905 and 1908. Crossbreeding of these animals has produced a Trinidadian type and these animals are the stock that has been imported into the US. The low reproductive efficiency in female buffalo can be attributed to delayed puberty, higher age at calving, long postpartum anoestrus period, long calving interval, lack of overt sign of heat, and low conception rate. In addition, female buffaloes have few primordial follicles and a high rate of follicular atresia. Understanding potential quantitative trait loci associated with economically important traits will help in producing genetically superior breeds.
3.2 REPETITIVE DNA SEQUENCES

A predominant part of eukaryotic genome consists of repetitive sequences of various types (Anna and Krzyzosiak, 2004). These are dynamic elements, which reshape their host’s genome by generating rearrangements, shuffling of genes and modulating pattern of expression. The dynamic nature of repeats leads to evolutionary divergence that can be used in species identification, phylogenetic inference and for studying process of mutation and selection. These repetitive sequences are mainly composed of interspersed transposons-derived repeats and tandem repeats (Slamovits and Rossi, 2002). The later includes satellite, minisatellites and microsatellites. Satellite DNAs, predominantly associated with centromeric heterochromatin is being increasingly utilized versatile tool for genome analysis, genetic mapping and for understanding chromosomal organization. On the other hand minisatellite and microsatellites dispersed throughout the genome, are highly polymorphic in all populations studied led to their extensive use as genetic markers for fingerprinting, genotyping, and for forensic analysis. On the basis of number of copies of a specific sequence, Repetitive DNA sequences are classed into two types:

- **Highly Repetitive Sequences** are short sequences (5 to 10 bp), amounting 10% of the genome is repeated large number of times, usually occurring as tandem repeats (present in approximately $10^6$ copies per haploid genome). They are not interspersed with different non-repetitive sequences. The sequence of each repeating unit is highly conserved. Most of the sequences in this class are located in heterochromatin regions of centromere or telomere of metaphase chromosomes. Highly repetitive sequences are involved in providing sites for proteins involved in “organizing chromosome pairing” during meiosis and recombination.

- **Moderately or Dispersed Repetitive Sequences** are short sequences (150 to 300 bp), amounting up to 40% of the genome or long (5 kb), amounting 1-2% of the genome, is found dispersed throughout the euchromatin (present in $10^2$-$10^5$ copies per haploid genome). These sequences are involved in the regulation of gene expression. In some case sequence of long dispersed repeats show homology with retro viruses, in particular flanking 300 to 600 bp direct repeats show similarity to the long terminal repeats (LTRs) seen in integrated retroviruses.
Fig. 1. Schematic diagram showing different categories of repetitive sequences.
3.3 REPETITIVE DNA ORGANIZATION

On the basis of their mode of amplification, two main organizations of repetitive DNA sequences, tandemly repeated and interspersed DNA has been reported (Slamovits and Rossi, 2002). Fig.1 gives brief detail of types of repetitive sequences.

3.3.1 Interspersed Repetitive DNA

Interspersed repeat sequences scattered throughout the genome have arisen by transposition, which is “ability to jump from one place to another in the genome” (Brown 2002). Even though the individual units of interspersed repetitive non-coding DNA are not clustered, taken together they account for approximately 45% of the human genome. By the mechanism of their transposition two main classes of interspersed repeats are classified:

3.3.1.1 RNA Transposons

RNA transposons also known as retroelements, found in eukaryotic genome require reverse transcription for their activity. Based on their structural relationship there are two general categories:

- **LTR elements, which includes:**
  - **Retroviruses** are viruses whose genomes are made of RNA. They infect many types of vertebrates.
  - **Endogenous Retroviruses (ERV)**s are retroviral genomes integrated into vertebrate chromosomes. Some are still active and might, at some stage in a cell’s lifetime, direct synthesis of exogenous viruses, but most are decayed relics that no longer have the capacity to form viruses (Patience et al., 1997). These inactive sequences are genome-wide repeats but they are not capable of additional proliferation.
  - **Retrotransposons** is the biggest class of transposons. An important characteristic of this type of transposable element is that they usually contain sequences with potential regulatory activity. They have features of non-vertebrate eukaryotic genomes (i.e. plants, fungi, invertebrates and microbial eukaryotes). These sequences allow the element to code for an mRNA molecule that is processed and polyadenylated. Retrotransposons have very high copy numbers in some genomes; for instance, they almost make up half of the maize genome. There are two types of retrotransposons:
Ty3/gypsy Family – repeats that are found in yeast and fruit fly respectively, which possess the env gene and can be looked upon as non-vertebrate retroviruses (Song et al., 1994; Peterson-Burch et al., 2000).

Ty1/copia Family – repeats that do not possess the env gene, therefore would not form infectious virus particles.

Non LTR Elements include:

- **LINEs (Long Interspersed Nuclear Elements)**
  LINEs are several thousand base pairs in size and make up 21% of the human genome (Pierce, 2005). They contain reverse-transcriptase-like gene (e.g. the pol gene) involved in retrotransposition process. The most abundant LINE family is the 7 kb, L1 repeat element (Furano, 2000) that has a copy number of >500,000 and accounts for approximately 15% of the human genome (Lander et al., 2001). Despite of its abundance LINE 1 repeat has no known function.

- **SINEs (Short Interspersed Nuclear Elements)**
  SINEs are small elements, usually 100 to 500 bp in length, accounting for 13% of the human genome (Pierce, 2005). SINEs do not have reverse transcriptase gene, instead they borrow reverse transcriptase enzymes from other retroelements. Well-known example of SINE in the human genome is Alu sequences (Capy et al., 1998), which has a copy of over 1 million. The Alu sequence is 300 bp long and occurs on average once every 3300 bps in human genome. They occur throughout the primate family and are descended from small, abundant RNA gene that codes for 300 nucleotide long RNA molecules known as 7SL. The 7SL RNA combines with six proteins to form a protein-RNA complex that recognizes the signal sequences of newly synthesized proteins and aids in their translocation through the membranes of endoplasmic reticulum to their ultimate destination in cell.

### 3.3.1.2 DNA Transposons

DNA transposons do not require RNA intermediate and transpose in a direct DNA-to-DNA manner. In eukaryotes, DNA transposons are less common than retrotransposons, but they have a special place in genetics because a family of plant DNA transposons - the Ac/Ds elements of maize. There are two types of DNA transposons that both require enzymes coded by genes within the transposon:
- **Replicative Transposons** – transposons involve direct interaction between the donor transposon and the target site, resulting in copying of the donor.

- **Conservative Transposons** – transposons that transposed by excision of the donor element and reintegration at a new site.

Furthermore, there are other kinds of DNA transposons known in *E.coli*, and fairly typical as in prokaryotes in general. They are composite transposons, Tn3-type transposons, and transposable phages (Brown, 2002).

### 3.3.2 TANDEM REPEATS

Tandem repeats consists of repeat arrays of two to several thousand-sequence units arranged in a head to tail fashion. Tandem repeats may be further classified according to the length and copy number of the basic repeat units as well as its genomic localization.

#### 3.3.2.1 Mega Satellite DNA

Mega Satellite DNA, are characterized by tandemly repeated DNA in which the repeat unit is approximately 50-400 times, producing blocks that can be hundreds of kilobases long. Some mega satellites are composed of coding repeats. For example: RNA genes, and the deubiquitinating enzyme gene USP17.

#### 3.3.2.2 Satellite DNA

Satellite DNA, are represented by monomer sequences, usually less than 2000 bp long, tandemly reiterated up to $10^5$ copies per haploid animals and located in the pericentromeric and/or telomeric heterochromatic regions (Charlesworth *et al.*, 1994). Satellite DNA constitutes from 1 to 65% of the total DNA of numerous organisms, including that of animals, plants, and prokaryotes. The term “satellite” in the genetic sense was first coined by the Russian cytologist Sergius Navashin in 1912, initially in Russian (“sputnik”) and Latin (*satelle*), and only later translated to “satellite” (Battaglia, 1999). This original usage referred to the morphology of a chromosome possessing a secondary constriction at a certain point along its length. The more familiar usage of "satellite" relates to a small band of DNA with a density different (usually lower, because of a high AT-content) from the bulk of the genomic DNA, and which becomes separated from the main band following CsCl centrifugation (Kit, 1961; Sueoka, 1961). Nucleotide changes and copy number variations fuel the process of their evolution within and across the species (Ugarkovic and Plohl, 2002). Satellite fraction(s), though not conserved evolutionarily (Amor and
Fig. 2. Representative diagrammatic illustration showing the most abundant classes of repetitive sequences with their approximate locations on the human chromosome 16.
Choo, 2002), are unique to a species and usually show similarity amongst related group of animals (Ali and Gangadharan, 2000; Henikoff et al., 2001).

- **Types of Satellite DNA**

  Satellite DNA constitutes important structural elements of heterochromatin region of centromeric, major pericentromeric and telomeric regions of eukaryotic chromosomes (Fig.2). The satellite DNA types are:

  - **Satellites I-IV:**
    Satellites I-IV were originally isolated by gradient centrifugation (Singer 1982), Satellite I is rich in As and Ts and is composed of alternating arrays of a 17- and 25 bp repeating unit. Satellite two and three are both derived from the simple five base repeating unit ATTCC. The tandemly reiterating arrays of satellite II and III are based on variants of the 5 bp GGAAT monomeric units (Frommer et al., 1982; Prosser et al., 1986). Satellite IV, higher-order repeat substructure exists in all species, with multimeric arrays ranging in size from 10 to 1500 kb. Satellite I, II, III occur as long tandem arrays in the heterochromatic regions of chromosomes 1, 9, 16, 17, and Y (Doggett, 2001) and the satellite regions on the short (p) arms of chromosomes 13, 14, 15, 21, and 22.

  - **Alpha Satellite Repeats:**
    A family of related species that occur as long tandem arrays, arranged in head-to-tail configuration (Manuelidis and Wu, 1978; Willard and Waye, 1987) at the centromeric region of metaphase chromosomes. The repeat unit is about 340 bp and is a dimer, i.e., it consists of two subunits, each about 171 bp long. The amount of alpha satellite DNA in different centromeres varies from approximately 250 kb to > 40 M bps. There are two major types of alpha satellite DNA:

    - **Higher-order Repeat**
      Higher-order repeat, consists of several monomeric repeat that are amplified as a unit, with the multimeric unit being arranged in a tandem head-to-tail configuration. The higher-order repeats are highly homogenous and are typically 97-100% identical. Although previously identified at all human
centromeres, higher-order alpha satellite has only been included in the assemblies of eleven chromosomes. There are several lines of evidences indicating that the higher-order alpha satellite DNA, not the monomeric alpha satellite DNA is associated with the functional centromeres (Schueler et al., 2001; Spence et al., 2002).

✓ Monomeric Repeat
Monomeric alpha satellites lack any higher-order periodicity and are on average approximately 70% identical. Monomeric alpha satellite typically lies at the edges of larger higher-order arrays, and has been included in all but three chromosome assemblies. Monomeric alpha satellite has a higher frequency of intra-chromosomal exchange than inter-chromosomal exchange. However, comparing orthologous regions of human and chimpanzee alpha satellite, monomeric alpha satellite is more conserved than higher-order alpha satellite.

• β-Satellite Repeats (BSR)
Beta satellite repeats, represent another family of sequences that show a predominant heterochromatic distribution, which includes pericentric regions of chromosomes 1, 13, 14, 15, 21, and 22 and Y (Waye and Willard, 1989; Agresti et al., 1987, 1989). They are GC rich with monomeric unit of 68 bp repeat, present in 30,000 - 60,000 copies. In addition, it is also found on long arm of chromosome 9 as well as distal cytological satellites of 5 acrocentric chromosomes lq12, 3q12, 9q12, and Yq11.

• γ-Satellite Repeats (GSR)
Gamma satellite repeats are newest family of repeats reported at the centromeres of human chromosomes 8 and X. Gamma-X satellite DNA is approximately 220 bp tandemly arranged repetitive DNA that comprise approximately 0.015% of each of the two chromosomes. A related repeat DNA has also been identified on long arm side of the major DYZ3 alpha-Y domain, outside the region previously defined as that required for mitotic centromere function of the human Y chromosome (Lee et al., 2000).
Review of literature

- **Telomeric Repeats**
  Telomeres are composed of multiple repeats of short sequence elements (typically 5 to 8 bp in length, with a GT-rich strand oriented 5' to 3' toward the end of the chromosome) and range in length from a few repeat units to >10 kb. Long simple sequence tandem repeats of interstitial TTAGGG arrays form a three-dimensional nuclear network of poorly transcribed domains, which involve gene silencing by repositioning. This network, as well as clusters of retroelements properly positioned in the nucleus, form unique lineage-specific structures that affect gene expression (Tomilin, 2008). The repeated sequence (TTAGGG)$_n$ is found at telomeres in all vertebrates, certain slime molds, and trypanosomes; (TTGGGG)$_n$ and (TTTGGGG)$_n$ are found in the ciliated protozoan *Tetrahymena* and *Oxytricha* species, respectively; and (TG$_1$)$_n$ is found in the yeast *Saccharomyces cerevisiae*. In organisms whose telomeres have been examined in detail, the GT strand extends 12 to 16 nucleotides (two repeats) beyond the complementary C-rich strand. The unique structure of telomere provides the mechanism for maintenance of the integrity of the chromosome ends.

- **Subtelomeric Repeats**
  Classes of repetitive sequences that are interspersed in the last 500,000 bases of non-repetitive DNA located adjacent to the telomere. Some sequences are chromosome specific and others seem to be present near the ends of all human chromosomes.

- **Satellite DNA Evolution and Centromere**
  Satellite DNA in pericentromeric heterochromatin as well as in genome in general, represents rapidly evolving components. Consequently, even among the most closely related species, they differ in nucleotide sequence, copy number, and/or composition of satellite families (Schmidt and Heslop-Harrison, 1998; Ugarković and Plohl, 2002). Rapid evolution of satellite DNA sequences is possible owing to the accumulation of nucleotide divergences, usually with a high rate and in a gradual manner (Bachmann and Sperlich, 1993). Gradual accumulation of mutations follows phylogeny at different hierarchical ranks. At the species level, centromeric satellites
DNAs were informative in phylogenetic studies of the *Drosophila obscura* group (Bachmann and Sperlich, 1993), or in the study of the fish family *Sparidae* (Garrido-Ramos *et al.*, 1999). Determination of ecotype-specific variants in the Arabidopsis thaliana 180 bp satellite indicated accumulation of divergences within the last ~5 Myr (Hall *et al.*, 2003; Ito *et al.*, 2007). Even within a genome, distinct forms of satellite DNAs can accumulate mutations with different rates, adding to the diversity of sequence patterns in pericentromeric areas. Alpha-satellite repeats, for example, occur as monomeric and higher-order units; these two distinct forms accumulate mutations with different evolutionary rates. Interestingly, centromerically located higher-order units diverge more rapidly than pericentromerically located monomeric repeats (Rudd *et al.*, 2006). Accumulation of mutations in satellite families is not the only way to alter specific profiles of satellite repeats in short evolutionary periods. Since more than one satellite family exists in a genome, expansions and contractions of satellite arrays can efficiently change a landscape of DNA sequences in heterochromatin by replacing one dominant (major) satellite repeat with another one less represented (Ugarković and Plohl, 2002). In this, unequal crossover is proposed to be the major mechanism responsible for dramatic fluctuations in the copy number of satellite DNAs (Smith, 1976). The occurrence of species-specific profiles as a consequence of copy number changes in a set of satellite DNAs shared by related genomes was originally explained through the library model (Fry and Salser, 1977), and experimentally verified in the study of satellite DNAs shared by species of the insect genus *Palorus* (Meštrović *et al.*, 1998). Copy number changes may be, but are not necessarily, accompanied by rapid evolution of nucleotide sequences, and can explain species-specificity of satellite profiles even when satellite sequences remain “frozen” during long evolutionary periods (Meštrović *et al.*, 1998; Mravinac *et al.*, 2002; Bruvo *et al.*, 2003). Not only distinct satellite DNAs, but also monomer variants from a single family can be distributed in genomes in the form of a library (Cesari *et al.*, 2003). In the constitution of a library, besides stochastic events, selection might represent a limiting factor for persistence of particular satellite sequences, as indicated by the study of intersatellite variability in a set of related repeats differentially amplified in a group of taxa (Meštrović *et al.*, 2006). In addition to nucleotide changes and expansions-contractions of satellite arrays, large-scale changes, such as segmental duplications, play an important role in the rapid evolution of DNA sequences in and around centromeric regions (Cardone *et al.*, 2004; Hall *et al.*, 2004;
Review of literature

Ventura et al., 2007). Satellite repeats may be the preferred form of DNA sequences in functional centromeres and their flanking regions just because of their unique characteristic to maintain sequence homogeneity over long stretches of DNA, and simultaneously to change rapidly in evolution, as explained in the above paragraphs. Concerted evolution and abundance of satellite repeats may stabilize interactions with DNA-binding proteins and eliminate effects of possible unwanted mutations, and at the same time the whole array can rapidly adopt to a new sequence variant which can better fit the mentioned interactions. In a recent model proposed by Dawe and Henikoff, 2006, centromeres are defined in a dual way, through an interplay of epigenetic factors and through interactions between “selfish” repetitive DNA sequences and protein components in a centromere, mediated by meiotic drive. Rapid evolution of satellite DNA sequences without impairment of centromeric function can be explained assuming that DNA binding centromeric proteins such as CENH3 and CENP-C co-evolve with satellite DNAs. This coevolution may be driven either by changes in satellite DNAs (Malik and Henikoff, 2001; Talbert et al., 2004), or by satellite repeats competition to better fit the chromatin environment (Dawe and Henikoff, 2006). In other words, both DNA and protein evolution drive each other in a centromere, thus providing a stable, but flexible system, able to work on genetic and epigenetic platforms and, if necessary, to rescue chromosomal function by forming new centromeres on non-specialized locations (Dawe and Henikoff, 2006). In agreement with the library model, it can be additionally proposed that one satellite family can replace another one if their sequences are of similar functional value, as discussed recently by Meštrović et al., 2006. In an ideal case, rapid replacements of equivalent DNA sequences would be possible without alterations in binding affinities. However, it can be reasonably expected that the evolution of DNA and DNA-binding centromeric proteins is an integrative result of all these processes in a particular organism. Whatever the scenario could be, a proposed direct consequence is that divergences in satellite sequences and corresponding proteins accumulated between individuals can cause incompatibilities in hybrids and eventually lead to reproductive isolation acting thus as a trigger in speciation process (Meštrović et al., 1998; Henikoff et al., 2001; Hall et al., 2005).
Satellite DNA and Transcription

The general significance of satellite DNA is rather controversial and several hypotheses have been developed to assign a role to this fraction of the eukaryotic genome. It has been suggested that it may play a role in the heterochromatin constitution, chromosome pairing and recombination, chromosome rearrangements, three-dimensional organization of the interphase nucleus, and in gene amplification events (Brutlag, 1980; Manuelidis, 1982; Strissel et al., 1996). Its role has also been suggested in centromere functions by their ability to associate with centromeric satellite-DNA binding proteins (CENP) (Marshall and Clarke, 1995; Sugimoto et al., 1994).

The transcription of satellite DNA has been described in vertebrates, invertebrates and in plants. In most species, satellite DNA is temporally transcribed at particular developmental stages or is differentially expressed in some cell types, tissues or organs. In chick and zebrafish, two types of transcript are identified; one that corresponds to α-repeat RNA and another group of mRNAs that contain an α-like satellite sequence in the 5’ and 3’ untranslated regions. The transcripts of an α-like satellite repeat detected during early embryogenesis were limited to the cardiac neural crest, head and the heart (Li and Kirby, 2003). Mouse γ-satellite DNA is differentially expressed during development of the central nervous system, as well as in the adult liver and testis (Rudert et al., 1995). In the crab Geocarcinus lateralis, a GC-rich satellite DNA is transcribed in a tissue and stage specific manner (Varadaraj and Skinner, 1994). Sam et al., 1996, have described an expressed repetitive sequence from the mouse genome, the expression of which is restricted during development and is induced in response to retinoic acid. The developmental, stage and tissue-specific expression of satellite DNAs in several species suggest that they have a regulatory role(s), although for most transcripts this role is still elusive and hypothetical.

Satellite DNA or its transcripts appear to be involved in heterochromatin formation and in chromatin-elimination processes. In Hymenoptera, satellite DNA transcriptions have been described in Diprion pini, Diadromus pulchellus, Diadromus collaris, Eupelmus vuiletti, and Eupelmus orientalis (Rouleux-Bonnin et al., 1996; Renault et al., 1999). The function of satellite DNA transcripts and their mechanism of transcription are not well known. It has been proposed that transcription of satellite DNA may be initiated from gene(s) or transposable element interdispersed in satellite DNA, as in lampbrush chromosomes (Hori et al., 1996; Solovei et al., 1996).
spermatocyte nuclei of *Drosophila melanogaster* exhibit three giant lambrush-like loops formed by the kl-5, kl-3 and ks-1 Y-chromosome fertility factors. These structures contain an abundantly transcribing highly repetitive simple sequence that accumulates large amounts of non-Y-encoded proteins (Pisano et al., 1993). Most of these transcripts are present as polyadenylated RNA in the cytoplasm but some are found exclusively in the nucleus, such as those associated with the Y chromosome of *Drosophila melanogaster* and *Drosophila hydei* (Trapitz et al., 1988; Bonaccorsi et al., 1990). Heat shock induces the transcription of a subset of satellite-III, which is located within pericentromeric heterochromatic regions of specific human chromosomes (Jolly et al., 2004; Rizzi et al., 2004). The satellite DNA transcripts from salamanders (Epstein and Gall, 1987), schistosomes (Ferbeyre et al., 1998) and crickets (Rojas et al., 2000) are expressed as long multimeric precursor RNAs that have the ability to adopt hammerhead-like secondary structures and can function as ribozymes with self-cleavage activity. Forced accumulation of 120-nt minor centromeric satellite transcripts in murine cells upon stress or differentiation, leads to defects in chromosome segregation and sister-chromatid cohesion, changes in hallmark centromeric epigenetic markers, and mislocalization of centromere-associated proteins that are essential for centromere function. These findings suggest that small centromeric RNAs may represent one of many pathways that regulate heterochromatin assembly in mammals, possibly through tethering of kinetochore and heterochromatin-associated proteins (Bouzinba-Segard et al., 2006). Transcripts of centromeric satellite DNA in maize were present within centromeric chromatin contribute to initiation and stabilization of kinetichore chromatin structure (Christopher et al., 2004). Satellite DNA transcripts have been proposed to play role in initiation of histone H3 methylation, a necessary preqisite in heterochromatin formation and maintenance (Usakin et al., 2007). In newt, snRNA-type promoter drives satellite DNA transcription and these RNAs are self-cleaving transcripts (Epstein and Coats, 1991; Cremisi et al., 1992; Green et al., 1993). This ribozymic activity has been demonstrated in several species of newt (Luzi et al., 1997; Denti et al., 2000). In higher vertebrates Dicer related RNA interference machinery is involved in the formation of the heterochromatin structure (Fukagawa et al., 2004). Recently, Saitoh et al., 2000 have identified a tandem-repetitive DNA sequence that encodes a novel deubiquitinating enzyme with a functional promoter designated as RS447 human megasatellite.
3.3.2.3 Minisatellite DNA

The repeat array typically comprises up to several hundred copies of a 6-100 bp sequence (Bois et al., 1998; Tautz, 1993), spanning from 0.5 to several kilobases in the eukaryotic genome (Vergnaud and Denoeud, 2000). The first minisatellite to be described was discovered in humans and was detected as specific loci (Wyman and white, 1980; Higgs et al., 1981; Bell et al., 1982). Later, Jeffrey and colleagues in 1985, found a minisatellite sequence in one of its non-coding (intron) regions of the human myoglobin gene. Since then similar DNA structures have been found in many organisms including bacteria (Skuce et al., 2002), avian (Reed et al., 1996), higher plants (Durward et al., 1995), protozoan (Bishop et al., 1998), and yeast (Haber and Louis, 1998) genomes. Comparison of the repeat units in classical minisatellites led to early notion of consensus or core sequences, which exhibit some similarities with the Chi sequences of λ phage (GCTGTGG). A minisatellite locus consists of multiple tandemly repeated copies of a DNA sequence, and is usually highly polymorphic due to variation in the number of repeats (Jeffreys et al., 1985). Minisatellites also known as Variable Number Tandem Repeat (VNTR) loci (Nakamura et al., 1988; Brown, 2002) are GC rich, with a strong strand asymmetry. Often minisatellites form families of related sequences that occur at many hundred loci in the nuclear genome. In human genome, number of minisatellite loci is estimated at approximately 3000, and each locus contains a distinctive repeat unit with respect to size and sequence content.

In humans, majority of minisatellites are clustered near sub-telomeric ends of chromosomes limiting their usefulness for extensive gene mapping (Royle et al., 1988), but there are examples of interstitial locations (alpha globin gene cluster (Proudfoot et al., 1982) and type II collagen gene (Stoker et al., 1985). Minisatellites of other species, such as mice or bovine (Georges et al., 1991; Julier et al., 1990), are not always preferentially clustered at chromosomal termini as in the human genome, but are distributed along the entire length of chromosomes. Unlike microsatellites, which usually alter during the DNA synthesis stage of the mitotic cell cycle, minisatellites alter during meiosis, undergoing changes in overall length and repeat composition (Jarman and Wells, 1989; Jeffreys et al., 1998). Minisatellite tracts have proven very useful for genomic mapping (Jeffrey’s et al., 1985) and linkage studies (Nakamura et al., 1987). Several probes have been identified which detect multiple
Table 1: Details of the sequences used as probe for hybridization

<table>
<thead>
<tr>
<th>Probe name</th>
<th>Sequence 5'-3'</th>
<th>Total length</th>
<th>Length repeat unit</th>
<th>No. repeat units</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-33.6-22</td>
<td>(CCTCCAGCCCT)_2</td>
<td>22</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>0-33.6-37</td>
<td>GCCCTTCCTCCGGAGCCCTCCTCCAGCCCTTCCTCCA</td>
<td>37</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>0-33.15-32</td>
<td>(CACCTCTCCACCTGCC)_2</td>
<td>32</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>0-33.15-80</td>
<td>(CACCTCTCCACCTGCC)_5</td>
<td>80</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>0-AY-29</td>
<td>GAGGARYAGAAAGGYGRGVRTGTTGCGGCG</td>
<td>29</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>0-YN-124</td>
<td>TCCTGAACACCCACTGTTACTTCCCA</td>
<td>27</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>0-33.1</td>
<td>GTGCTGCTTCCCTTCCCTCTCTTGTTC</td>
<td>27</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>0-34BHI</td>
<td>CCTGCTCCGCTCACGTGGCCCACGCAC</td>
<td>27</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>0-CCR-26</td>
<td>(CCR)_8CC</td>
<td>26</td>
<td>3</td>
<td>8.67</td>
</tr>
<tr>
<td>0-CCA-26</td>
<td>(CCA)_8CC</td>
<td>26</td>
<td>3</td>
<td>8.67</td>
</tr>
<tr>
<td>0-H-Ras</td>
<td>CACTCCCCCTTCTTCCAGGGGACGCCA</td>
<td>28</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>0-GACA-16</td>
<td>(GACA)_4</td>
<td>16</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0-GACA-24</td>
<td>(GACA)_6</td>
<td>24</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>
polymorphic loci in wide range of species (Burke and Bruford 1987; Bruford et al., 1992; Ruth and Fain 1993; Amos et al., 1991). List of minisatellite sequences widely used for fingerprinting studies are mentioned in Table 1 (Ali and Wallace, 1988).

3.3.2.4 Microsatellite / Short Sequence Repeats (SSRs)

The other well-known VNTR loci is the Microsatellites, or Simple Sequence Repeat loci (SSR), in which the repeat unit comprises fewer copies (usually 50-100) of much shorter di-, tri-, or tetranucleotide repeat unit (1-10 bp, usually 2-5 bp) (Tautz and Renz, 1984; 1989). Microsatellites are ubiquitously interspersed in coding and non-coding regions of eukaryotic and prokaryotic genomes (Gur-Arie et al., 2000; Toth et al., 2000; Katti and Gupta, 2001). All SSRs taken together occupy about 3% of the human genome in which they are widely dispersed and associated with many genes (Subramanian et al., 2003). The significance of microsatellite preference in different regions has not been completely understood. However, some microsatellites occurring in flanking regions of coding sequences are believed to play significant roles in regulation of gene expression by forming various DNA secondary structures and offering a mechanism of unwinding (Catasti et al., 1999). The variation of length and unit type of simple repeats in upstream activation sequences might influence transcriptional activity (Kim and Mullet, 1995; Epplen et al., 1996; Kashi et al., 1997; Martienssen and Colot, 2001), and affect interaction with different regulatory proteins during translation (Lue et al., 1989).

Microsatellites are usually characterized by low degree of repetition at a particular locus, but microsatellite containing identical motifs may be found at many thousand genomic loci. When the occurrence of SSRs in different functional genome regions is considered, it turns out that most of them show much higher density in non-coding regions. Exceptions to the rule are trimers and hexamers that are nearly two times more prevalent in exons compared to introns and intergenic regions. Their high frequency in coding regions may be explained by the fact that they do not change the reading frames and gene coding properties and, thus, are much better tolerated than other SSRs. Their positive selection in exons suggests also some function for the repeats. Examples well acknowledged in the humans are the trinucleotide repeat expansion and its association with a large number of genetic disorders, for example Fragile X chromosome (Handt et al., 2000), Huntington’s disease (Rubinsztein et al., 1995), Myotonic dystrophy (Winblad et al., 2006), Spinal-bulbar muscular atrophy
Review of literature

(Sinnreich et al. 2004), Spinocerebellar ataxia (Lindquist et al. 2006) and Epilepsy associated with EPMI gene (Lalioti et al. 1997). The high mutation rate of these repeats and their frequent length polymorphism have raised speculations that they may be involved in the regulation or “fine tuning” of gene expression and function, and have quantitative effects on phenotype.

- Few examples of repeat units used for fingerprinting and transcriptome analysis includes (GATA/GACA)n, CA, (AT)n, (GAA)n, (TCC)n, (GGAT)n, (GGCA)n, and (TTAGGG)n.

Minisatellite / Microsatellite Evolution

Initially, it was believed that the high rate of mutation in minisatellite was the product of unequal crossing-over between homologous chromosomes, but more recent evidence suggests that most mutations at these loci are intra-allelic and probably stem from sister chromatid exchange or replication slippage (Jeffreys et al., 1991). Furthermore, minisatellite variant repeat mapping and digital DNA fingerprinting of allelic variants at highly mutable human minisatellite VNTR loci has revealed a "gradient" of variability in monomer sequences across the tandem array. Jeffreys et al., 1991 suggested that this gradient of variability might be the result of mutation preferentially exerted at one end of the tandem array. An intriguing phenomenon reported by Rand 1992, is the high rate of point mutations, and small insertions and deletions, observed at one end of a VNTR in the mitochondrial DNA of crickets. This has been called repeat induced point (RIP) mutation or repeat associated polymorphism (RAP). Perhaps RIP or RAP is responsible for the gradient of variability seen across the tandem array of nuclear minisatellite. Regardless of the molecular mechanisms affecting mutation at the VNTRs, a continuum exists from the simple sequence repeats of microsatellites to the more complex sequences of minisatellites. What then is the evolutionary relationship between microsatellite and minisatellite VNTRs? Could microsatellites be the progenitors of minisatellites? Two hypotheses have been presented to account for the common "core" sequence identity of classes of minisatellite VNTRs. First, Jeffreys et al., 1985 proposed that minisatellite VNTRs of related core sequence are the result of transposition mediated by sequences flanking the minisatellite VNTR. For example, a 9 bp direct repeat was found flanking a hypervariable minisatellite array at the human myoglobin locus (Jeffreys et al., 1985). Additional reports have described the close association of
minisatellite VNTRs with dispersed repetitive elements such as the human Alu repeat (Armour et al., 1989) and transposon-like sequences (Rogaev, 1990). In this model, subsequent mutation at minisatellite VNTRs following transposition leads to sequence divergence that is carried to other repeat units of the tandem array by unequal exchange (Jeffreys et al., 1985). This "transposition" model, however, seems unlikely to account for the origin of all minisatellite VNTRs as many minisatellite VNTRs of related core sequence show no evidence of dispersed repeats flanking the tandem array (Jeffreys et al., 1985). The second hypothesis is that core sequences contain motifs that enhance the expansion of VNTRs independently at different loci (Jeffreys et al., 1985). In support of this "expansion hypothesis," (Levinson and Gutman, 1987) proposed that SSM is the most likely mechanism to explain the evolution of microsatellite VNTRs. The salient features of their model are as follows.

- The chance probability of the occurrence of short tracts of simple repeats in the genome is extremely high and serves as the raw material for expansion of simple sequence repeats by SSM.

- The finding of microsatellite arrays differing by a single base substitution is the consequence of several SSM events occurring before and after a single base change, the result of which is an augmented tandem repeat next to the progenitor sequence. For example, some microsatellites have been found to consist of mononucleotide repeats, adjacent to dinucleotide repeats, adjacent to tetranucleotide repeats, etc.

- Transitions will be more frequent than transversions resulting in the maintenance of polypyrimidine and polypurine tracts within a microsatellite array.

- The abundance of methylated C residues at CpG couplets (at least in vertebrate genomes) may favor the prevalence of TG repeats over GA repeats (Levinson and Gutman, 1987; Tautz et al., 1986). Since methylated C can undergo deamination, in a CpG repeat this would result in a transition of C to T and thereby increase the abundance of TpG couplets.

Several sequenced minisatellite VNTRs from different organisms (Jeffreys et al., 1985; Armour et al., 1992; Proudfoot et al., 1982; Bentzen and Wright, 1993) are, found juxtaposed to simple and cryptic microsatellites. For a bovine minisatellite VNTR, a complex 40 bp minisatellite tandem array is found interdispersed with a TG
repeat of variable length (Perret et al., 1990) further supporting the idea that microsatellite VNTRs may be progenitors of minisatellite VNTRs.

- **Minisatellite / SSRs and Transcription**

  Minisatellites are associated with chromosome fragile sites and are proximal to a number of recurrent translocation break points (Handt et al., 2000). Minisatellites / SSRs have also been implicated as regulators of gene expression, they are found to be present in transcribing regions of genomes, including protein-coding genes and expressed sequence tags (ESTs), although in general, repeat numbers and total lengths of SSRs in these regions are relatively small (Kantety et al., 2002; Thiel et al., 2003). For example, it has been found that ~12% of identified SSRs in Japanese pufferfish (Edwards et al., 1998), 10% in primate (Jurka and Pethiyagoda, 1995), 15% in rabbit (van Lith and van Zutphen 1996), and 9.1% and 10.6%, in pig and chicken, respectively (Moran 1993) are located in the protein-coding genes or open reading frames (ORFs). In cereals (maize, wheat, barley, sorghum, and rice) 1.5% to 7.5% of ESTs consist of SSRs (Kantety et al., 2002; Thiel et al., 2003). These ESTs have a range of functions such as metabolic enzymes, structural and storage proteins, disease signaling, and transcription factors, suggesting some role(s) of SSRs in plant metabolism and gene evolution. In protein-coding regions of all known proteins, 14% proved to contain repeated sequences, with a three times higher abundance of repeats in eukaryotes as in prokaryotes (Marcotte et al., 1999). Noteworthy, prokaryotic and eukaryotic repeat families are clustered to nonhomologous proteins.

  Tandem repeats residing within the genes involved in transcriptional regulation could lead to altered transcriptional activity (Yokota et al., 2003). The potential size expansion and contraction at 3’, 5’ regions and introns could lead to disruption of the original protein or formation of new genes by frame shift mutation. For example, Diabetes associated with the insulin gene, VNTR in insulin promoter, where different length of the repeat was associated with different types of diabetes (Eriksson et al., 2006; Bennett et al., 1995; Johnson et al., 1998) and cancer related to HRAS minisatellite locus (Krontiris et al., 1993; Jacqueline and Armour, 2003). A genome wide survey for tandem repeats located within the coding regions in Saccharomyces cerevisiae has thrown surprises that such intragenic repeats are mostly found within the stress induced and cell surface genes (Verstrepen et al., 2005; Bowen et al. 2005; Richard and Dujon 2006; Levdansky et al. 2007). Variation in FLO1 gene repeats in
yeast cells lead to gradual changes in the cell adhesion and their flocculation (Verstrepen et al., 2005). Similarly, in Canis familiaris, gene associated with tandem repeats functions as facilitators of evolution of limb and skull morphology (Caburet et al., 2005). The (TC)\textsuperscript{n} tract in promoter regions was found to serve as a transcriptional element for heat-shock protein gene hsp26 in Drosophila (Sandaltzopoulos et al., 1995), Aspergillus (Punt et al., 1990), Phytophthora (Chen and Roxby, 1997). Deletion of various di-, tri- and tetra SSR tracts markedly changed transcriptional activity. For instance, promoters transcriptional activity was dramatically decreased by deletion of a (TCCC)\textsuperscript{n} tract from the promoter regions of c-Ki-ras (Hoffman et al., 1990) and TGF-b3 gene in the CAT expression system (Lafyatis et al., 1991). Moreover, a (GT)\textsuperscript{n} repeat could enhance gene activity from a distance independent of its orientation, but more effective transcriptional enhancement resulted from the GT repeat being closer to the promoter sequences (Stallings et al., 1991). SSRs in intronic regions can also affect gene transcription. For example, a tetra-SSR HUMTH01 in the first intron of the tyrosine hydroxylase gene acts as a transcription regulatory element (Meloni et al., 1998). Gebhardt et al., 1999, found that transcription activity was affected by a (CA)\textsuperscript{n} tract located in the first intron of the epidermal growth factor receptor (EGFR) gene. They also showed that RNA elongation terminates at a site closely downstream of the SSR and that there are two separate major transcription start sites. Model calculations for the helical DNA conformation revealed high bend ability in the EGFR polymorphic region, especially if the CA stretch is extended. These data suggest that the (CA)\textsuperscript{n} SSR can act like a joint, bringing the promoter into proximity with a putative repressor protein bound downstream of the (CA)\textsuperscript{n} SSR. It is noteworthy that triplet SSRs may be preferentially located in regulatory genes related to transcription and signal transduction and under-represented in genes for structural proteins (Young et al., 2000), also suggesting an SSR effect on gene transcription.

### 3.4 BOVINE REPETITIVE SEQUENCES

Approximately 30% of the bovine genome consists of repetitive DNA sequences. Earlier hybridization kinetics experiments indicated that there are about 60,000 copies of these sequences per genome (Lenstra et al., 1993). Subsequently, it was shown that these sequences belong to two organizational classes, tandem repeats or isolated short, interspersed, repeated sequences. Five density satellites constitute about 20% of the genome, and most of the remainder consists of alternating repeating
and non-repeating sequences. The non-repeating sequences have a very broad size
distribution averaging 4,000 nucleotide pairs in length, with the longest exceeding
10,000-nucleotide pairs. The interspersed repetitive sequences are much more nearly
homogeneous in size, averaging 350 nucleotide pairs in length, and are divided into 8
to 14 sequence families (Mayfield et al., 1980).

Buffalo, cattle, goat, and sheep are members of a stable Robertsonian system,
the superfamily Bovidea, in which karyotypic evolution occurs almost exclusively by
centric rearrangements. In Bos taurus, based on buoyant densities in CsCl, eight
different satellites have been distinguished (Macaya et al., 1978; Kopecka et al.,
1978). Comparison of the primary structure of five of these satellites reveals a
common ancestral origin consisting of a dodecanucleotide sequence of
GATCGGCTAC (T/C) (Taporowasky and Gerbi, 1982). The C banding technique
reveals a constitutive heterochromatin in the centromeres of all the bovine autosomes,
but not X and Y-chromosomes. The location of four of the satellite sequences (I, II,
III and IV) on these chromosomes has been determined (Kurnit et al., 1973). Satellite
1 is present in the centromere of all the chromosomes. Two thirds of all satellite DNA
are observable in centromeres of the autosomes, with the remaining one third present
in the telomeric or interstitial regions. Satellite III and IV are localized in the
centromeres of most, but not all the chromosomes. There are few, if any satellite
DNA on the sex chromosomes.

In domestic animals, like cattle, the main purpose for generating polymorphic
DNA markers is for constructing genetic linkage maps, which in turn are developed
for identifying regions of the genome that influence economically important traits.
Mapping of underlying genes or quantitative trait loci (QTL) allows marker-assisted
selection (MAS), which is expected to increase the rate of genetic progress (Georges
et al., 1993). Examples of cattle QTL mapped by microsatellite analysis are horn
development (Georges et al., 1993b), weaver disease (Georges et al., 1993) and milk
production (Georges et al., 1995). Soon after the first reports about microsatellite
polymorphisms in humans (Litt and Luty, 1989), Weber and May 1989 showed that
dinucleotide blocks might be an abundant source of DNA polymorphism in cattle.
Vaiman et al., 1994 suggested that the cattle genome possesses less (CA)n/(GT)n
sequences than other mammalian species and gave an estimate of few thousand to
30,000 as the total number of microsatellites. The small and large scale development
of random microsatellite markers (e.g. Vaiman et al., 1994; Moore et al., 1994)
resulted in genetic linkage maps for cattle (Barendse et al., 1994; 1997; Bishop et al., 1994). Currently Ninety-nine loci have been assigned to river buffalo chromosomes, 67 of which are coding genes and 32 of which are anonymous DNA segments (microsatellites) (El Nahas et al., 2001). These markers used for buffalo studies mapped to the 29 cattle autosomes and the X and Y-chromosomes, the close chromosomal relationship between cattle and buffalo can thus be exploited to construct buffalo genome map.

3.5 THE DYNAMIC SPERMATOZOA AND TRANSCRIPTION

A mammalian ejaculated spermatozoon comprises highly differentiated cells that originate from the complex process of spermatogenesis, involving three major steps: the proliferation and differentiation of spermatogonia; the meiotic divisions during the spermatocyte stage; and, finally, the spermiogenesis. The transformation of spermatids into spermatozoa entails major morphological and molecular changes concerning acrosome and flagellar formation, cytoplasm elimination and mitochondria rearrangement and, finally, nuclear reshaping.

Spermatozoa transport not only paternal DNA but also protein-coding and non-coding RNAs to the oocyte. Since the early study by Pessot et al., 1989 several other studies have discussed the detection of specific transcripts in mammalian spermatozoa (Zhao et al., 2006). The majority of such transcripts have been linked to sperm development; however, with the advent of high throughput technology, such as microarray technology, the functional diversity of spermatozoal RNA is increasing. Spermatozoa, which are devoid of transcriptional and translational activity, are terminally differentiated cells produced in the complex process of spermatogenesis. Spermatogenic cells produce an impressive amount of polyadenylated poly (A) RNA, accounting for as much as 30% of the total RNA. Many researchers have stated that spermatogenic cells produce too much poly RNA for their own needs, especially given the reduced translation efficiency of these cells. Another peculiarity of spermatogenesis is that much of the RNA produced in spermatogenic cells is stable. Some RNAs transcribed in spermatocytes are preserved until late spermiogenesis (Langford et al., 1993). Various authors have used RNA profiles to qualify either a sperm function or a role in early embryogenesis. The messenger RNA (mRNA) distribution for specific transcripts between low- and high-motile sperm isolated from the same sample has been compared. The levels of different transcripts coding for
molecules involved in either nuclear condensation or sperm function are associated with the motility status of the sperm cells, although viability was similar in high- and low-motile fractions. Recent studies suggest that spermatozoal RNA might have a role in post-fertilization events. Miller et al., 2005 reported that some mRNAs detected in sperm cells and early embryos are absent in non-fertilized oocytes. Furthermore, microRNAs present in spermatozoa could also be involved in embryo development. Because of its origins and because of the peculiarities of spermatogenesis, spermatozoal RNA can be regarded not only as remnants that escape cytoplasmic extrusion during the last steps of spermiogenesis but also as a fingerprint of spermatogenesis quality and fertility outcome. An interesting clinical application of transcript profiling has been proposed which consists in assessing fertility potential using sperm samples as non-invasive biopsies. For example, abnormal \textit{KLHL10} gene (two missense mutations) was detected in spermatozoal RNA and associated with impaired fertility. Spermatozoal RNA profiles of normal fertile men have been established (Ostermeier et al., 2002), and is expected to play an important role in the assessment of idiopathic cases of infertility. Transcript analysis of spermatozoa content is now considered as a non-invasive approach for analyzing genetic defects of humans and other species.

### 3.5.1 Spermatozoal RNA in Fertility Research

Approximately 5000 spermatozoal transcripts are believed to be underpinning spermatogenesis and, by extension, male fertility (Ostermeier et al., 2002). Spermatozoal RNA isolated from ejaculate resents a unique opportunity to globally address the mechanisms that control the differentiation of the male gamete during normal, perturbed and diseased states. They provide a useful molecular record to assess environmental insult and/or genetic status since spermatogenesis is highly sensitive to environmental exposures including chemical, thermal and biological agents. Recently, two independent studies in humans have used RNA profiling techniques to address the relationship of motility and the RNA population between normal and motility-impaired sperm. Interestingly several transcripts in humans were identified that varied in a significant manner between the normal and motility impaired samples. These included testis specific protein 1 and lactate dehydrogenase C transcript variant 1 (Martins and Krawetz, 2005). This clearly points to the potential of this strategy to be used as a clinical assay to provide a panoramic view of testis...
gene expression that can be difficult to achieve from a testicular biopsy. Defining the “fertile male fingerprint” would have a significant impact on diagnosis, treatment and counseling. Realization would signal a major advance in the field of andrology.

3.5.2 Spermatozoal Functions

✓ Paternal chromatin packaging: the vast majority of DNA within the spermatozoon nucleus is packaged by highly charged protamines. However, a small proportion of the genome is packaged by histone-containing nucleosomes. It is proposed that these segments are marked with paternally derived RNA.

✓ Imprinting: there is mounting evidence that antisense RNAs play a critical role in establishing silenced chromatin domains. There is the possibility, therefore, that the recently discovered antisense spermatozoal RNAs could provide an epigenetic mark that is necessary for establishing and/or maintaining paternal imprints.

✓ Sperm maturation: under certain conditions, some cytoplasmic mRNAs are apparently translated de novo, possibly on mitochondrial polysomes. When this translation is prevented, sperm maturational events associated with capacitation are not observed. This would suggest that the production of proteins from spermatozoal RNAs is necessary for the maintenance of male fertility.

✓ Developmental maintenance: in addition to structural roles, it has been suggested that in common with oocyte RNA stores, spermatozoal transcripts may be functionally important to the zygote by actively promoting post-fertilization development. Recent evidence also would suggest that paternal RNAs could provide epigenetic marks to the developing embryo that influence the phenotype of the offspring.

✓ Embryonic spatial patterning: combining the identification of spermatozoal RNAs within the midpiece and the dependence of RNA partitioning by the paternally derived centrosome, it is tempting to speculate that paternal transcripts may play a key role in spatial patterning of the developing zygote.

3.6 RESTRICTION FRAGMENT LENGTH POLYMORPHISMS (RFLPS)

RFLPs were the original DNA markers developed in the late 1970s (Botstein et al., 1980). Their development was facilitated by the discovery of restriction enzymes, which cut DNA at specific sequences giving rise to restriction fragments.
The basic technique for detecting RFLPs involves the fragmentation of genomic DNA by a restriction enzyme. The resulting DNA fragments are then separated by length through a process known as agarose gel electrophoresis, and transferred to a membrane via the Southern blot procedure. Hybridization of the membrane to a labeled DNA probe then determines the size of the fragments, which are complementary to the probe. An RFLP occurs when the size of a detected fragment varies between individuals. Each fragment size is considered an allele, and can be used in genetic analysis. RFLP's are quick, simple and inexpensive ways to assay DNA sequence differences. It is the first DNA polymorphism to be widely used for genomic characterization, which detects variations ranging from gross rearrangements to single base changes. The polymorphisms are found by their effects on sites for restriction enzyme mediated cleavage of preparations of high molecular weight DNA. This method has been used extensively to reveal DNA fingerprinting, useful in the identification of samples retrieved from crime scenes, in the determination of paternity, and in the characterization of genetic diversity or breeding patterns in animal populations.

3.7 TRANSCRIPTOMES AND MINISATELLITE ASSOCIATED SEQUENCE AMPLIFICATION (MASA)

The term transcriptome is young and its mention first appeared in 1997 to describe the set of genes expressed from yeast genome (Velculescu et al., 1997). According to a more recent definition the term stands for set of all messenger RNA (mRNA) molecules, or "transcripts," produced in one or a population of cells. In addition to mRNA it also includes various non-coding RNAs playing either structural or regulatory functions in cells. Hence there are thousands of different transcriptomes in hundreds of different cell types and organs in their various physiological and pathological states. Recent studies using minisatellites markers have shown that mining satellite transcripts from somatic tissues, gonads and spermatozoa reveals various transcriptome(s) which shows tissue stage and specifies specific expression profile (Fortune et al., 2000). Minisatellites are ubiquitously distributed in eukaryotic genome and show extensive variation in number of repeats leading to multiallelic variation and high degree of heterozygosity. This extreme level of variation renders minisatellite invaluable as genetic markers. Minisatellite Associated Sequence amplification (MASA) involves random amplification of genomic or cDNA
sequences with minisatellite consensus sequence, that are evolutionarily conserved as a primer in the PCR reaction. Unlike restriction fragment length polymorphism (RFLP) which requires at least 100 folds intact DNA, MASA can be performed with very smaller quantity of target substrate. Jeffreys et al., 1985 described hypervariable minisatellites, which have proved to be useful tools for many studies in human molecular genetics. Two classic minisatellite probes, 33.6 and 33.15, derived from these minisatellite sequences showed a large number of highly polymorphic fragments in *HinfI* digests of human DNA. This extremely polymorphic pattern due to variation in the number of repeats of a 'core' 16-mer sequence allowed the precise identification of a particular individual by DNA fingerprint. Much of the studies so far have been concentrated using 16-nt (5' CACCTCTCCACCTGCC 3') consensus sequence of 33.15 repeat loci for MASA reaction. MASA with 33.15 in human genomic DNA detected a 513 bp Y chromosome specific band in males as well as a number of variable bands ranging from 260–2700 bp in both the sexes (Bashamboo and Ali, 2001). In Indian rhinoceros (*Rhinoceros unicornis*) genomic DNA amplification with 33.15 primer-identified band profiles specific to the *R. unicornis* and *D. bicornis* species (Kapur et al., 2003). In water buffalo *Bubalus bubalis*, 33.15 and GATA/GACA repeat loci generated a large number of known and novel genes with potential functions (Srivastava et al., 2006, 2008).

The novel part of the current approach is that functional, structural and regulatory genes associated with minisatellites are accessed without screening the conventional cDNA library proving this to be highly useful for such genome analysis where prior information is absent or inadequately available. The expression profile of genes based on MASA under normal and abnormal conditions is envisaged to be of great relevance for identification of event/stage specific mRNA transcripts. In the context of comparative genomics, mRNA transcripts commonly expressing in a large number of species may be segregated. Following this approach, genes with highest levels of expression in a given tissue may be easily identified and the information from different breeds of animals may be established. Thus, differential expression of genes accessed by MASA may be used to establish a genotype phenotype correlation in the context of genetic diseases, cancer biology, stem cell research, tissue engineering, organ transplantation, animal cloning, characterization of genetic integrity of different cell lines and conducting translational research. Thus MASA approach can be used as a reliable tool to develop species-specific marker profile.
Minisatellite 33.6

The consensus of minisatellite 33.6 used in the present is an 11 bp CCTCCAGCCCT repeat, originating from the human myoglobin gene (Jeffreys et al., 1986). Chimini et al., 1989, observed a single major hybridization peak when 33.6 was used for in situ hybridization to human metaphase chromosomes, thus permitting mapping to 1q23. Furthermore, hybridization to human DNA cleaved with 'rare-cutter' enzymes and fractionated on pulsed field gels also showed a fairly simple, largely monomorphic pattern which allowed chromosomal assignment using somatic cell hybrids. Polymorphic pattern generated by this probe has been used extensively in paternity or sibship testing and in a number of forensic studies. In wide range of animal species such as dog, cats, fishes, whales, honeybees, birds, butterfly (Jeffreys and Morton, 1987; Taggart and Ferguson 1990; Amos et al., 1991; Blanchetot 1991; Saccheri and Bruford, 1993) and plants (Tourmente et al., 1998; Broun and Tanksley, 1993), 33.6 probe has been used to generate DNA fingerprints. Minisatellite 33.6 produced individual specific band profile in the human (Azfer et al., 1999). This repeat used for MASA uncovered several bands across the tissues in buffalo, demonstrating its ubiquitous distribution. However at transcription level, information from this minisatellite is almost negligible. Clearly, MASA with minisatellite 33.6 using cDNA from different tissues and spermatozoa provides an opportunity to assess all the expressed genes and their mode of actions and interactions within bubaline genome. This approach will positively bridge the gap and facilitate much-needed advance research to identify genetically superior germplasm in context of animal genetics.