Published Online on 25 June 2015

Proc Indian Natn Sci Acad **81** No. 3 June 2015 pp. 609-628

DOI: 10.16943/ptinsa/2015/v81i3/48230

Review Article

Code in the Non-Coding

JAYA KRISHNAN and RAKESH K MISHRA*
Center for Cellular and Molecular Biology, Council of Scientific and Industrial Research, Uppal Road, Hyderabad 500 007, India

(Received on 31 March 2015; Revised on 13 May 2015; Accepted on 18 May 2015)

Genomes are comprised of both protein-coding and non-coding sequences. Strangely, most eukaryotic genomes are made up of huge amounts of non-coding regions while a relatively smaller part of the genome codes for proteins. The significance underlying the abundance of non-coding sequences has been elusive for many decades. The non-coding part of the genome comprises of sequences like transposons, satellite DNA, introns, pseudogenes, etc. With technological advances, we have now been able to know more about this part of the genome. Emerging studies show that these sequences perform various kinds of functions, many of which are regulatory in nature. Here, we present overviews of various kinds of non-coding elements found in eukaryotic genomes and discuss the roles that they perform in the genome. We suggest that a significant proportion of the non-coding DNA is an essential component of the genetic make-up in higher eukaryotes that has accumulated during evolution for regulatory function.

Key Words: Non-Coding DNA; Repetitive DNA; Regulatory Elements; Eukaryotic Genome

Introduction

A diploid human genome consists of 6 billion base pairs of DNA in the form of 23 pairs of chromosomes. One of the striking outcomes of the human genome sequencing projects was that the protein coding part constituted only ~2% of the entire DNA content. The remaining 98% of the genome, the non-coding part, has attracted significant degree of attention and debate over the past few decades. Comparison of genome from different organisms across the species shows that, while the genome sizes of organisms have increased with the increase in their complexity, the number of genes has not increased proportionately (Table 1). This observation has led to the hypothesis that much of the non-coding part of the genome has evolved under positive selection pressure and that it may have a functional relevance. The non-coding elements in genome can be divided into two categories – unique and repetitive sequences. Unique sequences include elements like promoters, enhancers, repressors, boundary elements, introns, conserved regions, pseudogenes and sequences that get transcribed into non-coding RNAs while repetitive sequences include transposable elements, satellite DNA, etc. (Table 2). In this review, we discuss the regulatory functions of a variety of DNA sequence motifs that constitute significant part of the genome as well as the major classes of non-coding transcripts.

Unique Non-Coding Sequences

Promoters and Associated DNA Elements

Each coding unit is associated with a region at the start of the unit that serves as the binding site for the transcription machinery for the formation of the transcription initiation complex. This region is the promoter. The first level of diversity in gene expression rests on the promoters. Characteristic features of promoters are different for different RNA

*Author for Correspondence: E-mail: *mishra@ccmb.res.in; Tel.: +91 40 27192658*

Table 1: Genome sizes of human and various model organisms showing that increase in genome size is not proportional to the increase in gene number which indicates accumulation of large amounts of non-coding DNA

Organism	Genes	Genome size (Mb)
Mycoplasma genetalium	517	0.58
Escherichia coli	4377	4.6
Sacchromyces pombe	4929	12.26
Sacchromyces cerevisiae	5770	12.49
Neurospora crassa	10000	39.9
Drosophila melanogaster	17000	122.6
Arabidopsis thaliana	27407	130
Caenorhabaditis elegans	21733	1000
Danio rerio	26206	1400
Mus musculus	23000	2800
Homo sapiens	23000	3300

Table 2: Components of the Human genome

Component	% in genome
Protein-coding genes	2
Introns	26
Long Interspersed Elements (LINEs)	20
Short Interspersed Elements (SINEs)	13
Heterochromatin and other sequences	13
Long Terminal Repeats (LTRs)	8
Other unique sequences	7
Conserved Non-coding Sequences (CNCS)	5
DNA transposons	3
Simple Sequence Repeats	3

polymerases (RNA Pol I, II and III). Pol II promoters form the most diverse class of promoters and consist of the DNA element that extends to about 35 bp upstream and/or downstream of the transcription initiation site, referred to as the core promoter. The core promoter is responsible for binding the polymerase and some core initiation factors. One of the best-studied core promoter elements is the TATA box that has the consensus sequence of TATATAAG, which is recognized by the TATA binding factor (TBP), a part of the TFIID complex, and directs the start of transcription from 25bp downstream of it (Goldberg, 1979; Smale and Kadonaga, 2003). In addition to the TATA box, metazoan core promoters can be composed of numerous other elements, including: Initiator element (Inr), Downstream Promoter Element (DPE) (Kutach and Kadonaga, 2000), Downstream Core Element (DCE) (Lee *et al*., 2005a), TFIIB-Recognition Element (BRE) (Lagrange *et al*., 1998), and Motif Ten Element (MTE) (Lim *et al*., 2004) (Fig. 1).

Many promoters also contain promoter proximal sequences that are the elements found within -200 of the TSS. They assist core elements in enhancing transcription. In vertebrates, promoters are associated with CpG islands that are short stretches of C-G nucleotides that are unmethylated when the associated gene is active and methylated when inactive (Deaton and Bird, 2011). Promoters give rise to a high level of diversity in gene expression and regulation. First level is achieved by mere sequence variations in them (Kim *et al*., 2008; Olivier, 2004). The sequence changes cause the transcription factors to bind with differential affinity, which thereby brings changes in gene expression. Secondly, certain promoters are also known to be tissue-specific in nature, which is brought about by the presence or absence of specific factors. Most promoters contain a unique combination of core elements, which also contributes to the tissue-specificity and differential gene expression (Ohtsuki

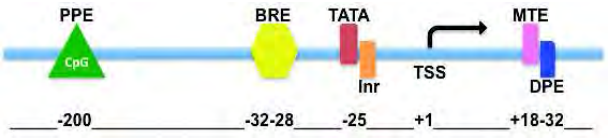

Fig. 1: Relative positioning of the various promoter elements with respect to the Transcriptional Start Site (TSS) or +1 nucleotide. MTE – Motif Ten Element, DPE – Downstream Promoter Element, TATA – TATA box Inr – Initiator, BRE – TFIIB Recognition Element, PPE – Promoter Proximal Element, CpG indicates enrichment of CpG islands on PPEs

et al., 1998). Thirdly, promoters are also targets of epigenetic modifiers to bring about gene regulation. Generally, H3K4me3 marks active promoters along with the presence of RNA Pol II while inactive ones have H3K27me3 mark. There are also poised promoters that carry bivalent marks of H3K4me3 and H3K27me3 (Bernstein *et al*., 2006; Ernst and Kellis, 2010). Lastly, transcription has been seen at the promoters that give rise to Promoter-Associated RNAs (PARs). These PARs comprise of both long and short RNAs and come from both the strands. Although the function of these RNAs has not been fully understood, some studies have shown that siRNAs target these PARs and thereby recruit repressive factors (Han *et al*., 2007; Kurokawa, 2011). Considering the average size and number, the promoters may constitute >1% of the euchromatic genome in higher organisms.

Enhancers

While promoters carry the potential to bring about diversity in gene expression and regulation, in the *in vivo* genomic context, when alone, they prove insufficient to carry out this job. Hence, genome has evolved another set of regulatory elements called the 'enhancers'. Precise temporal, spatial, and quantitative regulation of gene expression is essential for proper development. One of the elements in regulation of this precision is enhancer. The SV40 tumor virus DNA sequences were the first ones to be identified as enhancers (Banerji *et al*., 1981). Eukaryotic genomes are predicted to have thousands of enhancers but due to several reasons their identification has been a challenging task. Unlike promoter elements, enhancers can be located upstream, downstream or within an intron of a gene (Reviewed in Levine, 2010). Also, many enhancers are located far away from the target gene. Furthermore, enhancers, in general, do not have a consensus sequence and can be tissue specific, which may prevent their identification in the conventional transgenic assay. However, availability of certain tools has enabled us to identify enhancers in many genomes. For example, enhancer-traps in *Drosophila melanogaster* have been instrumental towards this end (Brand and Perrimon, 1993). Until recently, however, the target genes of many such enhancer-traps were unknown. A genome-wide enhancer characterization study has been carried out wherein expression patterns for more than 7000 such enhancer-trap *Drosophila* lines have been analyzed (Kvon *et al*., 2014). The study also identifies motif patterns present in the enhancers that correlate appreciably to their corresponding spatio-temporal expression domains.

More recently, the identification or prediction of enhancers in other genomes has been possible by use of epigenomic data. This includes histone marks and protein binding profile characteristic of enhancers. Conventionally, enhancers are sequences that have high concentration of transcription factors, like p300, Sox2 and Oct4 in mouse embryonic stem cells, binding to them. They are also marked by H3.3 and H2A. Z histone variants (Jin *et al*., 2009). There are also certain histone marks that are found to be associated with enhancers, for example, H3K27ac and H3K4me1. Another class of enhancers comprises the transcribed enhancers. It is now known that ~85% of the genome gets transcribed. A subset of these transcripts comes from enhancers that are known as eRNAs. These transcripts usually originate from active enhancers, which exhibit stronger signals for H3K4me1, H3K27ac and H3K79me2 marks (Kim *et al*., 2010; Orom *et al*., 2010). The genome-wide analysis, taking into consideration these criteria,predicts 0.4-1.4 million putative enhancers in the mammalian genome (ENCODE Consortium, 2012; Visel *et al*., 2009). A detailed analysis of the ChIP-seq data of the master transcription factors viz., Oct4, Sox2, Klf4, Esrrb and Nanog, which are known to bind to enhancers, revealed a new category of enhancer elements. These enhancer regions are large contiguous stretches bound by such proteins. When tested in reporter assays, they showed higher levels of gene activation as compared to the 'general enhancers'. Thus, these regions are termed as super-enhancers (Whyte *et al*., 2013).

Repressors

Repressor or silencer elements are DNA sequences that negatively regulate gene expression.They, like enhancers, can be present nearer or farther from their

target genes and function in an orientation independent manner. The best understood silencers are the Polycomb response elements or PREs. The PREs were first discovered in the *Drosophila* Hox complex wherein they help restrict the expression of each hox gene to its respective segment (Simon *et al*., 1993). These elements however have now been shown to be present all over the genome and are also known to function independently in transgenic contexts (Ringrose *et al*., 2003). Although PREs do not have any consensus sequence per say, it has become possible to map them to a great precision by using a combinatorial approach and looking for binding of various Polycomb group members along with the characteristic histone marks like H3K27me3 and H2A ubiquitinylation (Ringrose *et al*., 2003; Schwartz *et al*., 2012). The PcG proteins that bind on these elements are the major effectors of the PREs and they can be classified into three categories/complexes – PRCI, PRCII and the PhoRC complexes on the basis of the hierarchy in their function. The main members of the *Drosophila* PRC2 are the three PcG proteins – E(z) (Enhancer of zeste), Su(z)12 (Suppressor of zeste 12) and Esc (Extra sex combs), as well as Nurf55. The PRCII complex functions as a histone methyltransferase and it specifically methylates lysine 27 of histone H3. The major components of *Drosophila* PRC1 are the Ph (Polyhomeotic), Psc (Posterior sex combs), Sce (Sex combs extra; also known as Ring) and the founder member of the group, Pc. PRCI and PRCII do not have DNA binding factors (Bantignies and Cavalli, 2006). PRC1 is recruited to PREs by first binding with proteins like PHO, GAGA factor, Pipsqueak (PSQ), ZESTE, Sp1, DSP1, Grainyhead (GRH), etc. (Blastyak *et al*., 2006; Mishra *et al*., 2001; Brown *et al*., 2005; Brown *et al*., 1998), followed by the recruitment of PRC2 that methylates histone H3 at the lysine27 (H3K27me3). This mark is then recognized by the PRC1 that in turn establishes repressive chromatin and maintains it (Kahn *et al*., 2014; Muller *et al*., 2002) (Fig. 2). This model, although simple, does not explain the entire *in vivo* situations; for example PRC1 is also known to directly interact with Pho and thus has potential to get recruited onto the chromatin without the PRC2 mediation (Lanzuolo and Orlando, 2012). Thus the PREs along with the PcG protein help establish repressive chromatin and perform gene regulation. These elements are hence also known as the epigenetic DNA elements.

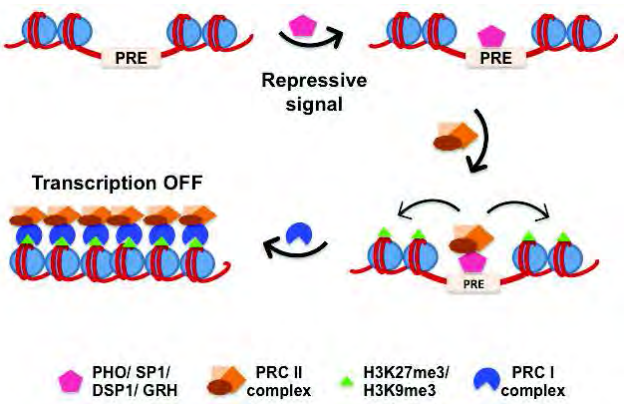

Fig. 2: Mechanism of repression by Polycomb Group proteins. Upon receiving repressive signals, proteins like PHO, SP1, DSP1 and/or GRH bind to the Polycomb Response Element (PRE). These proteins in turn recruit the PRCII complex that methylates the nearby histones. The histone marks are now recognized by the PRCI complex members, which in turn, repress the chromatin with the help of chromatin remodelers (not shown in figure)

Interestingly, the PRE elements are often juxtaposed to TRE (trithorax response element), which recruit activator proteins called the Trithorax group of proteins (trxG) (Schuettengruber *et al*., 2011). PRE/TREs along with these mutually antagonistic groups of proteins set the precise level of expression of target genes and form the epigenetic transcriptional memory of the cell. These elements, therefore, are also called the cellular memory modules (CMM). Once the gene expression state is established during early development, CCMs maintain that expression state through adulthood (Schwartz and Pirrotta, 2007). During replication, PcG and trxG complexes remain associated with CMM to remember and maintain the gene expression state (Petruk *et al*., 2012). PREs have been also identified in mammalian genomes indicating their conserved role in gene regulation (Liu *et al*., 2010; Mishra *et al*., 2007; Woo *et al*., 2010). There are also regions that have bivalent chromatin marks of H3K27me3 and K4me3, which are hypothesized to be the mammalian counterpart of

Drosophila PREs that switch gear upon differentiation (Bernstein *et al*., 2006). Apart from their role in early development, PRE/TRE also play important role in germ line stem cells (Chen *et al*., 2005), in tissue regeneration (Lee *et al*., 2005b) and several other developmental transitions (Bracken *et al*., 2006) and ageing (Mishra and Mishra, 2010).

Boundary Elements

Eukaryotes have thousands of regulatory elements, like enhancers and repressors, spread across the genome. These elements are able to precisely talk to their target genes/promoters. For example, while an enhancer may have the inherent capacity to talk to multiple promoters, in the genomic context it interacts only with its *in vivo* targets. Furthermore, eukaryotic genomes have two functionally distinct domains, heterochromatin and euchromatin, which are located and never cross-mingle. This is so because of the presence of regulatory elements called the insulators or the boundary elements. One of the first indications of the presence of such an element came from the observation in *Drosophila* polytene chromosomes that upon heat shock, the 87A7 locus showed puffing which is limited to the locus and led to the identification of two elements, one on either side of the puff, the scs and scs' (Specialized Chromatin Structures) elements (Udvardy *et al*., 1985). Such elements were termed as boundary elements. Thus, boundary elements are DNA sequences that divide the genome into functionally independent domains by delimiting the reach of the cis-regulatory elements within each domain. Depending on the assay used for the analysis, boundary elements are also known as 'enhancer blocker' insulators and barriers.

One of the best-characterized boundaries is the chicken beta globin insulator. This insulator has been shown to possess both enhancer blocking and barrier functions. The β-globin locus contains the clustered globin genes upstream to which are the folate receptor genes that are expressed when the globin genes are repressed and vice-versa. Further, downstream to the globin genes are the Odorant receptor genes that are repressed in erythroid cells. Thus, the globin gene locus forms a distinct expression domain and this is enabled by the presence of boundaries flanking it – 5'HS4 and the 3'HS1. The 5'HS4 has been extensively studied and has been found to be a CTCF-dependent boundary element. CTCF or the CCCTC binding protein was initially discovered as a negative regulator of *c-myc* and has now been shown to be involved extensively in boundary function and more recently as a genome architecture protein (Lobanenkov *et al*., 1990; Phillips and Corces, 2009). The 5'HS4 element has been extensively studied and validated in transgenic reporter constructs and shown to be functioning as an insulator. The element can also protect a transgene from position effects (Chung *et al*., 1993). The barrier function however has been shown to be independent of CTCF and resides in the first 250bp of the entire 1.2 kb region (Recillas-Targa *et al*., 2002). This activity requires a protein called USF1 that has been shown to localize at the 5'HS4 locus too (West *et al*., 2004).

Similarly, many boundary elements like the gypsy, Fab7, etc. in the fly, HML, HMR and tRNA-thr gene in yeast, tRNA gene in humans and several others have been identified. More recently, whole genome epigenetic profile and computational tools have been used for genome-scale predictions of boundary elements (Srinivasan and Mishra, 2012). An approximate estimate would suggest that boundaries may represent as much as 5% of the genome in higher eukaryotes. This meets the need of preventing long range or local misregulation by enhancers that are capable of driving any accessible promoter and restricting them only to legitimate target promoters (Mishra, 2014).

Boundary elements apart from regulating gene expression also help organize the genome with the help of boundary associated factors like CTCF, CP190, cohesion, etc. These proteins are, thus, also called architectural proteins. For instance, a recent study reported that CTCF is highly enriched in long-range interactions between transcription start sites (TSSs) and distal regulatory elements throughout ENCODE pilot regions spanning 1% of the human genome. The entire genome is organized into topologically associating domains (TADs) that are regions having high intra-domain long-range-

interactions. The boundaries of TADs have a high binding of CTCF, again indicating toward its role in genome packaging. Similarly, it has been seen in *Drosophila* that these TAD boundaries are also enriched in other proteins known to have boundary-associated functions like BEAF, CP190 and cohesin. This gives an indication that insulator-elements and insulator proteins do much more than boundary function and are infact the "genome organizers" (Gibcus and Dekker, 2013; Van Bortle *et al*., 2012).

Pseudogenes

Another significant contribution to non-coding DNA is by the pseudogenes that are disabled copies of genes that have lost their ability to code for proteins. Pseudogenes are predicted to be 10,000 to 20,000 in number in human genome (Torrents *et al*., 2003; Zhang *et al*., 2010). Pseudogenes are present in a wide range of species, including *Arabidopsis* (Benovoy and Drouin, 2006), *Drosophila* (Harrison *et al*., 2003) and *C. elegans* (Harrison *et al*., 2001). Pseudogenes can be classified into three types – unitary, processed and unprocessed. Unitary pseudogenes are those that have lost their ability for both transcription and translation. They are <100 in number in the human genome (Zhang *et al*., 2010). Processed pseudogenes lack non-coding sequences of introns from them indicating that they may have formed via reverse transcription of the original processed mRNA and got inserted in the genome. Processed pseudogenes have polyA tails and direct repeats at either end of the pseudogene (Maestre *et al*., 1995). The presence of direct repeats hints that the retrotransposition of mRNAs into the genome could have been mediated by LINE elements (Esnault *et al*., 2000). These pseudogenes can get transcribed if their integration occurs close to another promoter (Zheng *et al*., 2007). The third category contains unprocessed pseudogenes that arise by gene duplication events resulting in two or more copies of the same gene in the genome. This effectively creates a neutral gene copy that is free to accumulate mutations without affecting cell survival. This often leads to what is termed as 'neofunctionalization' of genes in which this neutral copy, after accumulating a set of mutations, transforms into a gene with a function entirely different from its original counterpart. However, the copies those fail to neofunctionalize or are still in that process get termed as pseudogenes (Brosius and Gould, 1992). This process has been one of the major driving forces for evolution and diversification of gene expression. While until recently these pseudogenes were considered to be genomic fossils, of late, numerous studies have reported their potential functions in the genome making them an important component of the genome. Many pseudogenes still retain their ability to get transcribed and these transcripts have the ability to influence the expression and regulation of its function copy in the genome. For example, overexpression of transcript of Oct4P1 – a pseudogene of Oct4, a pluripotency-associated transcription factor, leads to inhibition of mesenchymal stem cell differentiation and stimulating proliferation (Lin *et al*., 2007). Recently it was shown that transcript from the pseudogene can effectively act as a decoy for sequestering away repressive miRNAs and prevent them from silencing the expression from the functional copy (Poliseno *et al*., 2010). Many pseudogenes get processed into siRNAs. These siRNAs have been shown to be useful in keeping under check the mRNA levels of the original copy. An example of this is the HDAC1 gene that has several pseudogenes. Many miRNAs are produced from these pseudogenes and upon knocking-out Dicer, the level of HDAC1 was seen to get upregulated indicating that the siRNAs derived from the pseudogenes helped in regulation of HDAC1 gene expression (Tam *et al*., 2008). Although we are far from knowing the significance of the entire repertoire of pseudogenes, the functions we know today indicate that a large number of pseudogenes are not genomic fossils and are rather actively being used by the genome to aid in its regulation.

Conserved Non-Coding Sequences (CNCS)

CNCS are sequences that are >95% conserved over millions of years of evolution. One of the first such large-scale identification was done for human chromosome 21 (Frazer *et al*., 2001). CNCS make around 5.5% of the human genome. A subset of these shows 100% conservation across species. These are called Ultra-Conserved Elements (UCEs). A genome-

wide search identified 481 such genomic segments that are longer than 200bp and are 100% conserved between human, mouse and rat genomes. These elements are highly conserved (>95%) even among other species like chicken and fugu (Bejerano *et al.*, 2004). Out of these 481 ultraconserved elements, 111 map to protein-coding regions (including UTRs) and are transcribed. Many such elements are found near to developmentally regulated genes emphasizing their non-random presence in the genome (Sabarinadh *et al*., 2004; Sandelin *et al*., 2004). For example, a group of three such elements was found just downstream to the HoxD complex in mammals. The elements – CR 1, 2 and 3 – show 100% identity across vertebrates (Sabarinadh *et al*., 2004). The same region when compared in the shark genome was seen to be shorter indicating that additional unique sequences have been accumulated and conserved during evolution. The functional significance of UCEs was revealed when many of them were shown to function as regulatory elements. One such study reports that 45% of the 167 tested human elements functioned as tissue-specific enhancers in mouse transgenic assays (Pennacchio *et al*., 2006). In a later study, it was shown that 93% of the 481 identified UCRs were found to be transcribed at least in one of the various cancer tissues analyzed (Calin *et al*., 2007). Many of them were transcribed from both the strands. These transcripts showed distinct signatures in various cancers. In fact one of the tested UCR transcript, which was upregulated in colorectal cancer was found to increase the number of malignant cells by reducing apoptosis (Calin *et al*., 2007). This study clearly indicates the functional significance of these UCRs. Strangely, however, deletion of four such UCRs gave mice that were not only viable but also lacked any specific phenotype (Ahituv *et al*., 2007). This would imply that while such sequences are highly evolutionarily constrained and contribute to genome regulation in a variety of manners, they are not themselves necessary for survival.

Introns

Most eukaryotes have a split-gene system wherein the protein-coding exons are interrupted by sequence termed as intragenic regions or introns. Two groups discovered introns in adenovirus when they observed that hybrids of genomic DNA and mRNA of a gene showed up single stranded regions of DNA sequence in between (Berget *et al*., 1977; Chow *et al*., 1977). While most prokaryotes are devoid of introns, almost all eukaryotes – both single-celled and multicellular – have introns in their genes. The proportion of introns, however, varies in different genomes to the extent that they make up almost 25% of the human genome (Gregory, 2005).

Introns can be of four types based on their distribution and mechanism of splicing. Group I introns are mainly found in bacteria, phages, viruses, organelle genomes. These introns are usually found in rRNA and tRNA genes and are rarely found in protein-coding genes (Hausner *et al*., 2014). They are spliced out from the host mRNAs by a self-splicing mechanism, thus essentially categorizing them as ribozymes. Most of the Group I introns encode an endonuclease that aids in the mobility of these introns. These introns are capable of moving from its original position to an identical location into an intronless allele, a process termed as 'intron homing' (Dujon, 1989). Group II introns are found in the genomes of fungal and plant mitochondria, chloroplasts and eubacteria. These are self-splicing in nature and have conserved secondary structure consisting of stem-loop structures. Many Group II introns code for reverse transcriptase that is responsible for their retrotransposition and insertion into intronless alleles (Curcio and Belfort, 1996). Phylogenetic analysis shows that these introns are very similar to the LINE L1 transposons (Xiong and Eickbush, 1990). Archaeal genomes have introns that are found in tRNA genes. Many of them have this insertion just one nucleotide before the anticodon making its splicing out inevitable for the functioning of the tRNA, while others have insertions at places that do not affect the overall structure of the tRNA. The removal of these introns requires assistance from tRNA splicing endonucleases and ligases (Reviewed in Yoshihisa, 2014). Finally, the best-known introns are the spliceosomal introns, mainly found in the nuclear genomes of eukaryotes. They interrupt almost 90% of the protein-coding genes. Not being under stringent selection pressure, the introns have accumulated numerous mutations. The recognition

sites at the splice junctions that are essential for their splicing out, however, are conserved. These introns are spliced-out by large riboprotein complexes called the spliceosomes which follow a mechanism of splicing similar to that of groupII introns (Yoshihisa, 2014).

Introns, especially the spliceosomal introns, contribute to gene regulation in a variety of means. In addition to containing snoRNAs that are involved in the maturation of other rRNAs, introns are also a source of many small regulatory RNAs like miRNAs, snoRNAs, piRNAs and siRNAs (Rearick *et al*., 2011). Introns also house many regulatory elements like promoters and enhancers (Oshima *et al*., 1990; Pankov *et al*., 1994). The most important feature of introns is their role in bringing about complexity by alternative splicing. An extreme case of diversity achieved by alternative splicing is that of the *Drosophila* Down Syndrome Cell Adhesion Molecule (DSCAM) gene that can give 38000 isoforms (Schmucker *et al*., 2000). Alternative splicing illustrates the usefulness of non-coding introns in the context of complexity of higher eukaryotes as they add significantly to the complexity of their proteomes.

Genomic Regions/Elements Transcribing Regulatory RNAs

A large variety is seen in ncRNAs and they are found to occupy a major portion of the genomic space. Studies in the past two decades have indicated that this genome-wide transcription often produces regulatory RNAs contributing by diverse mechanisms to gene expression and regulation. These regulatory RNAs fall into two major categories on the basis of their average sizes – the small non-coding RNAs (sncRNAs) and the long non-coding RNAs (lncRNAs).

sncRNA

The sncRNAs constitute microRNAs (miRNAs), small interfering RNAs (siRNAs), PIWI interacting RNAs (piRNAs), etc.The sncRNAs are known mainly to repress gene expression except in a few cases where dsRNA has been shown to activate genes by a process termed RNA activation whose mechanistic details still remain to be elucidated (Portnoy *et al*., 2011).

miRNA: miRNAs are single-stranded RNA molecules that are approximately 21 or 22 nucleotides long. They are generated from hairpin-loop containing primary transcripts (pri-miRNAs) by the action of Drosha RNaseIII in the nucleus. miRNAs were the first class of sncRNAs that were discovered. They mainly repress gene expression by means of post-transcriptional gene silencing. The mature miRNA associates with ribonucleoprotein complex called the RNA Induced Silencing Complex (RISC) that consists of Argonaut and the Dicer proteins. The complex aids the miRNA to target mRNA to which it goes and hybridizes and causes the degradation of target mRNA. miRNAs can also cause gene silencing by inhibiting translation by preventing ribosome complex formation (Bartel, 2004).

siRNA: siRNAs are also 21-23 nt long RNAs that, like the miRNAs, silence gene expression by hybridizing to target mRNAs and causing their degradation via the RISC complex. While for miRNAs usually each pre-miRNA gives one mature miRNA, a single transcript can generate many siRNAs. And while siRNAs need perfect complementarity to target mRNAs, miRNAs require complementarity only 6-8 nt for seed pairing. A class of siRNAs is derived from repetitive sequences. These repeat associated siRNAs (rasiRNAs) perform the important function of keeping the heterochromatin regions of the genome transcriptionally inactive (Elbashir *et al*., 2001; Hammond *et al*., 2001).

piRNA: piRNAs were first identified through studies on the *Drosophila* Stellate locus, which is composed of repeated copies of a gene encoding a casein kinase II β-subunit homologue (Livak, 1990). piRNAs are 26-31nt long RNAs derived from clusters that make up almost 1% of the human genome. Most of the piRNAs map to transposons and other repeat elements and loss of piRNAs causes upregulation of transcripts derived from these transposons indicating that their major function is to keep the transposons inactive. They carry out this silencing with the aid of special proteins called the PIWI proteins. These

RNAs also bring about silencing via the RNA Induced Silencing Complex and thereby destruction of the target RNAs (Khurana and Theurkauf, 2010).

lncRNA

lncRNAs are > 200 nucleotides in length and, interestingly, resemble mRNAs in many of their characteristics, e.g., transcribed by RNA polymerase II, mostly 5' capped, spliced and polyadenylated at the 3' end. lncRNAs are functionally divided into 3 types – structural, repressive and activating. They can perform structural roles by providing a scaffold for the formation of paraspeckles or other structural features like the nuclear matrix (Pathak *et al*., 2013; Sasaki *et al*., 2009). There are several repressive lncRNA mediated functions such as recruiting repressive complexes at the target loci, causing transcriptional interference, allosterically modifying RNA binding proteins to subsequently inhibit transcription or preventing the formation of transcription initiation complex at the target loci. At the translational level lncRNAs can also degrade mRNAs and prevent protein synthesis (Martianov *et al*., 2007; Nagano *et al*., 2008; Wang *et al*., 2008). More recently, even activating role of lncRNAs have been reported which involves either preventing silencing or promoting activation (Bertani *et al*., 2011; Cesana *et al*., 2011; Krishnan and Mishra, 2014; Wang *et al*., 2011; Yang *et al*., 2011).

lncRNA has been the area of active research in recent past and several examples of such RNAs are now well studied. These include Xist and rox in dosage compensation in mammals and fruit flies, respectively, (Deng and Meller, 2006; Maenner *et al*., 2012), hsrω in stress response and variety of other regulatory processes (Jolly and Lakhotia, 2006; Lakhotia *et al*., 2012; Mallik and Lakhotia, 2009; Onorati *et al*., 2011) and several others reviewed recently (Fatica and Bozzoni, 2014; Krishnan and Mishra, 2014; Kung *et al*., 2013; Lakhotia, 2012; Mercer *et al*., 2009; Orom *et al*., 2010; Ponting *et al*., 2009). Considering that genomic representation of the lncRNA is still increasing, it is likely that it may constitute as much, if not more, as the protein coding part of the complex genomes.

Repetitive Non-coding Sequences

Transposable Elements

~50% of the human genome is composed of repetitive non-coding sequences. This portion of the genome comprises mainly of transposons, retrotransposons, and simple sequence repeats (Gregory, 2005). Transposons form the most abundant class of non-genic DNA making up almost 45% of the human genome, which include the Long Interspersed Nuclear Elements (LINEs), Short Interspersed Nuclear Elements (SINEs), Long Terminal Repeats (LTRs) and DNA transposons.

The LINE-1 (L1) is a 6 kb element that contains an internal RNA PolII responsive promoter, two open reading frames (ORFs) and a polyadenylation signal. The ORFs code for RNA-binding, endonuclease and reverse transcriptase proteins that make the element self-reliant or autonomous for transposition (Swergold, 1990). There are more than 500,000 copies of L1 elements that make up 17% of the human genome. However, only about a 100 of them are active at present. Alu and SVA (SINE-VNTR-Alu) elements are SINE elements that total to ~13% of the human genome. Alu elements are typically 300bp long and are dimers of monomers derived from 7SLRNA gene that are transcribed from an internal polIII promoter. These are primate specific elements and B1 elements are their counterparts in mouse (Batzer and Deininger, 2002; Deininger, 2011). SVA elements are hominid-specific and are ~2 kb in length each composed of an hexamer repeat region, an Alu-like region, a variable number of tandem repeats region, a HERV-K10-like region and a polyadenylation signal ending with an oligo dA-rich tail of variable length. It is not known as to which polymerase transcribes these elements, as neither polII nor polIII promoters have been detected in these elements. These elements, like Alu, are non-autonomous in transposition and likewise depend on L1 element machinery for their mobility (Ostertag *et al*., 2003; Wang *et al*., 2005).

LTRs are retrotransposons that have Long Terminal repeats on both the ends. Their size ranges from a few hundred base pairs to 25 kb. The long terminal repeats themselves vary considerably in size

from a few hundred base pairs to more than 5 kb, start with 5'-TG-3' and end with 5'-CA-3', and contain the promoters and terminators associated with transcription. They are similar to retroviruses except that they lack a functional *env* (envelope) gene. They can be further divided into three types based on sequence homology – Ty1-copia-like (Pseudoviridae), Ty3-gypsy-like (Metaviridae), and BEL-Pao-like groups. First identified in *Saccharomyces* and *Drosophila*, Ty1/Copia LTR retrotransposons are widespread in higher plants and vertebrates. Based on the divergence of their reverse transcriptase sequence, Ty1/Copia group represents the most ancient lineage of LTR retrotransposons (Wicker *et al*., 2007).

While most of the transposable elements have been rendered inactive due to accumulation of many mutations over time, interestingly, many of them have also acquired regulatory functions in the genome. Some indications for this came when it was observed that upon receiving stress like heat shock viral invasion or heavy metal poisoning massively induce SINE and LINE transcription suggesting that these elements could be playing role in stress response (Farkash and Luning Prak, 2006). Mouse B1 SINEs can also act as "methylation centers" and methylation spreading from human Alu elements has been implicated in silencing tumor suppressor genes (Arnaud *et al*., 2000). There are examples also where transposons have direct role in gene regulation. L1 and Alu elements can introduce new splice sites thus contributing to transcriptional diversity (Sorek *et al*., 2002). Many transposon elements are a source of small regulatory RNAs like piRNAs. In *Drosophila*, TE-encoded piRNAs are required to establish a gradient of maternal Nanos mRNA transcripts in the early *Drosophila* embryo (Rouget *et al*., 2010). In mammals, DNA methylation is established on the Rasgrf1 gene in the paternal germline and this requires TE-encoded piRNAs (Watanabe *et al*., 2011). There are also examples where many genes are coordinately regulated owing to their association with a common regulatory transposon element. A study shows that there is a small but distinct set of cells in mouse ES cell and iPS cell population that is similar to 2-cell stage embryo in which the blastomeres are totipotent. It was observed that these cells expressed a set of transcripts that initiated at LTR elements indicating some role of these elements in cell-fate regulation (Macfarlan *et al*., 2012). Interestingly, many of the genes have their origins in transposable elements. These include *Daysleeper* in *Arabidopsis thaliana* (Bundock and Hooykaas, 2005), *Tramp* (*Zbed1*) (Esposito *et al*., 1999), *Buster1-3, Zbed4* (Smit and Riggs, 1996) and *P52Ripk* (Gale *et al*., 1998) in mammals, and *Mustang*, a family of genes derived from transposase gene of Mutator transposon in angiosperms (Cowan *et al*., 2005). Certain SVA elements are capable of 3' transduction during transposition. A study shows that the *AMAC* gene was duplicated three times in the human genome through an SVA-mediated transduction event, creating a hybrid SINE-pseudogene. Surprisingly, two copies of the retroposed *AMAC* gene can be actively transcribed in different human tissues. It has been hypothesized that the SVA element could be acting as a promoter and bringing tissue-specific expression of a pseudogene (Xing *et al*., 2006). These observations establish the functional relevance of transposons, which is beginning to emerge from recent studies.

Satellite DNA

The term satellite DNA was given to the secondary band of DNA that used to separate in density gradient centrifugation of eukaryotic genomic DNA. These are sequences that are repeated in tandem in the genome. Satellite DNA can be classified on the basis of the length of the repeat sequence – satellites, which are repeat sequences that range from 5bp up to hundreds of base pairs in length; minisatellites, which are repeats of 6-~200 bp long sequences; and microsatellites that are repeats of 1-6 bplong DNA sequence. Satellites are usually found in telomeric and centromeric regions. D4Z4 is one such example, which is a 3.3kb unit repeated 11-100 times. Changes in repeat number of many satellite DNAs is associated with certain disorders and diseases. In this case, a decrease in number of the D4Z4 repeat at the subtelomeric region of 4q region to below 11 causes facioscapulohumeral muscular dystrophy (FSHD), which is an autosomal dominant disease characterized by progressive wasting and weakness of the facial,

shoulder and upper arm muscles (van Deutekom *et al*., 1993). Interestingly, while most satellites are non-coding in nature, each D4Z4 repeat at this region encodes for the double homeobox protein 4 (DUX4), a putative germ line transcription factor, which consequently has a role to play in the FSHD disorder (Geng *et al*., 2012). Other well-characterized/known satellites are the satellite III, which has the GGAAT repeat, the α-satellite, which has 171bp repeat motif and β-satellite. Satellite III is present in the centromeres of most of the human chromosomes and is thought to be its functional component (Blackburn, 1984). The α-satellite DNA is found in the centromeres and binds CENP-B protein on a 17bp motif within the satellite called the CENP-B box (Muro *et al*., 1992). In fission yeast, the binding of this protein on the satellite modifies histones and promotes heterochromatin formation (Nakagawa *et al*., 2002). The prototype of β-satellite has a 68bp repeat motif and is present mainly on chromosome 9 and other acrocentric chromosomes, viz., 13, 14, 15, 21 and 22 (Waye and Willard, 1989). While not much is known about the function of these repeats, an 18-copy sequence of this repeat that has got inserted into the transmembrane serine protease TMPRSS3 causes its disruption and thereby an autosomal disorder leading to deafness (Scott *et al*., 2001). Satellite DNA is also transcribed in many organisms. Some α-satellites transcripts have been detected in zebrafish embryos (Li and Kirby, 2003). In *Drosophila*, the satellite DNA on Y-chromosome loops produce transcripts in spermatocytes (Bonaccorsi *et al*., 1990). Satellite DNA is also a source of many small RNAs. In fission yeast, centromeric repeats generate 20-22 nt siRNAs that with heterochromatin protein HP1 modify the chromatin to form repressed heterochromatin (Schramke *et al*., 2005).

Minisatellites are defined as repeats of 6-100bp long motifs that can span from 0.5kb to several kilobases in the genome. They are also known as VNTRs for Variable Number of Tandem Repeats. Sequence comparisons have hinted a close similarity between minisatellites and the χ sequence (GCTGTGG) of λ-phage. Minisatellite was discovered in 1980 when they noticed a very high degree of polymorphism at a single locus (Wyman and White, 1980). They are routinely used as markers for genotyping owing to their high variability within populations. Alec Jeffreys' group developed a PCR-based method to use some of the hypervariable loci to perform genotyping studies (Jeffreys *et al*., 1991). VNTRs are also transcriptional regulators. Members of NFκB family of transcription factors can bind to repeat sequences that are present downstream to the HRAS gene and activate its transcription (Trepicchio and Krontiris, 1992). The insulin gene also has an associated minisatellite with a 14bp repeat unit whose repeat number is proportional to the susceptibility of insulin-dependent diabetes mellitus. Higher repeat numbers cause increase in insulin transcript in thymus thereby influencing the levels insulin-specific T-cells (Lucassen *et al*., 1993; Pugliese *et al*., 1997). In some cases a VNTR may be present with the coding region giving rise to a polymorphic peptide sequence. D4DR repeat is a 48bp or a 16-aminoacid repeat that shows polymorphism in population. In a population-based study it was shown that irrespective of ethnicity or gender, 7-repeat allele was strongly associated with novelty-seeking behavior (Ebstein *et al*., 1996). Apart from their role in genome organization and gene regulation, satellite DNA, especially minisatellites, are a rich repertoire of polymorphisms within populations. Thus they have been exploited as DNA markers for various applications.

Microsatellites, also known as Simple Tandem Repeats (STR) or Simple Sequence Repeats (SSRs), are clusters of 1-6 nt long motif repeats. These repeats are found in most genomes, both vertebrate and invertebrate, and are placed all over the genome. 3% of the human genome is made up of SSRs (Gregory, 2005). They are highly variable in nature, which is mainly due to variation in the repeat number rather than in their primary sequence. This indicates their near-neutral evolution. In fact, SSRs are among the fastest-evolving DNA sequences with high mutation rates: 10^{-2}–10^{-3} per locus per gamete per generation (Weber and Wong, 1993), which leads to their high polymorphism in terms of repeat number. However, most of these repeats are found in non-coding regions rather than in coding thus showing some bias in their selection across the genome (Metzgar *et al*., 2000). Many prokaryotic genomes are also rich in SSRs.

Comprehensive analyses of ~370 prokaryotic genomes revealed a bias in the enrichment of SSRs depending on whether the microbe is pathogenic. It was seen that SSRs composed of short monomers (1-4bp length) are often found in host-adapted pathogens that are not known to readily survive in a natural environment outside the host. On the other hand, SSRs with longer monomers (5-11bp length) are found mostly in non-pathogens. Even in eukaryotes, certain repeats are more prevalent in genome than others. For example, in human, among the four dinucluceotide SSRs, $(CA)_n$ is much more abundant than $(GC)_n$. Contrastingly, $(AT)_n$ is most abundant in plant genomes (Lagercrantz *et al*., 1993). Also, in some genomes, these repeats are often found near to transposon elements again supporting their non-random occurrence (Ramsay *et al*., 1999). Such observations point towards their positive selection in the genome (Lander *et al*., 2001). And indeed quite a number of reports support the notion that SSRs may be functional entities in the genome. One of the first reports used *Drosophila* as a model system and studied the role of a coding SSR in the *period* gene involved in circadian rhythm maintenance. The 17-copy repeat of the SSR coding for Thr-Gly is found mainly in southern Mediterranean and gives a circadian period of ~24 hours when the climate is warm (29°C) and a shorter period when the temperature drops to 18°C. The most prevalent $(Thr\text{-}Gly)_{20}$ allele that makes up 90% of the population, however, showed no significant difference in the circadian period between the two temperatures tested. Thus the difference in repeat number seems to be having a direct influence on the function of the protein and thereby the circadian cycle. Thus, $(Thr\text{-}Gly)_{20}$ is mainly found in regions with huge temperature variations so as to keep the circadian period constant despite temperature variation while, on the other hand, $(Thr\text{-}Gly)_{17}$ is seen in regions with lesser temperature variations (Sawyer *et al*., 1997). SSRs present not only in coding regions but also those in non-coding regions can affect gene expression. TG repeats were the first for which reporter assays predicted putative enhancer-like functions (Hamada *et al*., 1984). In another study, the vasopressin 1a receptor (V1aR) has been shown to influence social behavior in different rodent species (Young *et al*., 1997). This species-specificity in V1aR expression pattern and thereby the rodent behavior has been shown to be regulated by differences in a microsatellite in its 5' regulatory region. This microsatellite, which consists of GA repeats, is highly expanded in pro-social prairie and pine voles, while in the asocial montane and meadow voles it has shorter repeats (Hammock and Young, 2005). The longer GA repeats induce higher expression of luciferase gene in reporter assays. This may be due to the binding of GAGA Associated Factor, which is known to bind to GA repeat motif in *Drosophila* and murine cells, and depending on context also can enhance transcriptional output of a gene (Mahmoudi *et al*., 2002; Srivastava *et al*., 2013; van Steensel *et al*., 2003). Such repeats that enhance gene expression have also been seen within introns of protein-coding genes. One such example is of a tetranucleotide, TCAT, in the intron of the Tyrosine Hydroxylase gene (Meloni *et al*., 1998).

SSRs have been shown to be influencing many other cellular features. Certain repeats like AATGG have the ability to take up unusual DNA structures like hairpins (Catasti *et al*., 1999). Some repeats can contribute to formation of fragile sites as in the case of fragile-X syndrome. GAA repeat has also been shown to form loop-like triplex structures the formation of which have been shown to affect gene expression in reporter assays (Fabregat *et al*., 2001). DNA recombination is another process that is affected by the presence of microsatellite sequences and consequently they have been identified as hotspots for recombination (Jeffreys *et al*., 1998). One of the reasons is that some dinucleotide repeats are known to bind to recombination enzymes thus making them target these sequences (Biet *et al*., 1999). In RecA-dependent recombination, regions with high GC/GT repeats prevent complete strand exchange giving rise to recombinant alleles (Dutreix, 1997). Long stretches of these repeats can also cause the polymerase to slip and cause change in repeat number in the daughter strand, a process called replication slippage (Levinson and Gutman, 1987). It can lead to either increase or decrease in the repeat number leading to allelic variations. However, most-studied and well-documented reports of role of SSRs have been those

of triplet-expansion disorders. These are diseases caused by expansion of SSRs (trinucleotide-SSRs) beyond a threshold level. Fragile-X syndrome, Huntington's disease, Spinocerebellar ataxia 8 and Myotonic dystrophy are few of the around twenty such disorders known in humans (Orr and Zoghbi, 2007).

Genome-wide analysis of all 501 SSRs in human genome has revealed several interesting features of these sequences (Subramanian *et al*., 2003b). For example, 23 SSRs showed a bell-shaped enrichment curve showing enrichment of high repeat numbers. Another feature of SSRs is their enrichment near transcriptional start sites (TSS). 60% and 20% CCG and ACG, respectively, were found within 1kb of TSS in humans. A similar trend was seen in chimpanzee and mouse genomes. Interestingly, 34.2% and 72.4% of ACG and CCG repeat elements, respectively, overlapped with the predicted CpG islands. The presence of ACG and CCG near TSS and their overlap with CpG islands indicates the potential regulatory function that rests with these repeats (Ramamoorthy *et al*., 2014; Subramanian *et al*., 2003a). Further, detailed study on GATA/AGAT repeat revealed that it could function as an enhancer-blocker in both human cells and *Drosophila* (Kumar *et al*., 2013).These findings indicate that SSRs have the potential to perform cellular processes by playing a role in gene regulation and genome organization. We also saw how presence of SSRs in coding and non-coding regions can have different types of effects. While studying repeat sequences, in general, has been a challenging job, these examples do tell us how important this part of the genome is and the necessity to study them in more depth.

Conclusions and Perspectives

Genomes are full of a variety of elements that together bring about genome regulation and organization. We now know in greater details about 80% of the non-coding genome, which consists of pseudogenes, CNCS, transposable elements and repetitive DNA. Studies over the past three decades have been able to decipher to an appreciable extent how these different kinds of non-coding elements could perform such functions. Thus, the genome can be pictured as consisting of a small fraction as protein coding sequences and a much higher proportion of non-coding part that consists of regulatory sequences. There is a clear trend of increase of genome size when going from simple to complex organism, though it is not the same with number of genes (Table 1). Gene number goes up only little over 2 fold from *Neurospora* to human while the genome size goes up about 100 fold. Similarly, number of genes in the organisms from *Drosophila* to human is very similar (ranging from 1 to 1.5 fold) the genome size goes over 20 fold. The fact that number of genes has not shown an increase with evolution of complexity as is seen with the amount of non-coding DNA, it may not be incorrect to suggest that the non-coding sequences are being used as tools for evolving complex gene regulatory mechanisms and, thereby, complex organisms. One of the most interesting features of the eukaryotic genomes is their pervasive transcription. 85% of the human genome gets transcribed and number of these RNAs originate from known loci and are predicted to have functions.

A striking picture that comes out of number of studies during the past decade is the dual role played by regulatory elements-structural and functional. The genome-wide interaction data suggests that many of

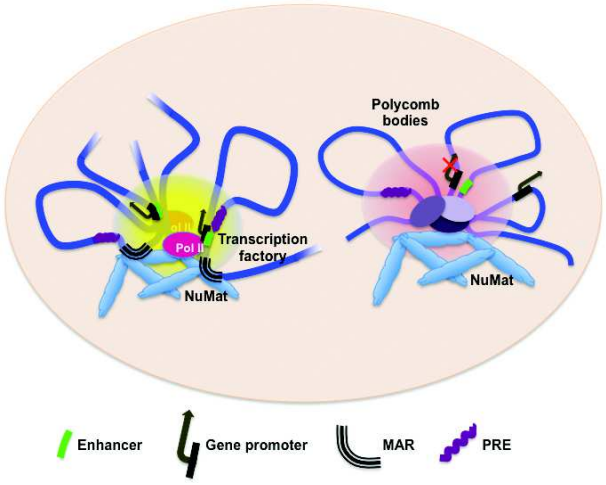

Fig. 3: Organization of active and repressed compartments in the nucleus. The schematic shows the active compartment – transcription factory and the inactive compartment – polycomb body. The DNA sequences that help form such compartments are the matrix-associated regions (MARs) and the scaffold is the nuclear matrix on to which the chromatin is organized

these cis-regulatory elements contact each other in long-range to bring about their function (Fig. 3). Data from genome-wide mapping of various histone marks, DNaseI hypersensitive sites and numerous transcription factors have indicated that a much larger proportion of the genome, than what was anticipated, is involved in one of the tested activities. For example, enhancers interact with their target promoters that may be several megabases away in the genome; or enhancers interact among themselves too. Similarly, boundaries and PREs also interact with each other and with enhancers and/or promoters as exemplified by the bithorax complex. Therefore, apart from functionally regulating the genome, these non-coding sequences also structurally organize the chromatin. Furthermore, with the help of similar DNA elements the chromatin is sequestered into active and inactive compartments in the nucleus.

Acknowledgements

Authors thank CSIR for financial assistance through NWP Genesis and EpiHeD.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio L A and Rubin EM (2007) Deletion of ultraconserved elements yields viable mice *PLoS Biology* **5** e234

Arnaud P, Goubely C, Pelissier T and Deragon J M (2000) SINE retroposons can be used in vivo as nucleation centers for de novo methylation*Mol Cell Biol* **20** 3434-3441

Banerji J, Rusconi S and Schaffner W (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences *Cell* **27** 299-308

Bantignies F and Cavalli G (2006) Cellular memory and dynamic regulation of polycomb group proteins *Curr Opin Cell Biol* **18** 275-283

Bartel D P (2004) MicroRNAs: genomics, biogenesis, mechanism, and function *Cell* **116** 281-297

Batzer M A and Deininger P L (2002) Alu repeats and human genomic diversity *Nat Rev Genet* **3** 370-379

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent W J, Mattick J S and Haussler D (2004) Ultraconserved elements in the human genome *Science* **304** 1321-1325

Benovoy D and Drouin G (2006) Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome *Journal of Molecular Evolution* **62** 511-522

Berget S M, Moore C and Sharp P A (1977) Spliced segments at the 5' terminus of adenovirus 2 late Mrna *Proceedings of the National Academy of Sciences of the United States of America* **74** 3171-3175

Bernstein B E, Mikkelsen T S, Xie X, Kamal M, Huebert D J, Cuff J, Fry B, Meissner A, Wernig M, Plath K *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells *Cell* **125** 315-326

Bertani S, Sauer S, Bolotin E and Sauer F (2011) The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin *Mol Cell* **43** 1040-1046

Biet E, Sun J and Dutreix M (1999) Conserved sequence preference in DNA binding among recombination proteins: an effect of ssDNA secondary structure *Nucleic Acids Res* **27** 596-600

Blackburn E H (1984) The molecular structure of centromeres and telomeres *Annu Rev Biochem* **53** 163-194

Blastyak A, Mishra R K, Karch F, and Gyurkovics H (2006) Efficient and specific targeting of Polycomb group proteins requires cooperative interaction between Grainyhead and Pleiohomeotic *Mol Cell Biol* **26** 1434-1444

Bonaccorsi S, Gatti M, Pisano C and Lohe A (1990) Transcription of a satellite DNA on two Y chromosome loops of Drosophila melanogaster *Chromosoma* **99** 260-266

Bracken A P, Dietrich N, Pasini D, Hansen K H and Helin K (2006) Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions *Genes Dev* **20** 1123-1136

Brand A H and Perrimon N (1993) Targeted gene expression as a means of altering cell fates and generating dominant phenotypes *Development* **118** 401-415

Brosius J and Gould S J (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA" *Proceedings of the National Academy of Sciences of the United States of America* **89** 10706-10710

Brown J L, Grau D J, DeVido S K and Kassis J A (2005) An Sp1/KLF binding site is important for the activity of a Polycomb group response element from the Drosophila engrailed gene *Nucleic Acids Res* **33** 5181-5189

Brown J L, Mucci D, Whiteley M, Dirksen M L and Kassis J A (1998) The Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1 *Mol Cell* **1** 1057-1064

Bundock P and Hooykaas P (2005) An Arabidopsis hAT-like transposase is essential for plant development *Nature* **436** 282-284

Calin G A, Liu C G, Ferracin M, Hyslop T, Spizzo R, Sevignani C, Fabbri M, Cimmino A, Lee E J, Wojcik S E *et al.* (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas *Cancer Cell* **12** 215-229

Catasti P, Chen X, Mariappan S V, Bradbury E M and Gupta G (1999) DNA repeats in the human genome *Genetica* **106** 15-36

Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A and Bozzoni I (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA *Cell* **147** 358-369

Chen X, Hiller M, Sancak Y and Fuller M T (2005) Tissue-specific TAFs counteract Polycomb to turn on terminal differentiation *Science* **310** 869-872

Chow L T, Gelinas R E, Broker T R and Roberts R J (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA *Cell* **12** 1-8

Chung J H, Whiteley M and Felsenfeld G (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila *Cell* **74** 505-514

Consortium E P (2012) An integrated encyclopedia of DNA elements in the human genome *Nature* **489** 57-74

Cowan R K, Hoen D R, Schoen D J and Bureau T E (2005) MUSTANG is a novel family of domesticated transposase genes found in diverse angiosperms *Molecular Biology and Evolution* **22** 2084-2089

Curcio M J and Belfort M (1996) Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA *Cell* **84** 9-12

Deaton A M and Bird A (2011) CpG islands and the regulation of transcription *Genes Dev* **25** 1010-1022

Deininger P (2011) Alu elements: know the SINEs *Genome Biology* **12** 236

Deng X and Meller V H (2006) Non-coding RNA in fly dosage compensation *Trends Biochem Sci* **31** 526-532

Dujon B (1989) Group I introns as mobile genetic elements: facts and mechanistic speculations—a review *Gene* **82** 91-114

Dutreix M (1997) (GT)n repetitive tracts affect several stages of RecA-promoted recombination *Journal of Molecular Biology* **273** 105-113

Ebstein R P, Novick O, Umansky R, Priel B, Osher Y, Blaine D, Bennett E R, Nemanov L, Katz M and Belmaker R H (1996) Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of Novelty Seeking *Nature Genetics* **12** 78-80

Elbashir S M, Harborth J, Lendeckel W, Yalcin A, Weber K and Tuschl T (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells *Nature* **411** 494-498

Ernst J and Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome *Nature Biotechnology* **28** 817-825

Esnault C, Maestre J and Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes *Nature Genetics* **24** 363-367

Esposito T, Gianfrancesco F, Ciccodicola A, Montanini L, Mumm S, D'Urso M and Forabosco A (1999) A novel pseudoautosomal human gene encodes a putative protein similar to Ac-like transposases *Human Molecular Genetics* **8** 61-67

Fabregat I, Koch K S, Aoki T, Atkinson A E, Dang H Amosova O, Fresco J R, Schildkraut C L and Leffert H L (2001) Functional pleiotropy of an intramolecular triplex-forming fragment from the 3'-UTR of the rat Pigr gene *Physiol Genomics* **5** 53-65

Farkash E A and Luning Prak E T (2006) DNA damage and L1 retrotransposition *Journal of Biomedicine & Biotechnology* **2006** 37285

Fatica A and Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development *Nat Rev Genet* **15** 7-21

Frazer K A, Sheehan J B, Stokowski R P, Chen X, Hosseini R, Cheng J F, Fodor S P, Cox D R and Patil N (2001) Evolutionarily conserved sequences on human chromosome 21 *Genome Research* **11** 1651-1659

Gale M, Jr Blakely C M, Hopkins D A, Melville M W, Wambach M, Romano P R and Katze M G (1998) Regulation of interferon-induced protein kinase PKR: modulation of P58IPK inhibitory function by a novel protein, P52rIPK *Mol Cell Biol* **18** 859-871

Geng L N, Yao Z, Snider L, Fong A P, Cech J N, Young J M, van der Maarel S M, Ruzzo W L, Gentleman R C, Tawil R *et al.* (2012) DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy *Developmental Cell* **22**

38-51

Gibcus J H and Dekker J (2013) The hierarchy of the 3D genome *Mol Cell* **49** 773-782

Goldberg M L (1979) *PhD Diss In Stanford Univ*

Gregory T R (2005) Synergy between sequence and size in large-scale genomics *Nat Rev Genet* **6** 699-708

Hamada H, Seidman M, Howard B H and Gorman C M (1984) Enhanced gene expression by the poly(dT-dG)poly(dC-dA) sequence *Mol Cell Biol* **4** 2622-2630

Hammock E A and Young L J (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits *Science* **308** 1630-1634

Hammond S M, Caudy A A and Hannon G J (2001) Post-transcriptional gene silencing by double-stranded RNA *Nat Rev Genet* **2** 110-119

Han J, Kim D and Morris K V (2007) Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells *Proceedings of the National Academy of Sciences of the United States of America* **104** 12422-12427

Harrison P M Echols N and Gerstein M B (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome *Nucleic Acids Res* **29** 818-830

Harrison P M Milburn D, Zhang Z, Bertone P and Gerstein M (2003) Identification of pseudogenes in the Drosophila melanogaster genome *Nucleic Acids Res* **31** 1033-1037

Hausner G, Hafez M and Edgell D R (2014) Bacterial group I introns: mobile RNA catalysts *Mobile DNA* **5** 8

Jeffreys A J, MacLeod A, Tamaki K, Neil D L and Monckton D G (1991) Minisatellite repeat coding as a digital approach to DNA typing *Nature* **354** 204-209

Jeffreys A J, Murray J and Neumann R (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot *Mol Cell* **2** 267-273

Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K and Felsenfeld G (2009) H33/H2AZ double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions *Nature Genetics* **41** 941-945

Jolly C and Lakhotia S C (2006) Human sat III and Drosophila hsr omega transcripts: a common paradigm for regulation of nuclear RNA processing in stressed cells pp 5508-5514

Kahn T G, Stenberg P, Pirrotta V and Schwartz Y B (2014) Combinatorial interactions are required for the efficient recruitment of pho repressive complex (PhoRC) to polycomb response elements *PLoS Genetics* **10** e1004495

Khurana J S and Theurkauf W (2010) piRNAs, transposon silencing, and Drosophila germline development *The Journal of Cell Biology* **191** 905-913

Kim B C, Kim W Y, Park D, Chung W H, Shin K S and Bhak J (2008) SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions *BMC Bioinformatics* **9** *Suppl 1* S2

Kim T K, Hemberg M, Gray J M, Costa A M, Bear D M, Wu J, Harmin D A, Laptewicz M, Barbara-Haley K, Kuersten S *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers *Nature* **465** 182-187

Krishnan J and Mishra R K (2014) Emerging trends of long non-coding RNAs in gene activation *FEBS J* **281** 34-45

Kumar R P, Krishnan J, Pratap Singh N, Singh L and Mishra R K (2013) GATA simple sequence repeats function as enhancer blocker boundaries *Nature Ccommunications* **4** 1844

Kung J T, Colognori D and Lee J T (2013) Long noncoding RNAs: past, present, and future *Genetics* **193** 651-669

Kurokawa R (2011) Promoter-associated long noncoding RNAs repress transcription through a RNA binding protein TLS *Advances in Experimental Medicine and Bbiology* **722** 196-208

Kutach A K and Kadonaga J T (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters *Mol Cell Biol* **20** 4754-4764

Kvon E Z, Kazmar T, Stampfel G, Yanez-Cuna J O, Pagani M, Schernhuber K, Dickson B J and Stark A (2014) Genome-scale functional characterization of Drosophila developmental enhancers in vivo *Nature* **512** 91-95

Lagercrantz U, Ellegren H and Andersson L (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates *Nucleic Acids Res* **21** 1111-1115

Lagrange T, Kapanidis A N, Tang H, Reinberg D and Ebright R H (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB *Genes Dev* **12** 34-44

Lakhotia S C (2012) Long non-coding RNAs coordinate cellular responses to stress *Wiley Interdiscip Rev RNA* **3** 779-796

Lakhotia S C, Mallik M, Singh A K and Ray M (2012) The large noncoding hsromega-n transcripts are essential for thermotolerance and remobilization of hnRNPs, HP1 and RNA polymerase II during recovery from heat shock in Drosophila *Chromosoma* **121** 49-70

Lander E S, Linton L M, Birren B, Nusbaum C, Zody M C, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.* (2001) Initial sequencing and analysis of the human genome *Nature* **409** 860-921

Lanzuolo C and Orlando V (2012) Memories from the polycomb group proteins *Annual Review of Genetics* **46** 561-589

Lee D H, Gershenzon N, Gupta M, Ioshikhes I P, Reinberg D, and Lewis B A (2005a) Functional characterization of core promoter elements: the downstream core element is recognized by TAF1 *Mol Cell Biol* **25** 9674-9686

Lee N, Maurange C, Ringrose L and Paro R (2005b) Suppression of Polycomb group proteins by JNK signalling induces transdetermination in Drosophila imaginal discs *Nature* **438** 234-237

Levine M (2010) Transcriptional enhancers in animal development and evolution *Curr Biol* **20** R754-763

Levinson G and Gutman G A (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12 *Nucleic Acids Res* **15** 5323-5338

Li Y X and Kirby M L (2003) Coordinated and conserved expression of alphoid repeat and alphoid repeat-tagged coding sequences *Developmental Dynamics : an Official Publication of the American Association of Anatomists* **228** 72-81

Lim C Y, Santoso B, Boulay T, Dong E, Ohler U and Kadonaga J T (2004) The MTE, a new core promoter element for transcription by RNA polymerase II *Genes Dev* **18** 1606-1617

Lin H, Shabbir A, Molnar M and Lee T (2007) Stem cell regulatory function mediated by expression of a novel mouse Oct4 pseudogene *Biochemical and Biophysical Research Communications* **355** 111-116

Liu Y, Shao Z and Yuan G C (2010) Prediction of Polycomb target genes in mouse embryonic stem cells *Genomics* **96** 17-26

Livak K J (1990) Detailed structure of the Drosophila melanogaster stellate genes and their transcripts *Genetics* **124** 303-316

Lobanenkov V V, Nicolas R H, Adler V V, Paterson H, Klenova E M, Polotskaja A V and Goodwin G H (1990) A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene *Oncogene* **5** 1743-1753

Lucassen A M, Julier C, Beressi J P, Boitard C, Froguel P, Lathrop M and Bell J I (1993) Susceptibility to insulin dependent diabetes mellitus maps to a 41 kb segment of DNA spanning the insulin gene and associated VNTR *Nature Genetics* **4** 305-310

Macfarlan T S, Gifford W D, Driscoll S, Lettieri K, Rowe H M, Bonanomi D, Firth A, Singer O, Trono D and Pfaff S L (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity *Nature* **487** 57-63

Maenner S, Muller M and Becker P B (2012) Roles of long, non-coding RNA in chromosome-wide transcription regulation: lessons from two dosage compensation systems *Biochimie* **94** 1490-1498

Maestre J, Tchenio T, Dhellin O and Heidmann T (1995) mRNA retroposition in human cells: processed pseudogene formation *The EMBO Journal* **14** 6333-6338

Mahmoudi T, Katsani K R and Verrijzer C P (2002) GAGA can mediate enhancer function in trans by linking two separate DNA molecules *The EMBO Journal* **21** 1775-1781

Mallik M and Lakhotia S C (2009) The developmentally active and stress-inducible noncoding hsromega gene is a novel regulator of apoptosis in Drosophila *Genetics* **183** 831-852

Martianov I, Ramadass A, Serra Barros A, Chow N and Akoulitchev A (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript *Nature* **445** 666-670

Meloni R, Albanese V, Ravassard P, Treilhou F and Mallet J (1998) A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro *Human Molecular Genetics* **7** 423-428

Mercer T R, Dinger M E and Mattick J S (2009) Long non-coding RNAs: insights into functions *Nat Rev Genet* **10** 155-159

Metzgar D, Bytof J and Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA *Genome Research* **10** 72-80

Mishra K and Mishra R K (2010) Polycomb Group of Genes and the Epigenetics of Aging *Epigenetics of Aging* 135-150

Mishra R K (2014) Restraining the enhancers from straying *Journal of Biosciences* **39** 739-740

Mishra R K, Mihaly J, Barges S, Spierer A, Karch F, Hagstrom K, Schweinsberg S E and Schedl P (2001) The iab-7 polycomb response element maps to a nucleosome-free region of chromatin and requires both GAGA and pleiohomeotic for silencing activity *Molecular and Cellular Biology* **21** 1311-1318

Mishra R K, Yamagishi T, Vasanthi D, Ohtsuka C and Kondo T (2007) Involvement of polycomb-group genes in establishing HoxD temporal colinearity *Genesis* **45** 570-576

Muller J, Hart C M, Francis N J, Vargas M L, Sengupta A, Wild B, Miller EL, O'Connor M B, Kingston R E and Simon J A (2002) Histone methyltransferase activity of a

Drosophila Polycomb group repressor complex *Cell* **111** 197-208

Muro Y, Masumoto H, Yoda K, Nozaki N, Ohashi M and Okazaki T (1992) Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box*The Journal of Cell Biology***116** 585-596

Nagano T, Mitchell J A, Sanz L A, Pauler F M, Ferguson-Smith A C, Feil R and Fraser P (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin *Science* **322** 1717-1720

Nakagawa H, Lee J K, Hurwitz J, Allshire R C, Nakayama J, Grewal S I, Tanaka K and Murakami Y (2002) Fission yeast CENP-B homologs nucleate centromeric heterochromatin by promoting heterochromatin-specific histone tail modifications *Genes Dev* **16** 1766-1778

Ohtsuki S, Levine M and Cai H N (1998) Different core promoters possess distinct regulatory activities in the Drosophila embryo *Genes Dev* **12** 547-556

Olivier M (2004) From SNPs to function: the effect of sequence variation on gene expression Focus on "a survey of genetic and epigenetic variation affecting human gene expression" *Physiol Genomics* **16** 182-183

Onorati M C, Lazzaro S, Mallik M, Ingrassia A M, Carreca A P, Singh A K, Chaturvedi D P, Lakhotia S C and Corona D F (2011) The ISWI chromatin remodeler organizes the hsromega ncRNA-containing omega speckle nuclear compartments *PLoS Genetics* **7** e1002096

Orom U A, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q *et al.* (2010) Long noncoding RNAs with enhancer-like function in human cells *Cell* **143** 46-58

Orr H T and Zoghbi H Y (2007) Trinucleotide repeat disorders *Annual Review of Neuroscience* **30** 575-621

Oshima R G, Abrams L and Kulesh D (1990) Activation of an intron enhancer within the keratin 18 gene by expression of c-fos and c-jun in undifferentiated F9 embryonal carcinoma cells *Genes Dev* **4** 835-848

Ostertag E M, Goodier J L, Zhang Y and Kazazian H H Jr (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans *American Journal of Human Genetics* **73** 1444-1451

Pankov R, Neznanov N, Umezawa A and Oshima R G (1994) AP-1, ETS, and transcriptional silencers regulate retinoic acid-dependent induction of keratin 18 in embryonic cells *Mol Cell Biol* **14** 7744-7757

Pathak R U, Mamillapalli A, Rangaraj N, Kumar R P, Vasanthi D, Mishra K and Mishra R K (2013) AAGAG repeat RNA is an essential component of nuclear matrix in Drosophila *RNA Biology* **10** 564-571

Pennacchio L A, Ahituv N, Moses A M, Prabhakar S, Nobrega M A, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis K D *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences *Nature* **444** 499-502

Petruk S, Sedkov Y, Johnston D M, Hodgson J W, Black K L, Kovermann S K, Beck S, Canaani E, Brock H W and Mazo A (2012) TrxG and PcG proteins but not methylated histones remain associated with DNA through replication *Cell* **150** 922-933

Phillips J E and Corces V G (2009) CTCF: master weaver of the genome *Cell* **137** 1194-1211

Poliseno L, Salmena L, Zhang J, Carver B, Haveman W J and Pandolfi P P (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology *Nature* **465** 1033-1038

Ponting C P, Oliver P L and Reik W (2009) Evolution and Functions of Long Noncoding RNAs *Cell* **136** 629-641

Portnoy V, Huang V, Place R F and Li L C (2011) Small RNA and transcriptional upregulation *Wiley Interdiscip Rev RNA* **2** 748-760

Pugliese A, Zeller M, Fernandez A, Jr Zalcberg, L J Bartlett, R J Ricordi, C Pietropaolo, M Eisenbarth G S, Bennett S T, and Patel D D (1997) The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes *Nature Genetics* **15** 293-297

Ramamoorthy S, Garapati H S and Mishra R K (2014) Length and sequence dependent accumulation of simple sequence repeats in vertebrates: potential role in genome organization and regulation *Gene* **551** 167-175

Ramsay L, Macaulay M, Cardle L, Morgante M, degli Ivanissevich S, Maestri E Powell W and Waugh R (1999) Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley *The Plant Journal : for Cell and Molecular Biology* **17** 415-425

Rearick D, Prakash A, McSweeny A, Shepard S S, Fedorova L and Fedorov A (2011) Critical association of ncRNA with introns *Nucleic Acids Res* **39** 2357-2366

Recillas-Targa F, Pikaart M J, Burgess-Beusse B, Bell A C, Litt M D, West A G, Gaszner M and Felsenfeld G (2002) Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities *Proceedings of the National Academy of Sciences of the United States of America* **99** 6883-6888

Ringrose L, Rehmsmeier M, Dura J M and Paro R (2003) Genome-wide prediction of Polycomb/Trithorax response elements

in Drosophila melanogaster *Developmental Cell* **5** 759-771

Rouget C, Papin C, Boureux A, Meunier A C, Franco B, Robine N, Lai E C, Pelisson A and Simonelig M (2010) Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo *Nature* **467** 1128-1132

Sabarinadh C, Subramanian S, Tripathi A and Mishra R K (2004) Extreme conservation of noncoding DNA near HoxD complex of vertebrates *BMC Genomics* **5** 75

Sandelin A, Bailey P, Bruce S, Engstrom P G, Klos J M, Wasserman W W, Ericson J and Lenhard B (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes *BMC Genomics* **5** 99

Sasaki Y T, Ideue T, Sano M, Mituyama T and Hirose T (2009) MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles *Proceedings of the National Academy of Sciences of the United States of America* **106** 2525-2530

Sawyer L A, Hennessy J M, Peixoto A A, Rosato E, Parkinson H, Costa R and Kyriacou C P (1997) Natural variation in a Drosophila clock gene and temperature compensation *Science* **278** 2117-2120

Schmucker D, Clemens J C, Shu H, Worby C A, Xiao J, Muda M, Dixon J E and Zipursky S L (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity *Cell* **101** 671-684

Schramke V, Sheedy D M, Denli A M, Bonila C, Ekwall K, Hannon G J and Allshire R C (2005) RNA-interference-directed chromatin modification coupled to RNA polymerase II transcription *Nature* **435** 1275-1279

Schuettengruber B, Martinez A M, Iovino N and Cavalli G (2011) Trithorax group proteins: switching genes on and keeping them active *Nature Reviews Molecular Cell Biology* **12** 799-814

Schwartz Y B, Linder-Basso D, Kharchenko P V, Tolstorukov M Y, Kim M, Li H B, Gorchakov A A, Minoda A, Shanower G, Alekseyenko A A *et al.* (2012) Nature and function of insulator protein binding sites in the Drosophila genome *Genome Research* **22** 2188-2198

Schwartz Y B and Pirrotta V (2007) Polycomb silencing mechanisms and the management of genomic programmes *Nat Rev Genet* **8** 9-22

Scott H S, Kudoh J, Wattenhofer M, Shibuya K, Berry A, Chrast R, Guipponi M, Wang J, Kawasaki K, Asakawa S *et al.* (2001) Insertion of beta-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness *NatureGenetics* **27** 59-63

Simon J, Chiang A, Bender W, Shimell M J, and O'Connor M (1993) Elements of the Drosophila bithorax complex that mediate repression by Polycomb group products *Dev Biol* **158** 131-144

Smale S T and Kadonaga J T (2003) The RNA polymerase II core promoter *Annu Rev Biochem* **72** 449-479

Smit A F and Riggs A D (1996) Tiggers and DNA transposon fossils in the human genome *Proceedings of the National Academy of Sciences of the United States of America* **93** 1443-1448

Sorek R, Ast G and Graur D (2002) Alu-containing exons are alternatively spliced *Genome Research* **12** 1060-1067

Srinivasan A and Mishra R K (2012) Chromatin domain boundary element search tool for Drosophila *Nucleic Acids Res* **40** 4385-4395

Srivastava S, Puri D, Garapati H S, Dhawan J and Mishra R K (2013) Vertebrate GAGA factor associated insulator elements demarcate homeotic genes in the HOX clusters *Epigenetics & Chromatin* **6** 8

Subramanian S, Mishra R K and Singh L (2003a) Genome-wide analysis of Bkm sequences (GATA repeats): predominant association with sex chromosomes and potential role in higher order chromatin organization and function *Bioinformatics* **19** 681-685

Subramanian S, Mishra R K and Singh L (2003b) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions *Genome Biology* **4** R13

Swergold G D (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter *Mol Cell Biol* **10** 6718-6729

Tam O H, Aravin A A, Stein P, Girard A, Murchison E P, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz R M *et al.* (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes *Nature* **453** 534-538

Torrents D, Suyama M, Zdobnov E and Bork P (2003) A genome-wide survey of human pseudogenes *Genome Research* **13** 2559-2567

Trepicchio W L and Krontiris T G (1992) Members of the rel/NF-kappa B family of transcriptional regulatory proteins bind the HRAS1 minisatellite DNA sequence *Nucleic Acids Res* **20** 2427-2434

Udvardy A, Maine E and Schedl P (1985) The 87A7 chromomere Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains *Journal of Molecular Biology* **185** 341-358

Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi J E and Corces V G (2012) Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains *Genome Research* **22** 2176-2187

van Deutekom J C, Wijmenga C, van Tienhoven E A, Gruter A M, Hewitt J E, Padberg G W, van Ommen G J, Hofker M H and Frants R R (1993) FSHD associated DNA rearrangements are due to deletions of integral copies of a 32 kb tandemly repeated unit *Human Molecular Genetics* **2** 2037-2042

van Steensel B, Delrow J and Bussemaker H J (2003) Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding *Proceedings of the National Academy of Sciences of the United States of America* **100** 2580-2585

Visel A, Blow M J, Li Z, Zhang T, Akiyama J A, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers *Nature* **457** 854-858

Wang H, Xing J, Grover D, Hedges D J, Han K, Walker J A and Batzer M A (2005) SVA elements: a hominid-specific retroposon family *Journal of Molecular Biology* **354** 994-1007

Wang K C, Yang Y W Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie B R, Protacio A, Flynn R A, Gupta R A *et al*. (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression *Nature* **472** 120-124

Wang X, Arai S, Song X, Reichart D, Du K, Pascual G, Tempst P, Rosenfeld M G, Glass C K and Kurokawa R (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription *Nature* **454** 126-130

Watanabe T, Tomizawa S, Mitsuya K, Totoki Y, Yamamoto Y, Kuramochi-Miyagawa S, Iida N, Hoki Y, Murphy P J, Toyoda A *et al.* (2011) Role for piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus *Science* **332** 848-852

Waye J S and Willard H F (1989) Human beta satellite DNA: genomic organization and sequence definition of a class of highly repetitive tandem DNA *Proceedings of the National Academy of Sciences of the United States of America* **86** 6250-6254

Weber J L and Wong C (1993) Mutation of human short tandem repeats *Human Molecular Genetics* **2** 1123-1128

West A G, Huang S, Gaszner M, Litt M D and Felsenfeld G (2004) Recruitment of histone modifications by USF proteins at a vertebrate barrier element *Mol Cell* **16** 453-463

Whyte W A, Orlando D A, Hnisz D, Abraham B J, Lin C Y, Kagey M H, Rahl P B, Lee T I and Young R A (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes *Cell* **153** 307-319

Wicker T, Sabot F, Hua-Van A, Bennetzen J L, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O *et al.* (2007) A unified classification system for eukaryotic transposable elements *Nat Rev Genet* **8** 973-982

Woo C J, Kharchenko P V, Daheron L, Park P J, and Kingston R E (2010) A region of the human HOXD cluster that confers polycomb-group responsiveness *Cell* **140** 99-110

Wyman A R and White R (1980) A highly polymorphic locus in human DNA *Proceedings of the National Academy of Sciences of the United States of America* **77** 6754-6758

Xing J, Wang H, Belancio V P, Cordaux R, Deininger P L and Batzer M A (2006) Emergence of primate genes by retrotransposon-mediated sequence transduction *Proceedings of the National Academy of Sciences of the United States of America* **103** 17608-17613

Xiong Y and Eickbush T H (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences *The EMBO Journal* **9** 3353-3362

Yang L, Lin C, Liu W, Zhang J, Ohgi K A, Grinstein J D, Dorrestein P C and Rosenfeld M G (2011) ncRNA- and Pc2 Methylation-Dependent Gene Relocation between Nuclear Structures Mediates Gene Activation Programs *Cell* **147** 773-788

Yoshihisa T (2014) Handling tRNA introns, archaeal way and eukaryotic way *Frontiers in Genetics* **5** 213

Young L J, Winslow J T, Nilsen R and Insel T R (1997) Species differences in V1a receptor gene expression in monogamous and nonmonogamous voles: behavioral consequences *Behavioral Neuroscience* **111** 599-605

Zhang Z D, Frankish A, Hunt T, Harrow J and Gerstein M (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates *Genome Biology* **11** R26

Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo S W, Lu Y, Denoeud F, Antonarakis S E, Snyder M *et al.* (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution *Genome Research* **17** 839-851.