Indian Journal of Biochemistry & Biophysics
Vol. 33, April 1996, pp. 93-102

Minireview

RNA polymerase II dependent genes that do not code for protein

S C Lakhotia

Cytogenetics Laboratory, Department of Zoology, Banaras Hindu University, Varanasi 221 005

Received 14 August 1995; accepted 21 December 1995

In recent years more and more examples of RNA polymerase II dependent non-coding transcripts have been described. Although these have frequently been ignored as "selfish DNA elements", it is becoming increasingly clear that many, if not all, of them have very important biological roles. Examples of such "genes" from *Drosophila*, mammals, other vertebrates, yeast etc. are considered. Although the specific mechanisms through which these non-coding transcripts function in the cell are not clear, comparisons reveal certain common themes, particularly the importance of secondary structures, rather than the primary base sequence of these transcripts. While some of these transcripts may function as ribozymes or as anti-sense regulators, most others may function more directly through their specific protein-binding properties. Since RNA is believed to be the first "living" molecule, it is very likely that some genes even today function only through this class of molecules. It is expected that instead of being ignored as examples of "selfish DNA", a more positive search for their functions will help unravel the significance of this novel class of genes.

Introduction

Eukaryotes have three principal RNA polymerases, each earmarked for transcribing specific class of genes. The RNA polymerase I is responsible for transcribing DNA sequences coding for the major ribosomal RNAs while the RNA polymerase III is used to transcribe the 5S ribosomal RNA, the various tRNAs and the small nuclear RNAs (snRNA). The transcripts produced by RNA polymerases I and III are not translated but are involved in processing and translating the transcripts made by the RNA polymerase II which transcribes all the protein-coding genes dispersed throughout chromosomal DNA. The RNA polymerase II transcribed genes share many features in common with regard to their transcriptional regulation, post-transcriptional processing of the primary transcript (heterogeneous nuclear RNA or hnRNA) and finally the transport of the processed messenger RNA (mRNA) to the cytoplasm for translation into their specific polypeptide products. The central dogma of molecular biology has ascribed such an important role to the translational activities of mRNA that any RNA polymerase II transcribed gene is expected to have a protein product. Those that seemed to not code for a protein have often been put aside as "selfish DNA"[1,2]. However, over the years more and more instances of genes that are transcribed by RNA polymerase II, whose products may display typical post-transcriptional processing events and are yet not coding for any protein, have been discovered in diverse organisms. This review will consider some of the known examples of "non-coding" genes.

Y-Chromosome loops or the fertility genes in *Drosophila* males

Ever since the classical studies of Bridges[3] on sex-determination in *Drosophila*, it has been known that the Y-chromosome is not essential for sex-determination but is essential for fertility of male flies. However, whole of the Y-chromosome is condensed, heterochromatic and genetically "inert" and, therefore, its essential role in male fertility remained enigmatic although conventional genetic studies did identify a number of Y-linked "fertility genes" having specific effects on spermiogenesis[4]. Meyer and his group[5,6] showed that during the primary spermatocyte stage in *D. hydei*, the Y-chromosome opens up into very large microscopically visible and transcriptionally very active "lampbrush" loops which are essential for normal differentiation of spermatids into mature motile spermatozoa. In *D. hydei*, six distinctive loops with characteristic morphology and transcript patterns are present while in *D. melanogaster*, the Y-chromosome loops in primary spermatocytes are not so distinct but are comparable in their general organization to that in *D. hydei*[7].

Y-Linked mutations that affect fertility have a complex relationship with the Y-chromosome associated loops. The most remarkable feature of the Y-chromosome linked fertility genes and the lampbrush chromosomes is the enormous size of the transcription units and the nature of DNA associated with them: the Y-chromosomal transcription units in primary spermatocytes are as large as 4000 kb in *D. melanogaster* and between 260 to 1500 kb in *D. hydei*[8,9]. Bulk of these huge transcription units is comprised of simple, satellite and of functional or degenerating transposable elements. None of these have any substantial open reading frames[8-12]. The only conventional protein coding gene so far known to be present on the *D. melanogaster* Y-chromosome is the one that produces a male-specific beta-heavy chain of dynein microtubule motor[13]. The *kl-5* complementation unit of the Y-chromosome, at which the dynein gene maps, is estimated to be about 1300 kb long while the dynein transcript is estimated to be only about 14 kb[10,11,13]. The physical arrangement of the dynein mRNA coding region with rest of the transcription unit is not well understood: it has been speculated that the dynein-coding region may actually be outside of the *kl-5* transcription unit[11] or may be scrambled in its 1300 kb region[10].

All the known Y-chromosomal transcription units, active only during the primary spermatocyte stage, share a few common enigmatic features. These include (i) production of very large sized transcripts that remain restricted to the nucleus, (ii) the transcription units comprise essentially of simple repetitive and retrotransposon elements which show remarkable sequence diversity between related species, (iii) the base sequences of these transcription units are full of stop codons in all possible reading frames and thus do not seem to code for any protein (except for the above noted dynein gene) and (iv) the various transcripts bind to specific proteins and are responsible for the characteristic shape and size of the different Y-chromosome lampbrush loops[8,10,14-18]. It appears that during the evolution of these male fertility genes, many retrotransposons got incorporated within the lampbrush transcription units; interestingly such insertions do not appear to have affected the functioning of these genes. Thus the male fertility gene Q in the lampbrush loops called "Nooses" has a large number of gypsy retroposons; sequence analysis of these gypsy elements reveals that all of them are truncated and have lost those sequences that may interfere with transcriptional continuity along the loop[18]. Transcription of these retroposons is required but not sufficient for the function of fertility genes like the Q gene (Nooses) of *D. hydei*[17,18]. It is also interesting to note that the many tandem repeats of retroposons that are present in the loop transcription units show identical sense orientation although outside the loop boundary the same retroposons show random arrangement[19].

What is the function of these unusually large and apparently untranslatable RNAs? The Y-linked fertility genes and the lampbrush loops are essential for spermiogenesis to proceed normally. The discovery of at least one protein-coding genes (dynein) on the Y-chromosome of *D. melanogaster* has encouraged hopes of finding more such protein-coding genes associated with the other fertility genes; such hopes have further encouraged suggestions that the huge transcription units comprising of satellite sequences and transposons, have evolved because the fertility genes are "the ultimate heaven for selfish genetic elements"[10]. However, an alternative and a challenging hypothesis suggests these transcripts to have more of structural and/or regulatory role by providing substrates to which other macromolecules may bind[9].

Other non-coding genes expressed in germ line cells of male *Drosophila*

Besides the above considered Y-chromosome loops, a few other non-coding "genes" active during spermatogenesis in *Drosophila* are also noteworthy. One of the them is the enigmatic "crystal (or *Suppressor of Stellate*)-*Stellate*" set of genes. *D. melanogaster* males without a Y-chromosome (X/O males) show presence of needle-like or star-like inclusions in their primary spermatocytes[5], a high level of non-disjunction of meiotic chromosomes, abnormal distribution of organelles in meiocytes and events of meiotic drive[20]. A Y-chromosome linked locus, the *crystal* or *cry*[21] (also named *Suppressor of Stellate*, *Su(Ste)*[22]) and an X-linked locus, the *Stellate* or *Ste*[20], are involved in this phenotype of X/O male *D. melanogaster* flies. These two X- and Y-linked loci contain arrays of partially homologous, tandemly repeated sequences[23]. The normal functioning of *cry* or *Su(Ste)* locus requires a critical number of subunits rather than physical integrity of the whole array of tandem repeats[20,24]. The *Ste* allele of *Stellate* locus carries about 200 copies of the repeat unit while the *Ste*$^+$ allele has only about 20 copies of the repeat units; in the absence of *Su(Ste)* gene on the Y-chromosome, the *Ste*$^+$ allele causes production of needle-like inclusions in

spermatocytes while the *Ste* allele causes star-like inclusions[20]. The *Stellate* locus codes for a protein which shares homology with the beta subunit of casein kinase II[22]. The *Su(Ste)* locus on the Y-chromosome suppresses activity of the *Stellate* locus on the X-chromosome so that in XY individuals, only a few stellate transcripts are made and even these are not properly spliced and processed but in the absence of Y-chromosome or in the presence of Y-chromosome deficient for the *Su(Ste)* locus, the level of stellate transcripts in spermatocytes is significantly elevated resulting in the formation of needle- or the star-like inclusions and to consequently affect disjunction and segregation of homologous chromosomes and cytoplasmic organelles like mitochondria[22,25]. The *Su(Ste)* locus besides sharing homology with the *Ste* locus, also shares sequence homology with the He-T family of transposons[26]. The 2800 bp repeat unit of the *Su(Ste)* locus has been shown to consist of (i) a region of homology to the *Ste* locus, (ii) a Y-specific AT-rich segment and (iii) a mobile element 1360 inserted in the *Ste* sequence[27]. It is not known if the *Su(Ste)* transcripts code for a protein but this appears unlikely in view of its structure. The suppression of *Ste* activity in the presence of *Su(Ste)* has been suggested[22] to be due to a mutual competition for a limiting set of transcription and splicing factors so that in the absence of *Su(Ste)*, these will be available to *Ste* for an abundant production of the stellate protein. This mutual inactivation of these two genes has also been considered an intriguing type of selfish genetic system[22] since related species of *Drosophila* do not have *Ste* or *Su(Ste)* loci. However, the important point to note is that in this case also, the *Su(Ste)* gene has very vital effect on the organism without possibly having a typical protein product.

Another gene active in the male germ line of *D. melanogaster*, but apparently not having a protein product, is the recently identified[28] *Mst* 40 sequence (*Male specific transcript* 40) located on the section 40 of the right arm of chromosome 2: the *Mst* 40 sequence is organized as tandemly arrayed 1.4 kb repeat unit with transcripts limited to nuclei in male germ line; the longest possible open reading frame is 43 amino acid long without homology to any known polypeptides; this gene was detected in all strains of *D. melanogaster* tested although other species of *Drosophila* did not show its presence[28]. A function for this gene is yet to be found.

Among the diverse variety of transposons that make up the bulk of Y-chromosome in *Drosophila* is the *micropia* family. These retroposons are present both on the autosomes, X-chromosome and the Y-chromosome of *D. hydei* as well as *D. melanogaster*. Lankenau *et al*.[29] have reported that in meiotic cells of male *D. hydei*, in addition to full length transcripts, *micropia* also encodes antisense transcripts complementary to the reverse transcriptase and RNAse H coding region; these antisense transcripts are present only in male germ line since they are produced from a testes specific promoter. Furthermore, while most of the sense transcripts are present as part of giant RNA molecules because of their location on the lampbrush loop forming sites, the antisense micropia transcripts are 1.0 and 1.6 kb long[29]. It appears that these full length antisense transcripts have an important role in the control of *micropia* encoded reverse transcriptase protein in male germ cells[29].

The *93D* or the *HSR ω* gene of *Drosophila*

One of the first non-protein coding genes to be characterized in some detail is the *93D* or the *hsr ω* gene of *D. melanogaster* and its homologues in other species of *Drosophila*. Transcriptionally, this is one of the most active genes following heat shock[30,31]. That this gene was different from other heat shock genes was revealed by its singular inducibility in polytene cells treated with a variety of agents like benzamide, colchicine etc. (see reviews in ref. 32-34). Since no new proteins were induced when this locus was selectively activated by benzamide, Lakhotia and Mukherjee[35] suggested that this gene does not code for any protein; this was confirmed when this gene and its homologues in other species were cloned and sequenced[36-39]. Sequence analysis revealed a remarkably conserved overall organization of this gene but equally remarkable divergence in the base sequence in different *Drosophila* species. In all species, the locus spans more than 10 kb and includes two exons and an intron in the proximal 1.9 to 2.0 kb followed (on the 3′ end) by a long stretch of tandem arrays of repeat units unique to this locus. The base sequence of the unique as well as of the repeat units is not strongly conserved between species except for certain small regions at the exon-intron junctions and for a 9 bp motif in the repeat units[39,40]. Although the repeat units do not share homology between different species, all the repeats in a tandem array in a species are highly homogeneous and maintain a certain minimum and maximum length[40]. The *hsr ω* locus produces two primary nuclear transcripts of > 10 kb and ~ 1.9 kb size, respectively;

the 1.9 kb transcript (spanning the two exons and one intron) is spliced to produce the 1.2 kb poly(A) cytoplasmic transcript. The 1.2 kb transcript has only one very short ORF but the amino acid sequence coded by this ORF is not conserved in different species[38,39,41,42]. The relative abundance of the three transcripts produced by this locus depends upon the nature of inducer[43,44]. Unlike most genes, the spliced out intron in this case is highly stable[43,44]. Besides being induced by the different inducers, this locus is also developmentally active in most tissue types of embryo, larva, pupa and adult[45] and although without a protein coding function, it is essential for survival of flies[46] and also for development of normal thermo-tolerance[32]. Hsp 83 is known to bind to the heat shock induced 93*D* locus[47] and recent observations in our laboratory suggest that the lethality due to deficiency of this locus is enhanced in HSP83 mutant heterozygotes[48].

A series of studies in our laboratory (reviewed in refs 32-34) showed that the 93*D* locus affected synthesis and/or turnover of the hsp 70 and $\alpha\beta$ repeat (heat induced but non-translatable, see below) transcripts. The rates of synthesis and/or turnover from the site of synthesis of the 87A-type and 87C-type hsp70 and the $\alpha\beta$ transcripts varied in relation to the specific profile of the 93D transcripts present or synthesized in response to a given condition of heat shock[49]. In this context, it is interesting to note that the five copies of hsp 70 genes in *D. melanogaster*, present at the 87A (2 copies) and 87C (3 copies) sites, share nearly identical coding and the 5′ upstream regulatory regions but their 3′ untranslated regions (3′ UTR) show considerable divergence. Sharma and Lakhotia[49] suggested that the differing 3′ UTRs may target the hsp70 mRNAs to different cellular compartments and that the 93D transcripts have a role in this process.

It is obvious that the 93*D* locus in *D. melanogaster* and its homologue in other species have important functions to perform in almost all tissue types during normal development as well as under various conditions of cellular stress[33,34]. It has been suggested that one of the functions of the cytoplasmic 1.2 kb transcript is to "monitor the health" of translational machinery while the nuclear >10 kb transcript may be involved in synthesis and turnover/transport of other transcripts like the hsp70[34,49].

The $\alpha\beta$ repeats of *D. melanogaster*

This is an interesting family of repetitive sequences that is present at several locations in the genome of *D. melanogaster* with some of them being heat shock inducible. The 87C locus, site for 3 copies of HSP70 genes, also harbours about 10-14 copies of $\alpha\beta$ repeats[50]. In addition, the $\alpha\beta$ repeats are also present at the heterochromatic chromocentre but these are not heat inducible. Those at the 87C site are heat inducible due to their being associated with sequence elements (the γ elements) that are identical to the *hsp* 70 promoter region[50,51]. The $\alpha\beta$ units that are immediately downstream of the 'γ' elements at the 87C site are induced by heat shock to produce multiple poly-A$^+$ transcripts of 2.5, 1.8, 1.4 and 1.1 kb sizes[52]. None of these appear to code for any protein[52-54]. The suggestion that these sequences are one more example of "selfish DNA" has gained support from observations that deletion of these sequences from the 87C site has no deleterious effect and that a sibling species, *D. simulans*, does not carry any heat inducible $\alpha\beta$ repeats at the 87C or at any other site[53,55]. Nevertheless as noted above, a series of studies in our laboratory has shown that the non-protein coding heat shock locus at 93D has specific effect on transcription of the $\alpha\beta$ sequences during heat shock[32-34,49,56]. Colchicine treatment which induces transcription at the 93D locus (see above) also leads to an increase in the level of $\alpha\beta$ transcripts at the 87C site[49]. Significance of these interactions remains unknown.

Non-protein-coding genes in mammals

In recent years, a number of genes associated with specific functions but apparently not coding for any protein have been identified in different mammalian genomes. These are briefly considered below.

The Xist gene

All mammals show inactivation of one of the two X-chromosomes in somatic cells of females to achieve dosage compensation of X-chromosome linked genes in males and females[57]. While the paternal X-chromosome is preferentially inactivated in Marsupials, the inactivation of one of the two X-chromosomes in eutherians is generally random in different somatic cells but once inactivated, the same X-chromosome continues to remain inactive in all cell generations. This inactivation affects condensation, transcription and replication of the entire chromosome and is apparently regulated by a single cis-acting centre, the X-inactivation centre (termed *Xic* in human and *Xce* in mouse), which is not only responsible for the initiation of inactivation but also for its spread to

the entire chromosome length (reviews in refs 58-60). A major breakthrough in understanding of this whole chromosome inactivation process was the cloning and characterization of a human as well as mouse gene that appeared to correspond to the *Xic* or *Xce*[61-64]. This gene, termed *XIST* (human) or *Xist* (mouse) for *X inactive specific transcript*, has attracted considerable attention not only for its remarkable role in inactivation of a whole chromosome but also for the way it achieves this role. The human Xist mRNA is 17kb long while the mouse Xist mRNA is 15 kb long but none of these appears to code for any protein[62,64] and in both cases, the transcripts are made only by the inactive X-chromosome, the allele on the active X-chromosome remaining completely silent. Xist transcripts are exclusively nuclear and appear to remain associated with the Barr body, which represents the inactive X-chromosome in interphase nuclei[62]. Evidence for involvement of Xist transcripts in inactivation of the X-chromosome appears complete since the appearance of these transcripts shows an absolute parallel with the pattern of X-inactivation. In mouse, humans and other eutherians, the first sign of X-inactivation is seen in extra-embryonic trophectoderm and primitive endoderm lineages with exclusive inactivation of the paternally derived X-chromosome in all cells[65]; X-inactivation in embryonic lineages occurs later and this is random with respect to the parental origin of the X-chromosome. Paternal X-chromosome derived Xist transcripts are first seen at 4-cell stage of female embryo prior to X-inactivation[66,67]. Specific inactivation of the paternal X-chromosome in the earliest stages of embryo correlates with the specific patterns of imprinting of paternal and maternal X-chromosomes: during spermatogenesis, the Xist locus is demethylated, passed on to the zygote in a hypomethylated state and, therefore, poised for transcriptional activity while the maternally derived Xist allele is fully methylated at this stage of embryonic development[68]. At a somewhat later stage of embryonic development in eutherians, the paternal imprinting is lost and Xist gene of one of the two homologues is remethylated randomly and this sets the stage for random inactivation of one of the two X-chromosomes[68-70]. It is notable that during meiosis in male mammals, the X-chromosome is inactivated and at this time its Xist gene is active. Therefore, Xist transcripts are believed to be responsible for inactivation of the X-chromosome during spermatogenesis in a manner analogous to the X-inactivation in somatic cells of female[71,72].

The mechanism of action of Xist transcripts in initiating inactivation of the X-chromosome from which these are produced is not known. Although Brown *et al.*[62] showed by *in situ* hybridization that in interphase nuclei from female, the Xist transcripts are seen more abundant in vicinity of the Barr body, it was not clear if this indicated binding of these transcripts to the inactive X-chromosome or to the nascent transcripts made by this chromosome. A structural feature of the inactive X-chromosome is worth noting: in interphase nuclei the two telomeric regions of the inactive X-chromosome (Barr body) remain closer[73]; the inactive X-chromosome in metaphase cells also shows a characteristic bend at the Xic locus[74] and in the primary spermatocytes, the inactive X-chromosome shows a similar spatial orientation within the sex-vesicle[71]. This bending at the Xic may facilitate non-homologous chromosome associations leading to heterochromatinization involving heterochromatin-specific proteins[71]. Whether this change in chromosome structure is due to the act of transcription at this locus or due to binding of the Xist transcripts or due to some other factors recruited by the Xist transcripts remains unknown[61,64,67,75]. Buzin *et al.*[76], using quantitative RT-PCR single nucleotide primer extension assay found only about 2000 Xist transcripts per cell and suggested that only models that do not require Xist RNA to cover the entire inactive X-chromosome are compatible with the number of these traanscripts present in a nucleus.

H19

This is another well known example of a non-coding gene in mammals. The *H*19 gene was first identified as a cDNA that was coordinately regulated along with the α-fetoproteins by trans-acting *raf* in murine fetal liver[77]. Subsequent cloning of the human *H*19 homologue[78] and comparison of sequence of the murine and human homologues revealed lack of conservation of the small open-reading frames although the organization of exons and introns was comparable and the base sequence of certain other regions was conserved[78]. These features led to the inference that *H*19 transcripts are not translated but as such function as RNA. In Southern blots, an *H*19 cognate could be detected in monkey, rat, and chicken but not in *Drosophila*[78]. A large proportion of *H*19 RNA in both human and mice cells exists in association with 28S cytoplasmic particle[78]. Subsequent studies showed the *H*19 gene to be imprinted with only the maternal allele normally expressed; *H*19 expression has an interesting regulatory effect on

expression of the adjacent group of imprinted genes[79]. The *insulin-like growth factor* 2 (*Igf*2) gene is immediately upstream of the *H*19 on human chromosome 7 and is expressed only from the paternal allele due to imprinting. A variety of studies have shown that *H*19 expression is specifically responsible for silencing of the neighbouring cis-located genes: since the paternal *H*19 is imprinted (methylated), this promoter is transcriptionally inactive while the non-methylated maternal allele is transcribed. This in turn inhibits the cis-located *Igf*2 and other adjacent genes on the maternal chromosome. These neighbouring genes are transcribed from the paternal chromosome on which the *H*19 allele is inactive[79,80]. In this respect, *H*19 functions in a manner reminiscent of *XIST*: while *XIST* activity inactivates a whole X-chromosome, *H*19 expression silences a nearby imprinted domain[80].

The *H*19 is expressed as abundant, spliced and poly(A) containing transcripts whose levels increase with cellular differentiation but are absent or reduced in several tumours. In agreement with these observations, Hao *et al.*[81] have shown that *H*19 RNA has a tumour-suppressor activity. *H*19 RNA has also been suggested to have important roles in differentiation of embryonic cytotrophoblasts of ectodermal origin[82].

An interesting feature of the *XIST* and *H*19 genes, shared with several other non-coding genes (e.g., the 93*D* locus of *Drosophila*), is the rather high degree of divergence of the base sequence (particularly at the open-reading frames) in spite of the structural organization of the locus being conserved. While in the case of human and mouse *H*19 genes, the exon-intron organization is highly conserved, in the case of *XIST* and *Xist* genes, the number of exons varies, although there is some degree of similarity[79]. The XIST and *H*19 transcripts share similar secondary structures with long energetically favourable stem-loop structures; in both cases, the longest of the stem-loops are present in the regions that show most conserved base sequence and therefore, appear to be functionally important[79]. The mouse and human *Xist*/*XIST* genes share certain short tandem repeats throughout their length at comparable positions although their total numbers vary between the two species. *H*19 transcripts do not have such extensive repeat motifs but both the human and mouse *H*19 transcripts carry 8-10 copies of TGGGGG motif in a short region near the 5′ end[79]. Conservation of these repeat motifis is reminiscent of the conservation of a 9 bp motif in the otherwise divergent repeat units at the 3′ end of the 93*D* locus of *Drosophila*[41]. In all probability such short repeat motifs help in functions of these transcripts perhaps by determining some aspect of the secondary structure of the RNA and/or by providing binding sites for other molecules.

Other examples of non-coding genes in mammals

While the above examples of non-coding transcripts are better known, more cases of non-coding transcripts have been reported from mammals and other organisms in recent years. Some examples of these are briefly considered below.

His-1 *locus in mouse*

The *His*-1 locus in mouse has been reported to be a common site for retroviral insertions leading to myeloid leukemias[83,84]. Viral insertion leads to activation of the *His*-1 locus which produces a 3 kb RNA derived from a gene consisting of 3 exons spanning 6 kb on mouse chromosome 2. This gene is conserved as a single copy gene in vertebrates and *Drosophila*; in mouse it is highly active in transformed myeloid cells but not in the normal cells examined and produces spliced poly(A) RNA which does not have any appreciable open-reading frame[84].

Synapse-associated non-coding RNA in rat

Velleca *et al.*[85] identified a novel synapse-associated RNA, the 7H4 RNA, in rat diaphragm muscle: this transcript is present selectively in association with synapses in the endplate zone of skeletal muscle of rat diaphragm and is upregulated during early postnatal development and after denervation. The 7*H*4 gene is without introns, yet produces 2 different sized transcripts with identical poly-adenylated 3′ ends. Sequence analysis revealed absence of any significant open-reading frames and is, therefore, believed to function through its RNA products[85].

Human UHG for U22 snoRNA

The nucleolus associated small RNAs (snoRNA) are involved in maturation of the 18S ribosomal RNA and are usually produced by processing of intron fragments of protein-coding host genes[86,87]. The U22 snoRNA (earlier called human RNA Y[88,89]) is highly conserved between man and *Xenopus*. A search for its host gene in humans (the *UHG*) whose intron is processed to produce the U22 snoRNA, revealed that the host gene specifies a poly(A) but non-protein coding RNA[89].

IPW

IPW or the imprinted gene in the Prader-Willi syndrome region was identified[90] to map to the smallest deletion overlap corresponding to the Prader-Willi syndrome region on the proximal human chromosome 15q. Only the paternal allele is expressed in lymphoblasts and fibroblasts due to imprinting; the transcript is spliced and poly-adenylated but its longest open-reading frame codes only for 45 amino acids. Therefore, this gene also appears to be making a non-protein coding transcript. It has been suggested that the Prader-Willi syndrome phenotype may be a direct consequence of lack of expression of this gene[90].

Non-translatable repetitive transcripts in eggs

Non-translatable RNA transcripts carrying single copy genomic sequences interspersed with repetitive sequences are commonly present in the cytoplasmic poly(A) RNA fractions in oocytes of sea urchins as well as *Xenopus*[91-94]. These interspersed (ISp) maternal RNAs show a high sequence complexity, are distinct from nuclear as well as embryonic messenger RNAs and are not associated with polysomes[95]. These ISp RNAs are not due to readthrough transcription of oogenetic mRNA promoters like the histone genes[96] but are specific transcripts in their own right[97]; the presence of identical transcripts in somatic cells of late embryos also suggests that these transcripts do not result from readthrough products characteristic of oocytes[95].

A very interesting class of repetitive transcripts in *Xenopus* oocytes is the Xlsirts (Xenopus laevis short interspersed repeat transcripts): these are a family of interspersed repeat RNAs that carry 3 to 13 repeat units of 79 to 81 bp length flanked by unique sequences[98,99]. The repeat units carry a sequence that is very similar to the consensus repeat found in the human *XIST* and mouse *Xist* genes[99]. The Xlsirt transcripts show a characteristic *in situ* localization in growing oocyte: particularly remarkable is their overlapping distribution at the vegetal cortex in stage 3-4 oocytes with the Vg1 and Xcat2 mRNAs (the Vg1 is a TGF-β-like molecule while the Xcat2 is a NANOS-like molecule and both are implicated in axial patterning of early amphibian embryo). Kloc and Etkin[100] showed that destruction of Xlsirts in growing oocytes by microinjected antisense oligonucleotides resulted in delocalization of the Vg1, but not of Xcat2 RNA, suggesting that the Xlsirts help anchor the Vg1 transcripts on the microfilaments at the vegetal pole in oocytes.

meiRNA in fission yeast

Nutrient starvation of diploid fission yeast (*Schizosaccharomyces pombe*) cells triggers them to enter meiosis through a cascade of events initiated by lowering of cAMP levels. One of cAMP regulated gene which is crucial for progression of meiosis in these cells is the *mei2*$^+$ gene[101] which makes an RNA-binding protein. *mei2*$^+$ function is required not only for pre-meiotic DNA replication but also for entry into meiosis I. Watanabe and Yamamoto[102] found that the RNA-binding mei2 protein interacts with 440 and 508 nucleotides long mei2RNA produced by the *sme2*$^+$ gene. Neither of these two transcripts have any appreciable open-reading frame. This interaction was necessary for entry of cells into meiosis I but not for initiation of pre-meiotic DNA synthesis for which a different RNA may associate with the mei2 protein[102,103]. It is notable in this context that meiosis-specific small nuclear RNA and a group of poly(A) RNA (zygRNA) has been described in lily[104]. The meiRNA appears to be different from the meiosis-specific RNAs of lily although it is possible that the zygRNA of lily includes a meiRNA counterpart[102]. It has been suggested[102,103] that the mei2-meiRNA complex may act as a meiosis-specific splicing factor. Whether this turns out to be the case remains to be seen but it is obvious that a non-coding RNA has a direct role in regulating sexual development in fission yeast.

Non-coding genes: selfish DNA or genes with important biological roles?

The belief that a spliced and poly(A) containing RNA must have a protein coding function is so deeply entrenched that for any newly discovered instance where the transcript does not appear to code for a protein, the authors tend to make an apologetic explanation that the RNA may code for a protein after some unknown kinds of editing or alternative splicing events or else the RNA may actually be a product of a selfish DNA. Fortunately, the increasing number of such genes being known in diverse organisms has lent credence to the concept of non-coding transcripts also having a biological role. The non-coding transcripts may function as ribozymes[105,106] or as antisense RNA regulating the activity of other transcripts as in the case of *lin*-4 gene of *C. elegans*[107,108].

In addition to roles as ribozymes or as antisense regulators, it is likely that the non-coding transcripts have more direct roles in cell regulation. Several possibilities exist. A large number of proteins are now known to perform significant bi-

ological roles through RNA-binding[109-114]. Contrary to earlier beliefs, it is now clear that in the ribosomes also, it is the RNA rather than the protein moiety that has the major catalytic activity[115-119]. Thus it remains possible that as in the case of the various RNA polymerase I or III transcription products, binding of RNA polymerase II transcribed non-coding RNAs to proteins may either alter the activity of the protein or may cause the RNA to have some activity about which we still do not know much.

A large number of coding transcripts have 3′ untranslated region (3′ UTR) of varying length; in a few cases these 3′ UTRs may be even longer than the coding region. Recent studies have shown the 3′ UTRs to perform very important roles in either targeting the transcript to specific cell compartments[114,120,121] or in controlling the kinetics of turnover of the mRNA[122] or even in transcriptional activation of own or other gene/s as trans-acting factors[123,124]. The 3′ UTR of alpha-tropomyosin of mice can suppress tumourigenicity[125]. In analogy with the 3′ UTRs, it is possible that the non-coding poly(A) containing transcripts may also carry out a variety of functions in the cell through their protein-binding properties. Certain zinc-finger proteins bind specifically to DNA-RNA hybrids with implications of their biological roles[126]. In many of these cases the binding is dependent upon the secondary structure of RNA rather than its primary base sequence[109]. Since most of the non-coding transcripts show conserved structure although not necessarily the primary base sequence, it is likely that these RNAs perform important biological functions through their structural motifs.

RNA directed *de novo* methylation of specific genomic sequences in plants has been demonstrated[127]. Various RNAs in crude HeLa cell extracts inhibit DNA methylase[128]. It is also known that certain unusual DNA structures favour cytosine methylation[129]. Thus it is possible that RNA-DNA hybrid may direct the DNA methylases[127]. It will be interesting to examine if the XIST RNA mediated inactivation of X-chromosome in mammals is due to some such properties. The RNA-DNA triple helices are stable under physiological conditions *in vitro*[130] and it has been suggested that RNA may regulate gene activity by binding to the major groove of double helix[127,130]. Sequence-specific binding of RNA or ribonucleoprotein to duplex DNA has also been considered to be important in gene regulation[131].

It is obvious that RNA polymerase II dependent non-coding transcripts are no longer mere curiosities or vagaries of the biological diversity. These seem to have established themselves as a distinct class of genes with very important functions. Understanding of the significance of such genes has been thwarted by the common "selfish genetic element" label applied to them. Recent years have witnessed an increasing understanding of the biological significance of heterochromatin[132] in view of which the biological relevance of the so-called "selfish" or "junk" DNA is debatable. It is to be hoped that with an increasing awareness of genetic functions through structural motifs as well rather than through the primary base sequence alone, will stimulate an appreciation and understanding of this interesting class of genes. With RNA being the first "living molecule", it is but to be expected that even today biological systems continue to utilize this versatile molecule directly. Progress in modelling the structure of RNA and RNP molecules will be of considerable help in this direction.

Acknowledgement

Work in my laboratory on the 93*D* heat shock locus of *Drosophila* has been supported by the Department of Science and Technology, Govt. of India (New Delhi) and by the Department of Atomic Energy, Govt. of India (Bombay). I thank Dr. Rajiva Raman for critical comments on the manuscript.

References

1 Doolittle W F & Sapienza C (1980) *Nature, London*, 284, 601-603.
2 Orgel L E & Crick F H C (1980) *Nature, London*, 284, 603-606.
3 Bridges C B (1916) *Genetics*, 1, 1-51.
4 Brosseau Jr G E (1960) *Genetics*, 45, 257-274.
5 Meyer G F, Hess O & Beermann W (1961) *Chromosoma*, 12, 676-716.
6 Hess O & Meyer G F (1968) *Adv Genet*, 14, 171-288.
7 Bonaccorsi S, Pisano C, Puoti F & Gatti M (1988) *Genetics*, 120, 1015-1034.
8 Hackstein J H P (1991) *Eur J Cell Biol*, 56, 151-169.
9 Hennig W, Brand R C, Hackstein J, Hochstenbach R, Kremer H, Lankenau D-H, Lankenau S, Miedema K & Potgens A (1989) *Genome*, 31, 561-571.
10 Hackstein J H P & Hochstenbach R (1995) *Trends Genet*, 11, 195-200.
11 Hennig W (1993) *Proc Natl Acad Sci USA*, 90, 10904-10906.
12 Lifschytz E (1987) *Int Rev Cytol*, 109, 211-258.
13 Gepner J & Hays T S (1993) *Proc Natl Acad Sci, USA*, 90, 11132-11136.
14 Hackstein J H P, Leoncini O, Beck H, Peelan G & Hennig W (1982) *Genetics*, 101, 257-277.
15 Grond C J, Siegmund I & Hennig W (1983) *Chromosoma*, 88, 50-56.

16 Hennig W (1990) in *Chromosomes: eukaryotic, prokaryotic and viral* (ed. Adolph K W), pp 213-238, CRC Press, Boca Raton, Florida, USA.
17 Hochstenbach R, Brand R & Hennig W (1994a) *Mol Gen Genet*, 244, 653-660.
18 Hochstenbach R, Harahangi H, Schouren K & Hennig W (1994b) *J Mol Evol*, 39, 452-465.
19 Hochstenbach R, Potgens A, Meijer H, Dijkhof R, Knops M, Schouren K & Hennig W (1993) *Chromosoma*, 102, 526-545.
20 Hardy R W, Lindsley D L, Livak K J, Lewis B, Siversten A L, Joslyn G L, Edwards J & Bonaccorsi S (1984) *Genetics*, 107, 591-610.
21 Pimpinelli S, Bonaccorsi S, Gatti M & Sandler L (1986) *Trends Genet*, 2, 17-20.
22 Livak K J (1990) *Genetics*, 124, 303-316.
23 Livak K J (1984) *Genetics*, 107, 611-634.
24 Palumbo G, Bonaccorsi S, Robbins L G & Pimpinelli S (1994) *Genetics*, 138, 1181-1197.
25 Hurst D L (1992) *Genetics*, 130, 229-230.
26 Danilevskaya O N, Kurenova E V, Pavlova M N, Bebehov D V, Link A J, Koga A, Vellek A & Hartl D L (1991) *Chromosoma*, 100, 118-124.
27 Balakireva M D, Shevelyov Y Y, Nurminsky D I, Livak K J & Ovozdev V A (1992) *Nucleic Acids Res*, 20, 3731-3736.
28 Russell S R H & Kaiser, K (1994) *Chromosoma*, 103, 63-72.
29 Lankenau S, Corces V G & Lankenau D-H (1994) *Mol Cell Biol*, 14, 1764-1775.
30 Ashburner M (1967) *Chromosoma*, 21, 398-428.
31 Mukherjee T & Lakhotia S C (1979) *Chromosoma*, 74, 75-82.
32 Lakhotia S C (1987) *J Genet*, 66, 139-157.
33 Lakhotia S C (1989) *Genome*, 31, 677-683.
34 Lakhotia S C & Sharma A (1995) *Genetica*, (in press).
35 Lakhotia S C & Mukherjee T (1982) *Chromosoma*, 85, 369-374.
36 Peters F P A M N, Lubsen N H & Sondermeijer P J A (1984) *Chromosoma*, 81, 271-280.
37 Ryseck R P, Walldorf U & Hovemann B (1985) *Chromosoma*, 93, 17-20.
38 Hovemann B, Walldorf U & Ryseck R P (1986) *Mol Gen Genet*, 204, 334-340.
39 Garbe J C, Bendena W G & Pardue M L (1989) *Genetics*, 122, 403-415.
40 Hogen N C, Slot F, Traverse K L, Garbe J C, Bendena W G & Pardue M L (1995) *Genetics*, 139, 1611-1621.
41 Fini M E, Bendena W G & Pardue M L (1989) *J Cell Biol*, 108, 2045-2057.
42 Pardue M L, Bendena W G, Fini M E, Garbe J C, Hogan N C & Traverse K L (1990) *Biol Bull*, 179, 77-86.
43 Bendena W G, Garbe J C, Traverse K L, Lakhotia S C & Pardue M L (1989) *J Cell Biol*, 108, 2017-2028.
44 Lakhotia S C & Sharma A (1995) *Chromosome Res*, 3, 151-161.
45 Mutsuddi M & Lakhotia S C (1995) *Dev Genet*, 17, 303-311.
46 Mohler J & Pardue M L (1984) *Genetics*, 106, 249-265.
47 Morcillo G, Diez M E, Carbajal M E & Tanguay R M (1993) *Chromosoma*, 102, 648-659.
48 Lakhotia S C & Ray P (1996) *J Biosci*, (in press).
49 Sharma A & Lakhotia S C (1995) *Chromosome Res*, 3, 386-392.
50 Lis J, Prestidge L & Hogness D S (1978) *Cell*, 14, 901-919.
51 Lis J, Neckameyer W, Mirault M E, Artavanis-Tsakonas S, Lall P, Martin G & Schedl P (1981) *Dev Biol*, 83, 291-300.
52 Lengyel J A & Graham M L (1984) *Nucleic Acids Res*, 12, 5719-5735.
53 Ish-Horowicz D, Holden J J & Gehring W (1977) *Cell*, 12, 643-652.
54 Hackett R W & Lis J T (1981) *Proc Natl Acad Sci, USA*, 78, 6196-6200.
55 Leigh-Brown A J & Ish-Horowicz D (1981) *Nature, London*, 290, 677-682.
56 Kar Choudhuri D & Lakhotia S C (1986) *Chromosoma*, 94, 279-284.
57 Lyon M F (1961) *Nature, London*, 190, 372-373.
58 Rastan S & Brown S D M (1990) *Genet Res*, 56, 99-106.
59 Brown S D M (1991) *Bioessays*, 13, 607-612.
60 Gartler S M, Dyer K A & Goldman M A (1992) *Mol Genet Med*, 2, 121-160.
61 Brown C J, Ballabio A, Rupert J L, Lafreniere R G, Grompe M, Tonlorenzi R & Willard H F (1991) *Nature, London*, 349, 38-44.
62 Brown C J, Hendrich B D, Rupert J L, Lafreniere R G, Xing Y, Lawrence J & Willard H F (1992) *Cell*, 71, 527-542.
63 Brockdorff N, Ashworth A, Kay G F, Cooper P J, Smith S, McCabe V M, Norris D P, Penny G D, Patel D & Rastan S (1991) *Nature, London*, 351, 329-331.
64 Brockdorff N, Ashworth A, Kay G F, McCabe V M, Norris D P, Cooper P J, Swift S & Rastan S (1992) *Cell*, 71, 515-526.
65 Takagi N & Sasaki M (1975) *Nature, London*, 256, 640-642.
66 Kay G F, Penny G D, Patel D, Ashworth A, Brockdroff N & Rastan S (1993) *Cell*, 72, 171-182.
67 Kay G F, Barton S C, Surani M Z & Rastan S (1994) *Cell*, 77, 639-650.
68 Norris D P, Patel D, Kay G D, Penny N, Brockdroff N, Sheardown S A & Rastan S (1994) *Cell*, 77, 41-51.
69 Zuccotti M & Monk M (1995) *Nature Genetics*, 9, 316-320.
70 Ariel M, Robinson E, McCarrey J R & Cedar H (1995) *Nature Genetics*, 9, 312-315.
71 Armstrong S J, Kirkham A J & Hulten M A (1994) *Chromosome Res*, 2, 445-452.
72 Das P & Raman R (1994) *J Biosci*, 19, 513-528.
73 Walker C L, Cargile C B, Floy K M, Delannoy M & Migeon B R (1991) *Proc Natl Acad Sci, USA*, 88, 6191-6195.
74 Flejter W L, Van Dyke D L & Weiss L (1984) *Am J Hum Genet*, 36, 218-226.
75 Goldman M A (1992) *Nature Genetics*, 2, 169-170.
76 Buzin C H, Mann J R & Singer S J (1994) *Development*, 120, 3529-3536.
77 Pachnis V, Belayew A & Tilghman (1984) *Proc Natl Acad Sci, USA*, 81, 5523-5527.
78 Brannan C I, Dees E C, Ingram R S & Tilghman S M (1990) *Mol Cell Biol*, 10, 28-36.
79 Pfeiffer K & Tilghman S M (1994) *Genes & Dev*, 8, 1867-1874.
80 Leighton P A, Ingram R S, Eggenschwiler, Efstratiadis A & Tilghman S M (1995) *Nature, London*, 375, 34-39.
81 Hao Y, Crenshaw T, Moulton T, Newcomb E & Tycko B (1993) *Nature, London*, 365, 764-767.
82 Rachmilewitz J, Gileadi O, Eldar-Geva T, Schneider T, de-Groot N & Hochberg A (1992) *Mol Reprod Dev*, 32, 196-202.

83 Askew D S, Bartholomew C, Buchberg A M, Valentine M B, Jenkins M A, Copeland N G & Ihle J N (1991) *Hematol Pathol*, 7, 1-22.
84 Askew D S, Li J & Ihle J N (1994) *Mol Cell Biol*, 14, 1743-1751.
85 Velleca M A, Wallace M C & Merlie J P (1994) *Mol Cell Biol*, 14, 7095-7104.
86 Fournier M J & Maxwell E S (1993) *Trends Biochem Sci*, 18, 131-135.
87 Sollner-Webb B (1993) *Cell*, 75, 403-405.
88 Tyc K & Steitz J A (1989) *EMBO J*, 8, 3113-3119.
89 Tycowski K T, Shu M-D & Steitz J A (1994) *Science*, 266, 1558-1561.
90 Waverick R, Kerns J A & Francke U (1994) *Human Mol Genet*, 3, 1877-1882.
91 Costantini F D, Britten R J & Davidson E H (1980) *Nature, London*, 287, 111-117.
92 Anderson D M, Richter J D, Chamberlin M E, Price D H, Britte R J, Smith L D & Davidson E H (1982) *J Mol Biol*, 155, 281-309.
93 Richter J D, Anderson D M, Davidson E H & Smith L D (1984) *J Mol Biol*, 173, 227-241.
94 McGrew L L & Richter J D (1989) *Dev Biol*, 134, 267-270.
95 Calzone F J, Lee J J, Le N, Britten R J & Davidson E H (1988) *Genes & Dev*, 2, 305-318.
96 Diez M O & Gall J G (1985) *Chromosoma*, 92, 243-253.
97 Davidson E H (1986) *Gene activity in early development*, Acad Press Orlando, Florida.
98 Spohr G, Reith W & Sures I (1981) *J Mol Biol*, 151, 573-592.
99 Kloc M, Spohr G & Etkin L D (1993) *Science*, 262, 1712-1714.
100 Kloc M & Etkin L D (1994) *Science*, 265, 1101-1103.
101 Watanabe Y, Lino Y, Furuhata K, Shimoda C & Yamamoto M (1988) *EMBO J*, 761-767.
102 Watanabe Y & Yamamoto M (1994) *Cell*, 78, 487-498.
103 Macneill S & Fantes P (1995) *Curr Biol*, 5, 4-6.
104 Hotta Y, Tabata S, Stubbs L & Stern H (1985) *Cell*, 40, 785-793.
105 Altman S, Kirsebom L & Talbot S (1993) *FASEB J*, 7, 7-14.
106 Long D M & Uhlenbeck O C (1993) *FASEB J*, 7, 25-30.
107 Wightman B, Ha I & Ruvkun G (1993) *Cell*, 75, 855-862.
108 Lee R C, Feinbaum R L & Ambros V (1993) *Cell*, 75, 843-854.
109 Mattaj I W (1993) *Cell*, 73, 837-840.
110 Fukami-Kobayashi K, Tomoda S & Go M (1993) *FEBS Lett*, 335, 289-93.
111 Scherrer K & Bey F (1994) *Progr Nucl Acid Res Mol Biol*, 49, 1-64.
112 Predkl P F, Nayak L M, Geottlieb M B C & Regan L (1995) *Cell*, 80, 41-50.
113 Curtis D, Lehmann R & Zamore P D (1995) *Cell*, 81, 171-178.
114 Izaurralde E & Mattaj I W (1995) *Cell*, 81, 153-159.
115 Noller H F, Hoffarth V & Zimmiak L (1992) *Science*, 256, 1416-1419.
116 Waldrop M M (1992) *Science*, 265, 1396-1397.
117 Pace N R (1992) *Science*, 256, 1402-1403.
118 Burma D P (1992) *Curr Science*, 63, 550-559.
119 Chatterjee D (1992) *Curr Science*, 63, 535.
120 St Johnston D (1995) *Cell*, 81, 161-170.
121 Lakhotia S C (1995) *Curr Science*, 69, 219-221.
122 Beelmann C A & Parker R (1995) *Cell*, 81, 179-183.
123 Rastinejad F & Blau H M (1993) *Cell*, 72, 903-917.
124 Wickens M (1993) *Nature, London*, 363, 305-306.
125 Rastinejad F, Conboy M J, Rando T A & Blau H M (1993) *Cell*, 75, 1107-1117.
126 Shi Y & Berg J M (1995) *Science*, 268, 282-284.
127 Wassenegger M, Heimes S, Riedel L & Sänger H L (1994) *Cell*, 76, 567-576.
128 Bolden A, Ward C, Siedlecki J A & Weissbach A (1984) *J Biol Chem*, 259, 12437-12443.
129 Smith S S, Kan J L C, Baker D J, Kaplan B E & Dembek P (1991) *J Mol Biol*, 217, 39-51.
130 Escude C, Francois J C, Sun J S, Ott G, Sprinzl M, Garestier T & Helene C (1993) *Nucleic Acids Res*, 21, 5547-5553.
131 McDonald C D & Maher III L J (1995) *Nucleic Acids Res*, 23, 500-506.
132 Zuckerkandl E & Hennig W (1995) *Chromosoma*, 104, 75-83.