Minireview

Employment opportunities for non-coding RNAs

Céline Morey, Philip Avner*

Génétique Moléculaire Murine CNRS 2578, Institut Pasteur, 25, rue du Docteur Roux, 75015 Paris, France

Received 19 March 2004; accepted 23 March 2004

Available online 28 April 2004

Edited by Horst Feldmann

Abstract Analysis of the genomes of several higher eukaryotic organisms, including mouse and human, has reached the striking conclusion that the mammalian transcriptome is constituted in large part of non-protein-coding transcripts. Conversely, the number of protein-coding genes was initially at least overestimated. A growing number of studies report the involvement of non-coding transcripts in a large variety of regulatory processes. This review examines the different types of non-coding RNAs (ncRNAs) and discusses their putative mode of action with particular reference to large ncRNAs and their role in epigenetic regulation.

Keywords: Non-coding RNA; Antisense transcript; Intergenic transcription; X-chromosome inactivation

1. Introduction and historical context

In the 1960s when Jacob and Monod [1] first defined the basic principles underlying bacterial genetic regulatory systems, the existence of two distinct types of genes was already clearly established: structural genes encoding mRNAs for proteins and regulatory genes producing regulatory RNAs that interact by base pairing with the operator elements of structural genes. For much of the last twenty years, at least as far as higher eukaryotes have been concerned, interest in the transcriptional activity of the genome has been focussed almost exclusively on gene discovery and protein-coding genes. Progressively, the dogma that the critical functions of the cell depend exclusively on proteins gained ground, the complexity of an organism depending exclusively on its repertoire of protein-coding genes. RNAs were considered solely as accessory molecules, involved mainly in mediating the processes of transcription and translation. This over-simplistic view was first called into question by the discovery of untranslated intronic sequences embedded within coding genes and the associated finding of splice variants which allow the synthesis of more than one protein product from a single gene (for review see [2]). Identification of the catalytic properties of the RNA subunit of ribonuclease P further underlined the notion that the functions of RNAs extend well beyond a transient role in ensuring the expression of protein-coding genes (for review see [3]).

Recent advances in genome sequencing and the availability of extensive finished sequence for the human and mouse genomes has not significantly improved our ability to identify putative regulatory RNAs. Much of the sequence related effort in genomics, which has been concentrated on developing methodologies to identify and define classical genes within genomic sequence, is based on the identification of conserved coding exons by comparative genome analysis [4] or on computational gene prediction which relies on gene-finding algorithms [5]. Such gene-finding algorithms are designed to identify open reading frames (ORFs), polyadenylation signals, conserved promoter regions and splice sites typically associated with protein-coding genes. They generally do not allow the detection of non-coding RNA (ncRNA) genes, which often show only weak primary sequence conservation, lack ORFs and are not systematically processed. ncRNAs remained closeted until very recently, when relevant computational and experimental approaches were at last initiated. In yeast, searches for polymerase III promoters, which often characterize small RNA genes, and analysis of expression profiles within "gaps" between predicted ORFs have identified new ncRNAs. Of particular interest is a recent comparison between the sequence annotations available for human chromosomes 21 and 22 and the expression profile of these chromosomes as established by hybridization of polyA$^+$ total RNAs extracted from 11 different human cell lines using oligonucleotide arrays, which cover the entire genomic DNA sequence of these chromosomes. These studies have led to the conclusion that there are 10-fold more transcription units than predicted coding genes in the human genome [6]. A small fraction of these additional transcription units is thought to represent pseudogenes, as sequence analysis has revealed that many of the 20 000 pseudogenes in the human genome possess functional promoters. (Note, however, that transcription of some pseudogenes may exert a regulatory function [7].) A larger fraction of this transcriptional activity is probably derived from repeat sequences that are transcribed such as long terminal repeats, small interspersed elements and long interspersed elements (LINEs). Many of these repeat elements contain active promoters; often containing canonical *Pol*II recognition sites.

[*] Corresponding author. Fax: +33-1-45-68-8656.
E-mail addresses: cmorey@pasteur.fr (C. Morey), pavner@pasteur.fr (P. Avner).

Abbreviations: ncRNAs, non-coding RNAs; LINE, long interspersed element; miRNA, microRNA; TGS, transcriptional gene silencing; PTGS, post-transcriptional gene silencing; RNAi, RNA interference; dsRNA, double-stranded RNA; siRNA, short interfering RNA; LCR, locus control region; NAT, natural antisense transcript; XCI, X-chromosome inactivation; *Xic*, X-inactivation centre; DCC, dosage compensation complex; MSL proteins, male sex lethal proteins

0014-5793/$22.00 © 2004 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2004.03.117

Others, such as many of the LINEs are truncated at their 5′ end. Accidental insertion of such elements, even when truncated, downstream of an endogenous promoter might however be expected to result in transcription. How many of the aberrant transcripts originating from such so-called "junk" DNA are retained and how many are recognized as illegitimate and rapidly destroyed by mechanisms such as the non-sense-mediated decay pathway, remains to be defined [8].

Originally, the term ncRNAs referred exclusively to polyadenylated eukaryotic RNAs transcribed by RNA polymerase II, carrying a 7-methylguanosine cap structure and lacking an ORF. Nowadays, this definition has been extended to designate all RNA transcripts without protein-coding capacity. Such ncRNAs can be divided into two classes: housekeeping RNAs and regulatory RNAs. Housekeeping ncRNAs (Table 1) are usually small, constitutively expressed and necessary for cell viability. They include not only ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) but also small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) implicated in diverse functions including splice regulation and rRNA modification. Housekeeping ncRNAs also include RNAs important for the transport and insertion of proteins into membranes and telomeric sequence addition (for review see [9]). A subset of the additional transcription units, that have been detected in global transcriptional profiling experiments, will probably correspond to novel ncRNAs falling into this category. The recent elaboration of software based on algorithms using "covariance models", which allow the detection of consensus secondary structures, may provide a way to identify and associate these small ncRNAs with cellular functions [10].

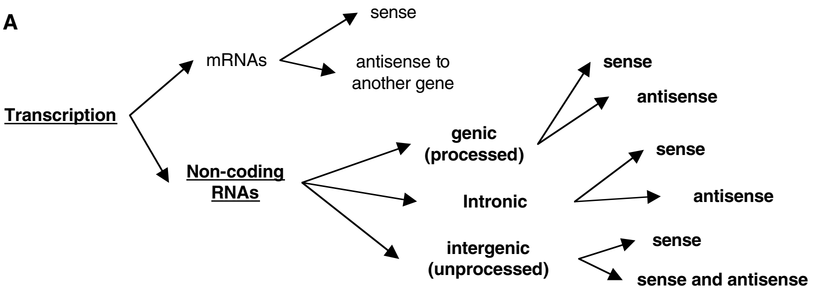
Regulatory ncRNAs include the microRNA (miRNA) family, which can induce post-transcriptional gene silencing (PTGS) by repressing mRNA translation (Fig. 1B). miRNAs are involved in the control of developmental timing and/or tissue-specific functions. The first miRNAs to be described, corresponded to the products of the *lin-4* and *let-7* genes in *Caenorhabditis elegans*. These miRNAs, 22 and 21 nucleotides in length, are processed from larger precursor RNAs and inhibit translation through antisense interactions with the 3′ untranslated region of the target mRNAs. miRNAs are widely distributed in all organisms. They are encoded in the genome as short inverted repeats which have a double-stranded RNA (dsRNA) stem loop about 70 bp long [11] and are found in introns as well as in intergenic clusters. Processing produces the small of 21–25 nucleotide effector molecules, which are usually derived from only one strand of the stem loop structure. Some miRNAs have homologues in both vertebrates and invertebrates although their small size renders the criteria of conservation between species often insufficient for the identification and isolation of new miRNAs. The number of miRNAs in human is thought to be 220–250 [12].

Besides such small ncRNAs, an even greater mystery envelops the role of larger regulatory ncRNAs. These RNAs show great diversity in their genomic organization. In some cases, they are produced from within a well-defined gene (in such cases, they are processed: spliced and/or polyadenylated), whilst in others, transcription either initiates within the intron of a host gene or may result from intergenic transcription (Fig. 1A). The observation that many regulatory elements such as locus control regions (LCRs), boundary elements, silencers and insulators are transcribed, points to the probable involvement of intergenic non-coding transcription in the function of these genomic regulators [13]. ncRNAs may also be transcribed from a single strand (sense ncRNAs) as well as in the opposite orientation when they may overlap with either protein-coding or non-coding genes (antisense RNAs/transcription). Many examples of *cis*-acting natural antisense transcripts (*cis*-NATs) have recently been described (for review see [14]). Two recent studies, one using the FANTOM2 mouse cDNA set, public mRNA data and mouse genome sequence data [15], the second combining a computational based search for sense-antisense transcripts pairs in human genome public databases with experimental assessment of the results using microarrays containing strand-specific oligonucleotides probes [16], have shown that the contribution of NATs has been largely under-evaluated in both mouse and human. It is now estimated that an antisense transcript is transcribed from some 10–20% of genes. Due to database bias in favour of polyA$^+$ RNAs, most of the antisense transcripts detected in these analyses were spliced and polyadenylated. This artefactual bias in favour of processed antisense RNAs may suggest that the overall proportion of antisense transcripts may be even higher. All possible types of arrangements of *cis*-NATs relative to their sense counterpart have been described: tail-to-tail, head-to-head or arrangements with the antisense transcript totally included within an intron of the sense transcript have been described [15]. Antisense ncRNAs appear to be especially abundant at imprinted loci, which are usually organized in

Table 1
Functional classification of housekeeping ncRNAs

Type	Function	Databases or search tools
rRNA	Translation of genetic information	http://intra.psb.ugent.be:8080/rRNA/
tRNA	Translation of genetic information	http://rna.wustl.edu/tRNAdb/
snRNA	Pre-mRNA splicing; spliceosome components	http://psyche.uthct.edu/dbs/uRNADB/uRNADB.html
snoRNA	RNA modifications, 2′-*O*-methylation and pseudouridylation	http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_dataBase.html
tmRNA	*trans*-translation	http://www.bioinf.au.dk/tmRDN/ http://psyche.uthct.edu/dbs/tmDB/tnRDB.html http://www.ag.auburn.edu/mirror/tmRDB
Telomerase RNA	Telomeric DNA synthesis	http://mbcr.bcm.tmc.edu/smallRNA/Database/Telomere-RNA
Signal recognition		http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html
Particle		http://bio.lundberg.gu.se/dbs/SRPDB.SRPDB.html
Ribonuclease RNA	RNA processing	http://jwbrown.mbio.ncsu.edu/RNaseP/home.html

Non-coding housekeeping RNAs are outside of the focus of this review. Details about this class of transcripts can be found in the indicated databases.
Abbreviations: rRNA, tRNA, snRNA, snoRNA, tmRNA, transfer-messanger RNA.



B

Type	Examples	Organism	Size	Regulation system	Effects/mechanisms	Chromatin	Databases or search tools
sense	*Xist*	Mammals	15-17 kb (Mouse)	X-chromosome inactivation (Dosage compensation)	Long-range *cis*-action transcriptional silencing	heterochromatin	
	roX1 *roX2*	*Drosophila*	3,7 kb 0,6 kb	X hypertranscription (Dosage compensation)	Long range *cis*- and *trans*-action transcriptional activation	relaxed chromatin	
antisense	*Tsix (Xist)*	Mouse	> 40 kb	X-chromosome inactivation (Dosage compensation)	*cis*-repression of the sense counterpart random X-chromosome inactivation	open chromatin	***in silico*** **detection of antisense** http://www.hgmp.mrc.ac.uk/Research/Antisense/
	Air (Igf2r)	Mammals	108 kb (Mouse)	Genomic imprinting	*cis*-repression of the sense counterpart monoallelic expression of imprinted genes		http://arep.med.harvard.edu/twister/antisense.html http://labonweb.com/antisense/
	Ube3a-as *(Ube3a)*	Mammals	450 kb (Human)	Genomic imprinting (human AS/PWS locus)	*cis*-repression of the sense counterpart		http://genome.gsc.riken.go.jp/m/antisense/ http://bio.ifom-firc.it/ANTIHUNTER/
	otr	Yeast		silencing of centromeric repeats	TGS (RNAi)	heterochromatin	
Intergenic transcripts	*Xite* region B	Mouse		X-chromosome inactivation (Dosage compensation)	*Tsix* regulation ?	open chromatin ?	
	β-globin LCR β-globin locus	Mammals		developmental regulation of the β-globin locus	Long range *cis*-activation of target genes 5'->3' interactions	open chromatin open chromatin	
	iabs (BX-C)	*Drosophila*		Regulation of the BX-C	5'->3' interactions		
miRNAs (intergenic or intronic)	*lin-4* ; *let-7* (*C.Elegans*)	all organisms	21-25 nt	developmental regulation tissue-specific functions	PTGS of a wide variety of target genes RNAi pathway	?	***in silico*** **detection of miRNAs** http://sci.bio.argon.acad.bg/mirna/MIST2eng.htm http://biobases.ibch.poznan.pl/ncRNA/

General databases for ncRNAs:
http://biobases.ibch.poznan.pl/ncRNA/
http://www.sanger.ac.uk/Software/Rfam/
http://rfam.wustl.edu
http://infernal.wustl.edu/

Fig. 1. Classification of regulatory ncRNAs. (A) The different classes of regulatory RNAs that may emerge from the act of transcription have been subdivided according to their localization, both with regard to genomic organization and the orientation of transcription. (B) This table shows examples of regulatory ncRNAs from different eukaryotic organisms, their function and associated chromatin structure. Databases ressources are provided.

clusters and are regulated by putative imprinted centres [17]. Antisense transcripts have been found for 22 of the 58 mouse imprinted genes examined [15].

One of the best-characterized regions producing different subtypes of ncRNAs and containing non-coding transcription units is the mouse X-inactivation centre (*Xic*) (Fig. 3), the master control region implicated in the regulation of X-chromosome inactivation (XCI) in mammals [18]. Through reference to this system and other particularly informative examples, we will focus the rest of this review on the functions potentially exerted by large regulatory ncRNAs and the associated underlying molecular mechanisms.

2. ncRNAs and the long-range establishment of high-order chromatin structures

In mammals, dosage compensation of X-linked gene products between the sexes is achieved by the transcriptional silencing of a single X-chromosome during early female embryogenesis. Initiation of XCI requires the counting of X-chromosomes and the random choice of the X-chromosome to inactivate (for review see [19]). Once initiated, the inactivation signal is propagated along the chromosome inducing its heterochromatinization. This inactive state is stably maintained through subsequent cell divisions. The *Xic*, which controls XCI initiation and spreading, contains the X-inactive specific transcript (*Xist*) gene. *Xist* was first discovered in 1991 and corresponds to one of the first gene described to encode an untranslated RNA [20–22]. The gene is conserved between species at the level of its genomic organization but shows only weak sequence homology, possibly implicating a role for secondary structure [18]. *Xist* ncRNAs are 15–17 kb long in mice, spliced, polyadenylated and restricted to the nuclear compartment [23,24]. Prior to the onset of XCI, *Xist* RNA is synthesized from both X chromosomes in the female embryo. *Xist* up-regulation on the putative inactive X-chromosome and RNA coating of this chromosome constitute the first detectable signs of inactivation. This coating and *Xist* upregulation then trigger the enrichment of the targeted X by the lysine-27 methylated form of histone H3 (H3K27). The establishment of this modification is ensured by the polycomb group complex *eed enx1*, which possesses a histone H3K27 methyltransferase activity and which transiently accumulates on the future inactive X at the time of XCI [25,26]. H3K27 hypermethylation is accompanied by other chromatin changes, including H3K4 hypomethylation, H3K9 hypermethylation and H3–H4 hypoacetylation. These modifications appear concomitantly with the transcriptional silencing of X-linked genes. It remains unclear whether this silencing is directly triggered by *Xist* accumulation or is subsequent to chromatin modifications. Replication timing of the inactive X then becomes retarded into late S-phase, whilst the CpG islands associated with the promoters of X-linked genes are methylated [27]. These numerous and successive layers of modification lead to the establishment of a silent chromatin structure on the elected X and, in turn, lock the inactive X into a stable heterochromatic state, which ensures the faultless maintenance of the inactive state throughout the cell cycle and its transmission at mitosis.

Unlike classical mRNAs which are exported to the cytoplasm or to other ribonuclear particles and which are characterized by their high mobility [28], the *Xist* RNAs remain localized within the nucleus and tightly associated with the inactive chromatin. Experiments aimed at dissociating the *Xist* containing chromatin fraction have indicated that *Xist* RNAs is bound to the nuclear matrix and suggested an association of *Xist* transcripts with nuclear matrix attachment proteins [29]. The accumulation of scaffold associated factor A (SAF-A) on the inactive X may indicate that this factor ensures the link between *Xist* and the nuclear matrix [30]. Restriction of *Xist* RNAs to the nucleus may also involve a specialized nuclear compartimentalization process.

The *Xist* transcript represents the canonical example illustrative of the ability of some ncRNAs to 'travel' long distances in order to induce chromatin modifications. It is interesting to note that despite its long range effect, the *Xist* RNA propagates itself only along the chromosome from which it originates. The molecular mechanisms responsible for this *cis*-spreading and the consequent chromatin refolding remain poorly understood. *Xist* RNA propagation is thought to involve relay or entry sites interspersed along the chromosome. These specialized sequences may work as nucleation centres which concentrate chromatin remodelling complexes and facilitate the propagation of the silent chromatin structure to surrounding sequences (Fig. 2A). One such nucleation centre has been proposed to correspond to a 150-kb region lying just 5′ to the *Xist* gene which is constitutively enriched in H3K9 methylation [31] (Rougeulle et al., in press) (Fig. 3).

The inactive X chromatin is also enriched in the histone variant macroH2A, which accumulates within a nuclear structure known as the macrochromatin body. *Xist* RNA is necessary for the localization of macroH2A to the inactive X [32]. Due to its histone nature, macroH2A has been proposed to be involved in linking the X-chromatin to *Xist* RNA. Since, in the ES cell model, macroH2A only becomes associated with the inactive X, a considerable time after the onset of XCI, it is however more likely to be involved in the maintenance of the transcriptionally silent chromatin structure.

Site-specific transgenes expressing different forms of *Xist* RNA have allowed the function of specific parts of the *Xist* transcript to be analysed [33]. Interestingly, a repeat sequence located at the 5′ extremity of the *Xist* RNA which is able to adopt a double hairpin conformation seems to be responsible for the silencing function of *Xist*. In the absence of this motif, *Xist* decorates the X-chromosome without however inducing transcriptional repression. In contrast, the coating function appears to be mediated by many different and redundant sequences within the transcript acting in a cooperative manner [33].

In Drosophila, unlike the situation pertaining in mammals, dosage compensation is achieved by a 2-fold upregulation of transcription of genes on the single X-chromosome present in the males. Intriguingly, however, the fly dosage compensation system also involves two ncRNAs: *roX1* and *roX2* (RNA on the X) which show features similar to *Xist*. *roX1* and *roX2* are members of the dosage compensation complex (DCC), which also contains the male specific lethal (MSL) proteins. An intact DCC complex is required for hyper-transcription of genes on the X-chromosome. A translation block of any one of these subunits results in the failure of the DCC assembly in females (for review see [34]). MSL proteins include chromodomain proteins which can interact with RNA in vitro and an histone acetyl transferase responsible for the decondensed chromatin structure which characterizes the dosage compensated Drosophila X-chromosome. In genetic backgrounds where some of the MSL proteins have been mutated, a partial DCC can still be detected associated with approximately 35 primary "chromatin entry sites" distributed along the X [35]. These "entry sites" have been shown to act as initial docking sites for the DCC assembly from where the complex spreads *in cis* to hundreds of other sites which ensure the propagation of a decondensed chromatin structure along the entire X-chromosome (Fig. 2A).

These two examples serve to emphasize the capacity of some ncRNAs to spread in *cis* over long distances (it is, however, noteworthy that *roX*s ARNs can also act in *trans* when inserted at an ectopic location). Interestingly, the RNA molecules described in these examples seem to be capable of multiple interactions with chromatin modifying enzymes; these interactions mediating the chromatin organization at the chromosomal level. The recent identification of the involvement of an RNA component in the establishment of histone modification patterns at mammalian pericentric heterochromatin may point to a more general implication of RNA molecules in high-order chromatin structures [36].

3. Antisense ncRNAs and transcriptional repression

Xist RNA is the major effector responding to the initiation signal(s) involved in XCI. Intriguingly, *Xist* expression itself is controlled by another non-coding transcription, antisense to *Xist*, named *Tsix* (Fig. 3). *Tsix* transcription is mainly initiated 12 kb downstream of *Xist* and gives rise to approximately equivalent amount of primary and spliced transcripts which span a 40-kb region and overlap the entire *Xist* locus. Like *Xist*, *Tsix* is expressed on all Xs prior to the onset of XCI. When XCI initiates, *Xist* RNA up-regulation on the future inactive X-chromosome is accompanied by transcriptional repression of *Tsix*, whereas on the active X *Tsix* expression persists [37].

Genomic imprinting, like XCI, leads to monoallelic expression of target genes. Non-coding antisense transcription is frequent as already mentioned at imprinted loci and shows a mutually exclusive expression pattern with its sense counterpart. Antisense transcription at imprinted loci can extend over several hundred kilobases (*Air*, Antisense transcription at the *Igf2r* locus, 108 kb; *Ube3a-as*, antisense at the Angelman Syndrome/Prader-Willi Syndrome locus (AS/PWS locus), 450 kb or more, Fig. 1B) The extended nature of the

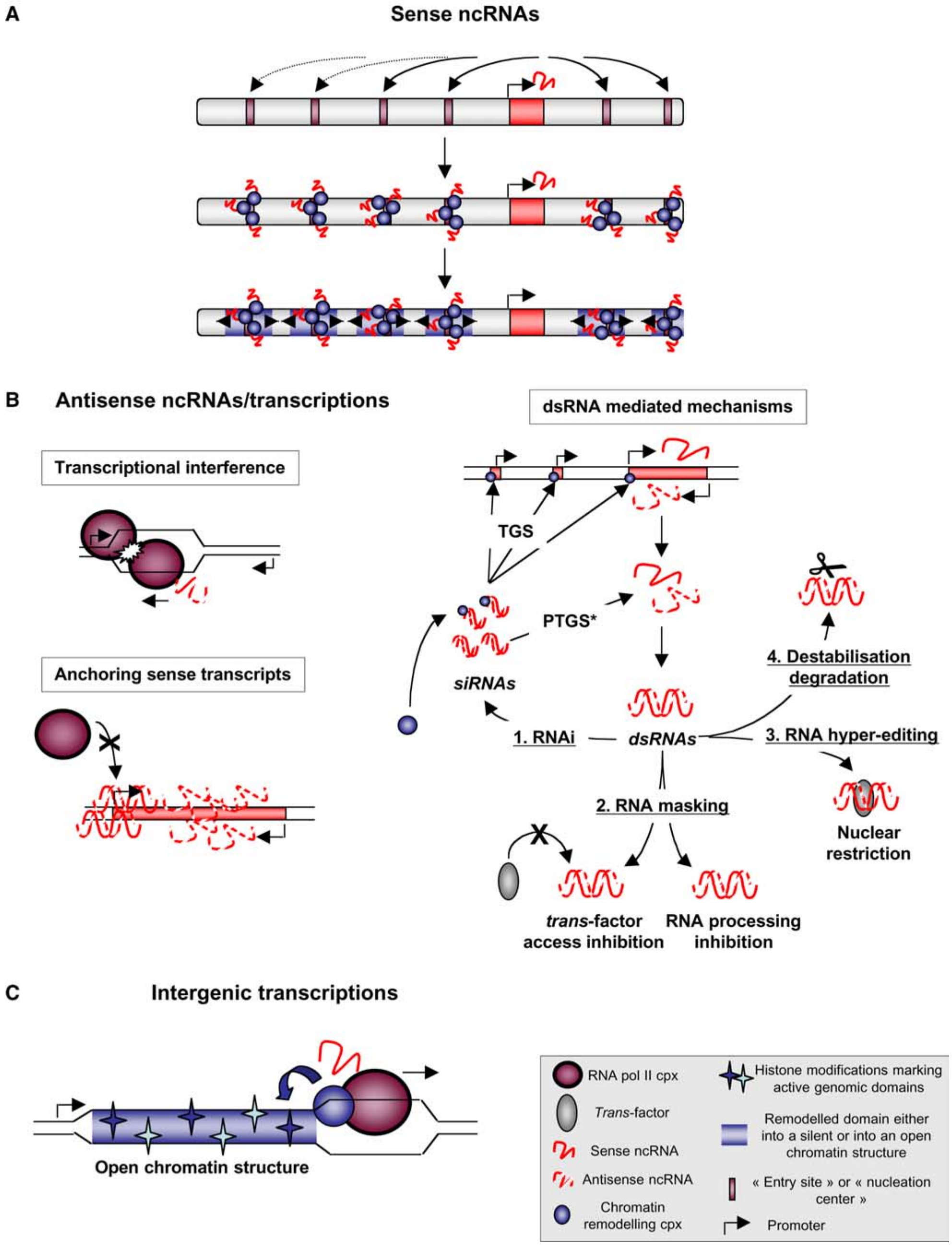

A
Sense ncRNAs
B
Antisense ncRNAs/transcriptions
dsRNA mediated mechanisms
Transcriptional interference
Anchoring sense transcripts
TGS
PTGS*
siRNAs
1. RNAi
dsRNAs
2. RNA masking
3. RNA hyper-editing
4. Destabilisation degradation
Nuclear restriction
trans-factor access inhibition
RNA processing inhibition
C
Intergenic transcriptions
Open chromatin structure
RNA pol II cpx
Trans-factor
Sense ncRNA
Antisense ncRNA
Chromatin remodelling cpx
Histone modifications marking active genomic domains
Remodelled domain either into a silent or into an open chromatin structure
« Entry site » or « nucleation center »
Promoter

Fig. 2. Putative regulatory mechanisms of ncRNAs. (A) Sense ncRNAs may participate in the large-scale establishment of specific chromatin structures through the recruitment of chromatin remodelling complexes (cpx) at specific sites interspersed along the chromosome and the subsequent spreading of the refolded chromatin to adjacent regions. (B) Antisense ncRNAs or transcription may repress the expression of their sense counterpart either through interference with the sense and antisense transcriptional machineries either by anchoring and retention of the nascent sense transcripts thereby impeding the access of the *Pol*II complex (cpx) to the promoter of the sense transcription unit or alternatively by process(es) involving dsRNAs intermediaries (for details see the text). (C) Intergenic transcription may serve to define domains of open chromatin structure through the deposition of histone modifications linked to the passage of the *pol*II complex (cpx) and associated chromatin remodelling factors.

transcripts may account for and mediate the long-range *cis*-effects associated with the antisense transcription. Similarities between the regulation at the *Xic* and imprinted regions may indicate that non-coding antisense transcription at both exploit the same underlying molecular mechanism(s). Truncation of both the *Air* and the *Tsix* antisense transcripts by insertion of premature polyadenylation signals [38,39] induces the reactivation of the sense transcription at both loci concomitant, respectively, with the loss of imprinted gene expression and distortion of random XCI. These results suggest a direct implication of antisense transcription in monoallelic repression of overlapping gene function. However, it is important to appreciate that these experiments do not allow discrimination between the act of antisense transcription per se as opposed to the direct implication of antisense RNA molecules.

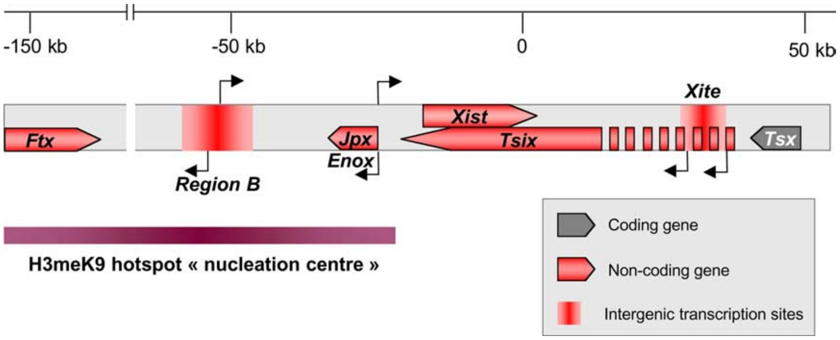


Fig. 3. Map of the mouse *Xic*. This simplified map shows the 200 kb of the Xic around the *Xist* locus and the locations of known non-coding transcription units within this region. The extent of the H3K9 methylation hotspot, which may act as a nucleation centre allowing the spreading of the X-inactivation signal, is also shown.

Although recent studies have identified many NATs, the molecular mechanisms underlying NAT-mediated gene repression are still unknown. Several different models based on results obtained in eukaryotic organisms have been proposed. Transcription-dependent mechanisms hypothesize that the repressive function originates in the movement of the RNA polymerase machinery in the opposite direction to sense transcription leading to topological constraints and/or molecular collisions involving sense transcription (Fig. 2B). This hypothesis is favoured by the observation that antisense transcripts tend to overlap the 5′ region of their sense genes which regroup essential elements such as the promoter, translation initiation sites and enhancers, and would be expected therefore to be particularly vulnerable to antisense transcription mediated disruption. However, this type of mechanism cannot fully explain the repressive effect at the *Air Igf2r* locus as genes outside of the region overlapping *Air* are also transcriptionally repressed. Since, prior to XCI, *Xist* and *Tsix* are co-expressed on both X-chromosomes, for a transcriptional interference model to be relevant, it would be necessary to postulate that the *Xist* and *Tsix* genes are transcribed alternately.

Another class of model proposes a direct role for the antisense ncRNA molecules (Fig. 2B) probably via a pairing between sense and antisense transcripts leading to the formation of dsRNAs. The formation of dsRNAs may either, through a destabilization process, induce the degradation of the sense/antisense complex, or alternatively mask access to functional domains of either or both transcripts by *trans*-acting factors. This RNA masking might interfer with any of the multiple steps involved in post-transcriptional RNA-processing including splicing, polyadenylation, transport and translation. In the case of *Xist*/*Tsix*, it is interesting to note that the absence of *Tsix* is associated with the dispersion of *Xist* RNAs outside of their transcription site even prior to the onset of XCI [40]. This observation suggests an additional role for antisense transcripts in anchoring sense RNAs at their site of transcription or at sites of anchorage. The significant molar excess of *Tsix* RNAs over *Xist* transcripts [41] may allow the trapping of all the *Xist* RNAs produced and in a certain manner, obstruct the access of the transcriptional machinery to the *Xist* promoter, leading to partial repression of *Xist* expression (Fig. 2B).

The very widespread use of RNAi-based mechanisms in different biological systems and recent evidence demonstrating that at least in some species, a part of the RNAi pathway may occur in the nuclear compartment [42] suggest a role for this process in antisense-mediated gene repression. According to this view, dsRNAs would be cleaved into short interfering RNAs (siRNAs) by Dicer or other RNaseIII family members, which would mediate the silencing of the surrounding genes presenting a minimal sequence homology (for review [43]). In the PTGS model, inhibition of the target genes results from degradation of the RNA products by siRNA. In the transcriptional gene silencing (TGS) situation, silencing is mediated by DNA methylation and chromatin changes. Links between RNAi and chromatin states have recently been established in fission yeast, where the RNAi machinery can apparently direct H3K9 methylation and repress H3K4 methylation at centromeric repeats thereby inducing heterochromatic silencing [42]. At outermost region (*otr*) repeats, complementary sense and antisense transcriptions induce the formation of siRNAs, which are thought to guide histone modifications and the recruitment of heterochromatin proteins. A provocative correlation between H3K4 methylation and *Tsix* expression has recently been observed within the *Xist* gene, which may suggest the involvement of an RNAi pathway in XCI. However, neither dsRNAs nor siRNAs corresponding to the *Xist*/*Tsix* overlapping region have as yet been described despite intensive investigation. The size of the region and the nature of the transcripts obviously complicate such analysis. It is, however, noteworthy that the XCI process would necessitate the involvement of a *cis*-limited RNAi pathway.

4. Non-coding transcriptions, open chromatin structures and trans interactions

Within known regulatory complexes, additional transcription units involving transcription from intergenic loci are increasingly being described. Three examples of this phenomenon from amongst many (for review [13]) will be discussed on here. Within the *Xic*, two novel sites of intergenic transcription have been reported so far. The first, known as the “B region”, is located approximately 50 kb 5′ and upstream to *Xist* (Fig. 3). Transcriptional activity is initiated from this region in both orientations [18]. The second, the X-inactivation intergenic transcription element (*Xite*) locus, lies 30 kb 3′ and downstream to *Xist* and is associated with a series of *DNase*I hypersensitive sites. Both the transcriptional activity and *DNase*I sensitivity of the *Xite* locus are developmentally regulated with an expression profile that parallels that of the *Tsix* gene during

XCI [44]. This observation has led the authors to propose that *Xite* might act as an enhancer or as a LCR modulating *Tsix* expression.

Such intergenic transcription units produce ncRNAs of variable size. The role of these molecules has not been addressed up until now as it has been considered that their function is likely linked to the use of transcription-dependent mechanisms rather than being RNA-dependent per se. This assessment has been comforted by the discovery of an association between the C-terminal domain of the elongating RNA polymerase II holoenzyme and histone modifying enzymes marking active chromatin domains such as histone acetyl transferases [45,46] and H3K4 histone methyl transferases [47,48]. This link between the transcriptional machinery and chromatin remodelling enzymes has led to the seductive piggybacking hypothesis [49,50]. According to this model, the movement of the RNA polymerase II elongation complex along the chromatin fibre would allow the deposition of histone modifications (Fig. 2C). Seen from this perspective, non-coding transcription would have as its ultimate function the definition of domains of open chromatin necessary for the facilitated binding of *trans*-acting factors and other *trans*-interactions with other genomic elements. Transcription at the *Xite* locus could create docking sites accessible for transcription factors and would favour interactions between these factors and *cis*-regulatory elements associated with target genes thereby allowing transcriptional activation.

Tsix antisense transcription itself may also participate in the definition of an open chromatin domain, as the loss of *Tsix* antisense transcription is associated with a loss of H3K4 methylation within the *Xist* gene [51]. This may suggest that the *Tsix* non-coding antisense transcription intervenes both in the repression of *Xist* expression and in the formation of an open chromatin domain which is potentially involved in other parts of the XCI process. XCI by its nature must involve *trans*-interactions between X-chromosomes as well as interactions with autosomal factors that could require the establishment of an open chromatin structure around *Xist* to expose target sequences. More generally, one might expect that non-coding intergenic transcription will, through chromatin remodelling, be involved in *cis* and *trans* communications between elements distributed over large genomic domains such as the *Xic*.

The correlation between intergenic transcription and regulatory activity is also exemplified by regulation at the Bithorax complex (BX-C) during Drosophila embryo development (Fig. 1B). The BX-C complex, which extends over 300 kb, includes three coding genes: the homeotic genes Ultrabithorax, Abdominal-A (*abdA*) and Abdominal-B (*AbdB*) [52]. The sequential expression of these genes along the antero-posterior embryo axe determines the fate of each parasegment of the embryo [53]. Regulation of the *abdA* and *AdbB* genes is under the control of the infra-abdominal (*iab*) region. This 100-kb *cis*-regulatory region has been genetically divided into numerous *iab*s subdomains (*iab* 2–8) containing both enhancers and silencers. The *iab*s are transcribed following the same colinearity rule as homeotic genes [54]. Alteration of transcription in one *iab* subdomain induces a homeotic transformation of the more posterior segment under its control, suggesting that intergenic transcription plays a crucially important role in *iab*s activity [55].

A final example concerns the human β-globin locus (Fig. 1B). This locus consists of five genes tightly regulated during the erythropoietic development (for review see [56]). The regulation of the β-globin locus is governed by an LCR, located 8 kb upstream of the first β-globin gene and which contains five *DNase*I hypersensitive sites. This LCR shows constitutive intergenic transcription activity, which is thought to intervene in the sequential gene activation of β-globin genes by helping in the establishment of three functional domains. The extent of these domains is precisely delineated by intergenic transcription occurring at the β-globin locus downstream of the LCR and correlates exactly with the activation patterns of the β-globin genes themselves. These domains are characterized by chromatin modifications, which also correlate with the intergenic expression state [57].

Many of these results suggest that the mechanism of repression mediated by ncRNAs may vary with the regulatory system. Whilst antisense transcription seems to be especially frequently employed in the establishment of allele-specific expression of target genes and many of our notions concerning ncRNAs have been derived from the study of such systems, the analysis of antisense transcripts at additional loci may well reveal other modes of functioning and regulatory involvement.

The discovery that ncRNAs and non-coding transcription play a role in varied and highly disparate regulatory systems and exploit an extended variety of mechanisms suggests that ncRNAs are likely implicated and integral to the overall molecular architecture of organisms (for review see [58]). The expanding numbers of new, functional ncRNAs together with the observation that regulatory RNA molecules can be produced from the introns of coding genes suggests a role for ncRNAs as signalling molecules that can be both easily produced and destroyed at a smaller energetic cost than can protein signalling molecules which require both protein synthesis and proteolysis. In this view, the transcriptional background provided by "junk" DNA may not only participate in the maintenance of genome wide low-level transcriptional activity but also be required for efficient and integrated cellular function. The baroque hypothesis of a primordial "RNA world" clearly deserves to be revisited. We may also be obliged to reevaluate our notions of what constitutes 'junk' DNA.

Acknowledgements: This work was supported by the Association pour la Recherche contre le Cancer (ARC) and the French Ministry of Research under the Action Concertée Incitative contract no. 032526.

References

[1] Jacob, F. and Monod, J. (1961) J. Mol. Biol. 3, 318–356.

[2] Hastings, M.L. and Krainer, A.R. (2001) Curr. Opin. Cell Biol. 13, 302–309.

[3] Levy, M. and Ellington, A.D. (2001) Curr. Biol. 11, R665–R667.

[4] Roest Crollius, H. et al. (2000) Nat. Genet. 25, 235–238.

[5] Lander, E.S. et al. (2001) Nature 409, 860–921.

[6] Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P. and Gingeras, T.R. (2002) Science 296, 916–919.

[7] Hirotsune, S. et al. (2003) Nature 423, 91–96.

[8] Iborra, F.J., Jackson, D.A. and Cook, P.R. (2001) Science 293, 1139–1142.

[9] Eddy, S.R. (2001) Nat. Rev. Genet. 2, 919–929.

[10] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Nucleic Acids Res. 31, 439–441.

[11] Carrington, J.C. and Ambros, V. (2003) Science 301, 336–338.

[12] Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B. and Bartel, D.P. (2003) Science 299, 1540.

[13] Cook, P.R. (2003) J. Cell Sci. 116, 4483–4491.

[14] Vanhee-Brossollet, C. and Vaquero, C. (1998) Gene 211, 1–9.

[15] Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S. and Hayashizaki, Y. (2003) Genome Res. 13, 1324–1334.

[16] Yelin, R. et al. (2003) Nat. Biotechnol. 21, 379–386.

[17] Rougeulle, C. and Heard, E. (2002) Trends Genet. 18, 434–437.

[18] Chureau, C. et al. (2002) Genome Res. 12, 894–908.

[19] Avner, P. and Heard, E. (2001) Nat. Rev. Genet. 2, 59–67.

[20] Brown, C.J., Ballabio, A., Rupert, J.L., Lafrenière, R.G., Grompe, M., Tonlorenzi, R. and Willard, H.F. (1991) Nature 349, 38–44.

[21] Borsani, B. et al. (1991) Nature 351, 325–329.

[22] Brockdorff, N. et al. (1991) Nature 351, 329–331.

[23] Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S. and Rastan, S. (1992) Cell 71, 515–526.

[24] Hong, Y.K., Ontiveros, S.D., Chen, C. and Strauss, W.M. (1999) Proc. Natl. Acad. Sci. USA 96, 6829–6834.

[25] Plath, K. et al. (2003) Science 300, 131–135.

[26] Silva, J. et al. (2003) Dev. Cell 4, 481–495.

[27] Chaumeil, J., Okamoto, I., Guggiari, M. and Heard, E. (2002) Cytogenet Genome Res. 99, 75–84.

[28] Calapez, A. et al. (2002) J. Cell Biol. 159, 795–805.

[29] Clemson, C.M., Mc Neil, J.A., Willard, H.F. and Lawrence, J.B. (1996) J. Cell Biol. 132, 259–275.

[30] Helbig, R. and Fackelmayer, F.O. (2003) Chromosoma 112, 173–182.

[31] Heard, E., Rougeulle, C., Arnaud, D., Avner, P., Allis, C.D. and Spector, D.L. (2001) Cell 107, 727–738.

[32] Costanzi, C. and Pehrson, J.R. (1998) Nature 393, 599–601.

[33] Wutz, A., Rasmussen, T.P. and Jaenisch, R. (2002) Nature Genetics 30, 167–174.

[34] Akhtar, A. (2003) Curr. Opin. Genet. Dev. 13, 161–169.

[35] Smith, E.R., Allis, C.D. and Lucchesi, J.C. (2001) J. Biol. Chem. 276, 31483–31486.

[36] Maison, C. et al. (2002) Nat. Genet. 30, 329–334.

[37] Lee, J.T., Davidow, L.S. and Warshawsky, D. (1999) Nat. Genetics 21, 400–404.

[38] Sado, T., Wang, Z., Sasaki, H. and Li, E. (2001) Development 128, 1275–1286.

[39] Sleutels, F., Zwart, R. and Barlow, D.P. (2002) Nature 415, 810–813.

[40] Morey, C., Arnaud, D., Avner, P. and Clerc, P. (2001) Hum. Mol. Genet. 10, 1403–1411.

[41] Shibata, S. and Lee, J.T. (2003) Hum. Mol. Genet. 12, 125–136.

[42] Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I. and Martienssen, R.A. (2002) Science 297, 1833–1837.

[43] Hannon, G.J. (2002) Nature 418, 244–251.

[44] Ogawa, Y. and Lee, J.T. (2003) Mol. Cell 11, 731–743.

[45] Wittschieben, B.O. et al. (1999) Mol. Cell 4, 123–128.

[46] Kornberg, R.D. (1999) Trends Cell Biol. 9, 46–49.

[47] Krogan, N.J. et al. (2003) Mol. Cell Biol. 23, 4207–4218.

[48] Ng, H.H., Robert, F., Young, R.A. and Struhl, K. (2003) Mol. Cell 11, 709–719.

[49] Travers, A. (1999) Proc. Natl. Acad. Sci. USA 96, 13634–13637.

[50] Gerber, M. and Shilatifard, A. (2003) J. Biol. Chem. 278, 26303–26306.

[51] Morey, C., Navarro, P., Debrand, E., Avner, P., Rougeulle, C. and Clerc, P. (2004) Embo J. 23, 594–604.

[52] Sanchez-Herrero, E., Vernos, I., Marco, R. and Morata, G. (1985) Nature 313, 108–113.

[53] Martin, C.H. et al. (1995) Proc. Natl. Acad. Sci. USA 92, 8398–8402.

[54] Bae, E., Calhoun, V.C., Levine, M., Lewis, E.B. and Drewell, R.A. (2002) Proc. Natl. Acad. Sci. USA 99, 16847–16852.

[55] Drewell, R.A., Bae, E., Burr, J. and Lewis, E.B. (2002) Proc. Natl. Acad. Sci. USA 99, 16853–16858.

[56] Engel, J.D. and Tanimoto, K. (2000) Cell 100, 499–502.

[57] Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R. and Fraser, P. (2000) Molecular Cell 5, 377–386.

[58] Mattick, J.S. (2001) EMBO Rep. 2, 986–991.