INTRODUCTION
I. INTRODUCTION

Maximum amount of mammalian genome is composed of repeat sequences and minimum fraction is entrusted for protein coding function. For example, >50% human genome is repetitive sequences and <2% is protein-coding [Lander et al., 2001]. Out of this repeat sequences, about 20% are transposon-like sequences belonging to the long interspersed nuclear elements (LINEs), which are non-LTR retrotransposons functioning as autonomous polyA-transposons. There are also small interspersed nuclear elements (SINEs) up to 13%, which also act as retrotransposons, with the help of LINEs. Besides these, there are several other types of repeat sequences, e.g., simple repeats, microsatellites, minisatellites and others [Lander et al., 2001; Waterston et al., 2002; Gibbs et al., 2004; Goodier and Kazazian 2008].

Table I and Figure I show repetitive DNA in mammalian genome. The human, mouse and art genomes are composed of about 20%, 20% and 23% LINEs consisting of LINE-1, LINE-2, LINE-3 subclasses of non-LTR, autonomous, poly A retrotransposons respectively. Similarly, the SINEs (Alu and MIR sequences) compose about 13%, 7.8% and 7% of the human, mouse and rat genomes, respectively. Besides this, the LTR-elements are up to 8%, 10% and 9%. The DNA-elements are 2.8%, 0.86% and 0.8% respectively. The total interspersed repeats are up to 44.8%, 39% and 40% in the human, mouse and rat genomes, respectively. Therefore, repetitive DNA in mammalian genome is of a wide variety but their function is not yet understood. Function of repetitive DNA is much less clearly understood. Repeat sequences are present in chromosomal structures, e.g., centromere and telomere, and contribute to chromosomal stability, segregation and genomic homeostasis. Centromere and telomere are also involved in chromosomal abnormality and instability and therefore, diseases like cancer [McCord and Broccoli, 2008]. Although, repeat sequences as DNA structures at centromere and telomere are well known, but repeat sequences as RNAs are much less explored. Repeats can make elaborate self-folding and parallel-strand, triplex (H-DNA), G-4 DNA etc [Neidle and Parkinson, 2003]. Secondary structures RNA can also attain elaborate secondary and tertiary structures and biochemical properties, activities due to the presence of 2'-OH, modified as well as unusual bases (e.g., DHU, A6m, G7m, Ψ, I) [Tinoco Jr et al., 2008]. It can make bulge, stem-loop, hair-pin, crucifer, hammerhead, pseudoknot and many other structures [Hartman et al. 2009, Brown 1999]. RNA can act as a catalyst (ribozyme) and riboswitch [Strobel and Cochrane, 2007]. It has A.U and G.U base pairing. RNA can act as a messenger (mRNA).
### Table I. Comparison of repeat sequences in human, mouse and rat genome

<table>
<thead>
<tr>
<th>Types of Repeats</th>
<th>Fraction of Human Genome (2.9 Gb)</th>
<th>Fraction of Rat Genome (2.8 Gb)</th>
<th>Fraction of Mouse Genome (2.6 Gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINEs</td>
<td>20.42%</td>
<td>23.11%</td>
<td>20.10%</td>
</tr>
<tr>
<td>LINE-1</td>
<td>16.89%</td>
<td>22.73%</td>
<td>19.65%</td>
</tr>
<tr>
<td>LINE-2</td>
<td>3.22%</td>
<td>0.33%</td>
<td>0.38%</td>
</tr>
<tr>
<td>LINE-3</td>
<td>0.31%</td>
<td>0.06%</td>
<td>0.06%</td>
</tr>
<tr>
<td>SINEs</td>
<td>13.14%</td>
<td>7.05%</td>
<td>7.78%</td>
</tr>
<tr>
<td>Alu</td>
<td>10.60%</td>
<td>1.65%</td>
<td>2.56%</td>
</tr>
<tr>
<td>MIR</td>
<td>2.20%</td>
<td>0.51%</td>
<td>0.56%</td>
</tr>
<tr>
<td>LTR elements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERV_class I</td>
<td>8.29%</td>
<td>9.04%</td>
<td>10.28%</td>
</tr>
<tr>
<td>ERV_class II</td>
<td>2.89%</td>
<td>0.97%</td>
<td>0.79%</td>
</tr>
<tr>
<td>ERVL (III)</td>
<td>1.44%</td>
<td>3.24%</td>
<td>4.13%</td>
</tr>
<tr>
<td>MaLRs</td>
<td>3.65%</td>
<td>0.84%</td>
<td>1.08%</td>
</tr>
<tr>
<td>DNA elements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclassified</td>
<td>2.84%</td>
<td>0.81%</td>
<td>0.86%</td>
</tr>
<tr>
<td>Total interspersed repeats</td>
<td>44.83%</td>
<td>40.31%</td>
<td>39.45</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>0.45%</td>
<td>2.38%</td>
<td>2.41%</td>
</tr>
</tbody>
</table>

Ref: [http://www.nature.com/nature/journal/v428/n6982/full/nature02426.html](http://www.nature.com/nature/journal/v428/n6982/full/nature02426.html), Lander et al., 2001; Waterston et al., 2002; Gibbs et al., 2004.
Figure 1. Composition of human, mouse and rat genomes. Various types of sequences and their relative percentage is mentioned. (A), (D) human, (B), (E) mouse and (C), (F) rat genomes. Abbreviation: LINE: Long interspersed nuclear element, Alu, MIR: Mammalian wide interspersed repeat, ERV: Endogenous retroviruses, http://www.ncrna.org/statgenome/, Human Mar 2006 assembly, Mouse Feb 2006 assembly, Rat Nov 2004 assembly; Lander et al., 2001; Waterston et al., 2002; Gibbs et al., 2004
Fraction of repeats in human genome

Lander et al., 2001
Figure I.

Waterston et al., 2002

Gibbs et al., 2004
adapter (tRNA), scaffold/matrix (rRNA), catalyst (RNase P, Group I, II introns, 23 SrRNA, peptidyl transferase etc), guide (guide RNA during RNA editing), template (e.g., telomere RNA, LINE RNA), catalytic center (e.g. U2-U6 SnRNAs), regulator of gene expression (e.g. miRNAs), signaling molecule (dsRNA/TLR-3 receptor), control for translation (e.g., miRNA, ScRNA), toxic pathogenic molecule (e.g., (CAG)n repeat-RNA in neurodegenerative diseases) [Wahl et al. 2009; Orr and Zoghbi, 2007], genome (e.g., HIV-1, Measles viruses), weapon against transposons (e.g., piRNA), switch (e.g., riboswitch) and many others, yet to be discovered [Mello and Conte, 2004; Carl and Phillip, 2004; O'Donnell and Boeke, 2007].

RNA structure is important for cellular functions. For example, transfer RNA (tRNA) maintains its characteristic tertiary structure due to its intra-strand H-bonding to form base pairing and the loops containing various unusual and modified bases preventing base pairing. tRNA structure is important to bind with aminoacyl-tRNA synthases, get charged with activated amino acids (aa) by the enzyme, proofreading of aa correct charging, interacting and specifically recognizing the codons with the help of its anticodon, maintaining the flexibility for degeneracy due to the “Wobble effect” for efficient decoding of the genetic code, interact with ribosomal A site, P site and E site to undergo the process of initiation, elongation, translocation during protein synthesis. The 3’-CCA 2’-OH of the peptidyl tRNA, also takes part in peptide-bond formation, so much so that EF-G, RF1, RF2 mimic tRNA-structure in order to function during translocation and termination steps of translation. RNA structure is important for splicing of mRNA, e.g., the 2’-OH nucleophillic attack by the branch point A in group-II and premRNA introns effects the first trans-esterification reaction during cleavage of the 5’-p of the intron, self-splicing of Group-I introns require the specific G-pocket RNA structure for protein-free catalytic activity, RNase P-M1 RNA needs Mg²⁺-coordinated structure to catalyze the cleavage, structure of U1, U2, U4, U5 and U6 snRNAs are important for the spliceosome assembly and nuclear pre-mRNA splicing. Both the consensus sequences and the secondary structures are essential in RNA processing.

RNA structure is important in rRNAs and ribosome biogenesis and function. Mutations and binding of antibiotics to some rRNA sequences and structures abolish protein synthesis. The specific endonucleolytic cleavage and 2’-O-CH₃-ribose modification of rRNA structures are essential for rRNA processing, maturation and ribosome biogenesis. The 16S rRNA 3’-end structure and its interaction with the 5’-end of mRNAs at the Shine-Dalgarno sequence in
prokaryotes and the counterpart in 18S rRNA with the 5'-end Kozak sequence in
eukaryotic mRNAs are essential for recognition of the translational start codon
(AUG) and assembly of pre-initiation complex. Eukaryotic mRNAs through their
elaborate 5'-untranslated region (UTR) and 3'-UTR structures bind specific
proteins to effect translation and regulate it. The 3'-UTR also regulates half-life of
the messanger RNA [Ross 1996].

Recently a novel class of small regulatory RNAs have been reported, e.g.,
siRNA, rasiRNA, miRNA, piRNA, tiRNA, which are generated through elaborate
RNA processing pathways to regulate gene expression. They have been also linked
to disease processes, e.g., cancer and neurodegeneration [Ryazansky and Gvozdev,
2008; Ventura and Jacks, 2009; Stefani and Slack, 2008]. The unstable repeats in
the mRNA, e.g., (CAG)n, (CGG)n, trinucleotide repeats in codons of specific
mRNAs, sometimes undergo trinucleotide repeat-expansion mutations and cause a
variety of neurodegenerative diseases. Such mutations in the 5'-UTR and introns
have also been shown to cause such diseases. Thus, there is a field of RNA-
mediated diseases [Kumari and Usdin, 2009; O’Rourke and Swanson, 2009].

Although, it was earlier known that mammalian genomes contain repetitive
DNA, the post-genome sequencing phase (2001 onwards) really brought a big
surprise that the human genome contains >50% repetitive DNA, similar
information came from mouse and rat genomes, which were sequenced and
available following years [Lander et al., 2001; Waterston et al., 2002; Gibbs et al.,
2004]. A large amount of repeats selected along with evolution of the mammalian
genome pleads for its advantage for the organism as opposed to a passively carried
over “selfish DNA” concept. This may be linked to evolution of structural
organization of the DNA in the form of chromosomes and partitioning the
chromatin into euchromatin and heterochromatin [Shapiro and Sternberg, 2005].
This may also be related to complex developmental processes, e.g., dosage
compensation and X-chromosome inactivation in mammalian females with XX
chromosomes [Payer and Lee, 2008], maintaining the heterochromatic Y
chromosomes [Steinemann and Steinemann, 2000]; silencing gene expression for
differentiation of cells into a large number of functionally specialized cells,
chromosome condensation at metaphase to facilitate segregation of a diploid set of
chromosomes during mitosis in somatic cells and reduction division during meiosis
in germ cells to generate haploid male and female gametes; allowing exchange of
genetic information between homologous sequences/genes; maintaining ploidy of
the organism and developing sexual barrier between species, therefore, bringing in
the time component of evolutionary scale into complex genomes. The complex
large genomes have many functions where repetitive DNA may be involved such as insulating genes and their expression patterns from one locus to another, developing islands of locus control regions (LCRs) for coordinating groups of similar genes expressing in a coordinated, developmentally regulated manner (e.g., globin gene locus); anchoring genomes to nuclear matrix, allowing accessibility of various protein factors to select chromatin regions, preventing chromosomes from fusing with each other or translocating from one region to another, facilitating methylation as a mechanism for genome imprinting [Green et al. 2007], transposition of genomic DNA of ‘jumping genes’, which Barbara McClintock, who discovered these elements during 1930-1950, organization of polytenized chromatids, e.g., the polytene chromosome of *Drosophila*, evolution of new species by polyploidy (e.g., Rice and Wheat, the two major plants providing food to the mankind); developing consistency of chromosome number and forms (karyotype) of species; higher order packaging of chromatin and may be many others.

Similarly, abnormality in the repeat sequences is manifested by multiplication of centrosome leading to aneuploidy during cancer [Alberts et al., 2008], shortening of telomeric repeats during aging [Stewart and Weinberg, 2006], amplification of telomere during cancer, instability and expansion of trinucleotide repeats of the nature (CAG)n and (CGG)n, which make Glutamine (Q) leading to poly Q in neurodegenerative diseases, e.g., Fragile X-associated Tremor/Ataxia Syndrome (FXTAS) AND Fragile X syndrome (FXS) involving (CGG)n repeat expansion in the 5'-UTR of FMR1 gene [O'Rourke and Swanson, 2009]. In FXTAS disease it does not encode a protein, instead the CGG repeat in RNA are the likely cause of neurodegeneration in FXTAS O'Rourke and Swanson, 2009. These are example of simple sequence repeats in RNA, which is toxic to cells, causing many neurological disorders [Orr and Zoghbi, 2007]. Therefore, it is too simple to call repeats as ‘Junk DNA’. Small regulatory RNAs e.g., siRNAs produced from the centromeric repeats have been shown to cause histone methylation and heterochromatin formation. The position variegation effect, the white eye colour in *Drosophila*, is also known to be regulated by heterochromatin environment and suppression of translation from a normal gene for orange red eye colour gene. Repeats in RNA have been shown to affect splicing by affecting interaction between splicing signals and spliceosome. They can bind to protein factors to alter transcript processing, can form secondary structures resulting in sequestration of splicing signal by altering existing RNA-RNA, RNA-protein or protein-protein interactions [Hefferon et al., 2004]. Repeats in 5'-UTRs and 3'-
UTRs of mRNAs have been shown to regulate translation and half-life of the message [Ross, 1996; Alberts et al., 2008]. Thus repeats at both DNA and RNA levels have effects on cellular functions. In addition, certain simple repeat sequences have been shown to bind nuclear proteins [Epplen et al. 1996; Ishikawa et al. 1993]. About 6400 polymorphic microsatellites (simple repeats) have been used to prepare the physical map of human genome. The Alu (SINE) sequences in the human genome are about 300 nt long and each Alu is unique for its position and sequence [Batzer and Deininger, 2002]. We hypothesize that noncoding repetitive DNA sequences in mammalian genome transcribe into RNAs, which make elaborate structures and provide them as interface for intermolecular interactions of the nature of RNA:RNA and RNA:protein. This is a key to noncoding RNA function under cellular conditions. RNA structures can also bind small molecules e.g., Mg$^{2+}$, GTP, ATP, drugs, antibiotics etc [Sucheck and Wong, 2000]. Therefore, looking for repeats in small and large RNAs may be rewarding.

Earlier in our laboratory, in an attempt to identify, isolate and characterize cDNAs containing (GA/CT)$^n$ simple repeats, a λgt11 human testis cDNA library and a λgt11 rat testis cDNA was screened by a 227 bp rat genomic simple repeat DNA (GenBank Accession No. X 97459). The 227 bp DNA has a central (GA)$_7$A (AG)$_7$ dinucleotide mirror repeat with the potential to form a triplex (H-DNA)-like structure in vitro and shows homology with various eukaryotic RNAs due to the presence of (GA/CU)$_n$ and other repeat motifs [Dey and Rath, 2005]. Fifteen positive human λgt11-cDNAs and 27 rat λgt11-cDNAs were isolated. Out of these three rat cDNAs: 4.1, 5.5 and 11.4 and two human cDNAs: TPIP-C2 and H3-SRY were chosen and characterized for their function in this study. The five cDNA sequence has been analyzed in silico, the RNA expression patterns have been studied in rat, mouse and human cells. Functions of the RNA/protein products have been analyzed by cell-transfection or recombinant protein characterization assays. Small RNAs have been studied in case of two noncoding RNAs.