1

Non-coding RNA, Its History and Discovery Timeline

The complexity of eukaryotic transcriptomes and the rapid development of high throughput sequencing technology have led to an explosion in the number of long non-coding RNA (or lncRNA) recently identified and as yet undescribed. Current challenges include not only their precise labeling but also their functional characterization and the clinical pertinence of these biological objects. To begin with, it is important to return to the characteristics of RNA, a molecule essential to cellular functionality.

The timeline of discoveries linked to non-coding RNA is shown in Figure 1.1 and its history will be detailed in the first half of this chapter, from DNA to the first non-coding transcripts. Next, the role of global genomic and transcriptomic studies in changing our vision of RNA's capacity in gene expression regulation circuits and cellular complexity will be discussed. This functional diversity has given rise to various extensive classifications of lncRNA.

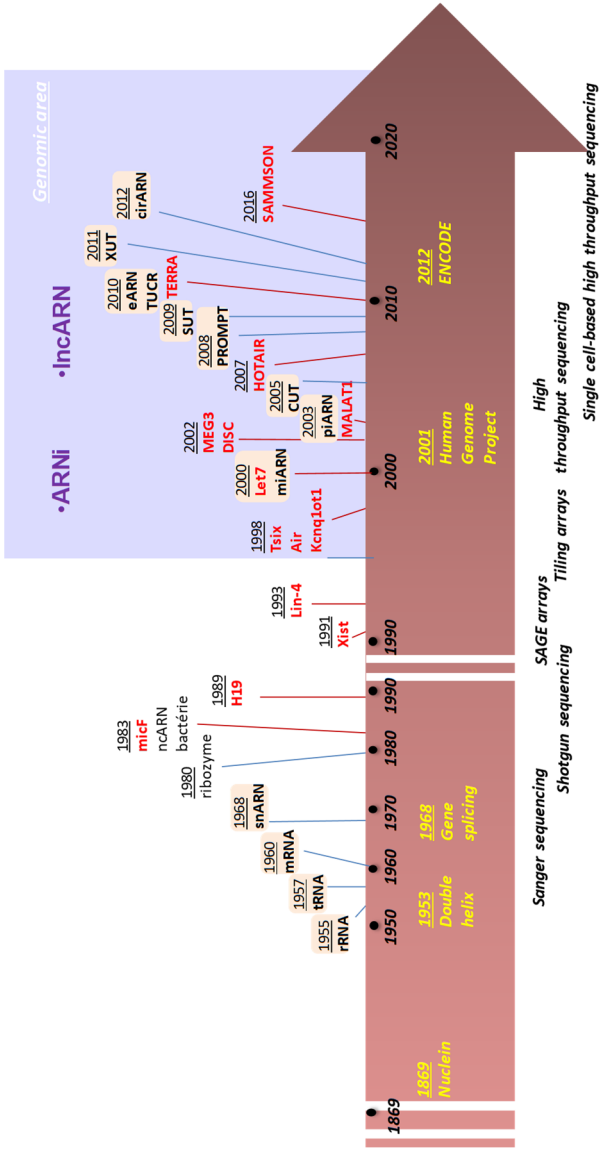


Figure 1.1. *Timeline of the main discoveries of RNA biology and in particular of eukaryotic non-coding RNA. For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*

NOTES ON FIGURE 1.1.– In light pink are the discoveries of large non-coding RNA families. In red are typical examples of specific non-coding RNA. In yellow are a selection of the big discoveries in molecular biology. Beneath the timeline are the novel technologies that enabled the characterization of non-coding RNA families. In violet are the discoveries of mechanisms regulated by non-coding RNA.

1.1. The biology of RNA, a century of history

The defining and conceptualization of RNA molecules in cell biology date back to 1869 and the discovery of nucleic acids. It took more than a century for researchers to finally identify non-coding transcription and to begin to suggest regulatory possibilities.

1.1.1. *From nuclein to the double helix*

At the end of the 19th Century, several essential discoveries foreshadowed the era of molecular biology. Although Friedrich Miescher isolated the contents of the cell nucleus, which he named "nuclein" (containing nucleic acid) (Dahm, 2005) in 1869, the interests of scientists at the time focused on the proteins thought to carry genetic information. It was only in 1944 that the link between nucleic acid (of the nucleus) and genetic information was made when Oswald Avery suggested that DNA was the carrier of genetic information (Avery *et al.*, 1944). As such, the association between the Mendellian genetic model and Miescher's "nuclein" was missing for more than half a century.

Following the visualization of the double helical structure of DNA by James Watson and Francis Crick in 1953, it was suggested, in 1961, that RNA might provide the intermediary molecule in the flow of information between

DNA and proteins (Cobb, 2015). Outlined in 1958 by Francis Crick and then by François Jacob and Jacques Monod, the central dogma of molecular biology included gene transcription of DNA to RNA in the nucleus followed by protein synthesis. It was also confirmed that the flow of information was only from DNA to RNA, then from RNA to the protein and never the reverse (Cobb, 2015). The mediator role of RNA has become a central focus for research, essential to the development of modern molecular biology.

1.1.2. *The "RNA world" concept*

It was only in 1955 that Georges Palade identified the very first non-coding RNA (ncRNA) that made up part of the most abundant RiboNucleoProtein (RNP) complex of a cell: the ribosome. Soon after, in 1958, a second class of ncRNAs were discovered by Mahlon Hoagland and Paul Zamecnik for their role as intermediaries between amino acids and RNA: transfer RNA (tRNA).

In 1960, François Jacob and Jacques Monod defined "messenger RNA" (mRNA) as the intermediate molecule carrying genetic information for protein synthesis. Following this, Crick and his team established that the genetic code is a universal non-overlapping triplet code in which three nucleotides code for one amino acid (Crick, 1968).

The discovery of heterogenous nuclear RNA (hnRNA) in the late 1960s led to the study of rRNA maturation and the discovery of splicing (Berk, 2016; Lewis *et al.*, 1975; Weinberg and Penman, 1968; Zieve and Penman, 1976).

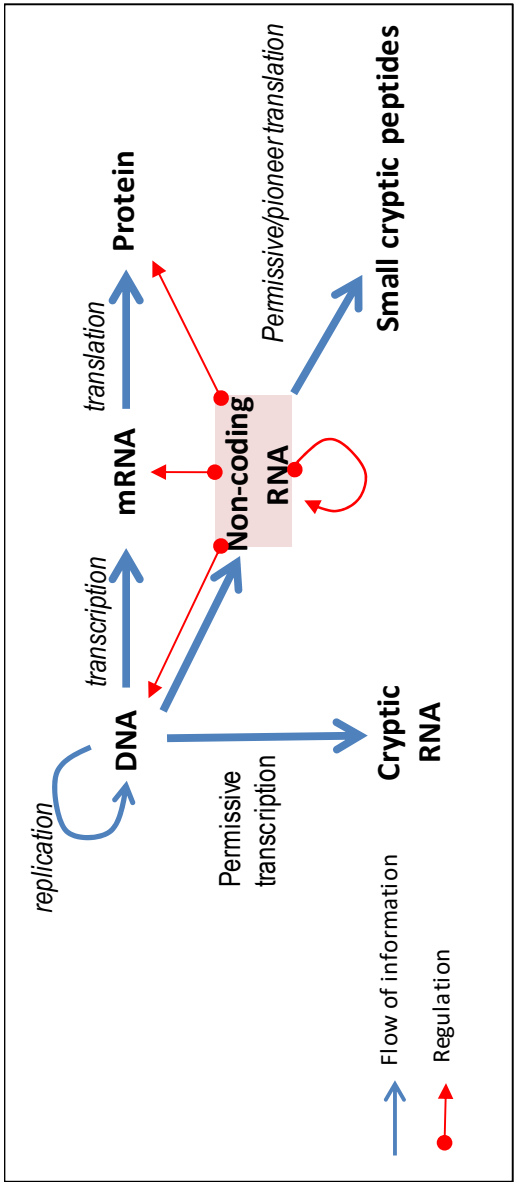


Figure 1.2. *Flow of genetic information and its regulators. For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*

The regulators in red in Figure 1.2 were not known in 1958 during the formulation of the initial dogma of information flow, nor were the permissive transcription and translation processes.

Although Jacob, Monod and Crick had all stated that RNA was not just a messenger, for a long time many scientists treated it as a simple unstable intermediate molecule, neglecting the potential active roles of other classes of ncRNA. However, this concept was partially changed in 1980 when Thomas Cech and Sidney Altman discovered that RNA molecules could also act as catalysts for chemical reactions such as self-splicing (Kruger *et al.*, 1982) or RNA degradation through ribonucleotide nuclease P, or RnaseP (Guerrier-Takada *et al.*, 1983). These RNA enzymes, or ribozymes, have since been acknowledged as key actors in the flow of genetic information (Figure 1.2), as part of both the ribosome and the spliceosome (Butcher, 2009; Cech, 2000).

The discovery of ribozymes led to the hypothesis of the "RNA world" which proposed that pre-biotic life revolved around RNA. Further studies of its roles in cell biology have revealed that RNA is necessary for DNA replication and that ribonucleotides are precursors of DNA's deoxyribonucleotides.

Furthermore, as previously mentioned, RNA plays an important role in all stages of protein synthesis whether as a template (mRNA) or as an actor (ncRNA: rRNA, tRNA, etc.) (Bernhardt, 2012). The latter are constitutively expressed in the cell and are necessary for vital cell functions. These maintenance ncRNAs are the subject of many specialized publications and will not be explored in this book. Other classes of regulatory ncRNA were discovered in the 1990s and will be broadly discussed here. The origin of those non-coding RNA and the fact they constitute an additional layer of genetic expression started to be discussed at this date.

Several hypotheses were raised, especially the interconnections with introns (Mattick, 1994; Morris and Mattick 2014). The origin and extension of the "RNA world" concept was increasingly acknowledged. These ncRNAs are expressed very specifically throughout the stages of embryonic development, in certain tissues or pathological states and also play multiple roles in the regulation of gene expression and genomic stability.

1.1.3. *Small bacterial RNA: pioneers of non-coding RNA*

The very first ncRNA regulator defined was micF, from the Escherichia coli bacteria. It was described in 1987 by Masayuki Inouye and his team (Inouye and Delihas, 1988) as the first RNA that regulates the expression of a gene through sequence complementarity, and represents the main class of bacterial small ncRNA regulators (sRNA). It was shown that the ncRNA of micF suppresses the translation of a target mRNA coding for a porine (outer membrane protein F, OmpF) that is involved in passive transport across cell membrane. When micF transcription is activated, it inhibits the expression of the ompF gene both at the mRNA level and within the protein.

Following the *in vitro* characterization of RNA's duplex structure, it was shown that micF binds to the ribosome binding site (RBS) of the ompF mRNA on the Shine-Dalgarno sequence, thus inhibiting the translation of ribosomes (Figure 1.3).

More recently, it has been shown that the regulation of gene expression by micF extends to other genes, including the mRNA lrp (Leucine responsive protein) (Corcoran *et al.*, 2012). Lrp is a transcription factor that regulates the expression of genes in the E. coli bacterium in reponse to osmotic changes and the availability of nutrients. Remarkably, Lrp regulates the expression of micF and vice

versa, thus creating a feedback loop and proving the important role of micF in the overall regulation of genes and metabolism. Similar mechanisms have been found in the Salmonella bacteria, indicating a conservation of this regulatory pathway through evolution (Delihas, 2015). Since then, many other sRNA, whose lengths vary from 50 to 500 nucleotides (nt), have been discovered, including ncRNAs that act in *trans* or *cis*, RNA thermometers and riboswitches. They all work in pairs, thus inhibiting the translation of targeted mRNAs and inducing their degradation.

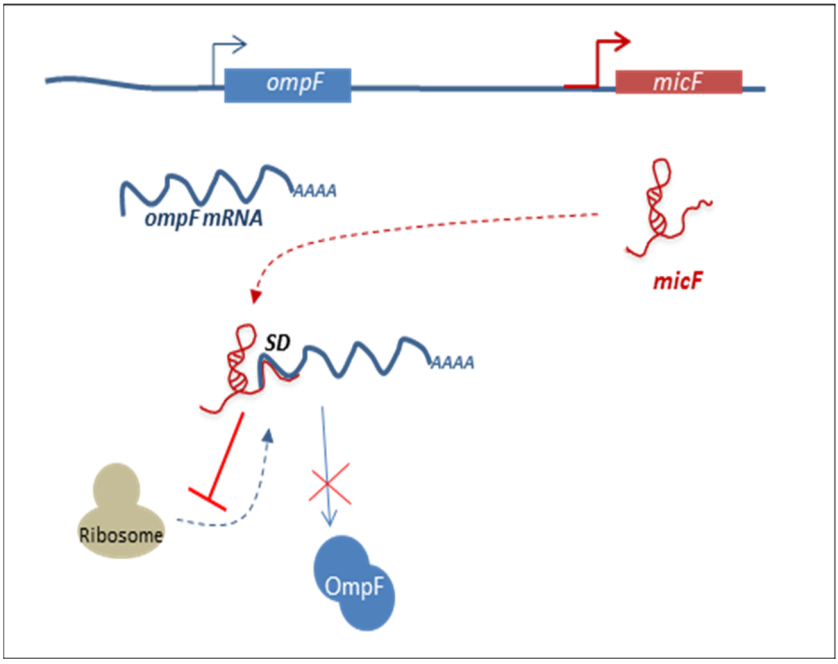


Figure 1.3. *Regulation of ompf expression by the non-coding RNA micF. For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*

NOTES ON FIGURE 1.3.– micF partially binds to the 5' region of the ompF mRNA blocking recognition of the Shine-Dalgarno (SD) site by the ribosome, therefore also blocking the translation of mRNA into the OmpF protein.

1.1.4. *Micro-RNA and RNA interference*

In the early 1990s, through co-expression of transgenes or viral infection experiments in different eukaryotic organisms, several scientists observed a phenomenon of protein synthesis inhibition mediated by RNA. The regulatory effects of these molecules would dramatically change the perspective of RNA as a simple messenger. Initial studies described the phenomenon as "co-suppression" in plants, as "post-transcriptional gene silencing" (PTGS) in nematodes or as "attenuation" in fungi, but none of them suspected it to be RNA itself. The first micro (mi)RNA was characterized in the Caenorhabditis (C.) elegans nematode in 1993 by Victor Ambros and his colleagues. Ambros discovered that the lin-4 gene produced small RNAs of 22 and 61 nts from a longer non-coding precursor. The precursor RNA forms a stem-and-loop structure that is cut to create shorter RNA strands with an antisense complementarity in the untranslated 3' region (UTR: Un-Translated Region) of the lin-14 transcript (Lee *et al.*, 1993). The pairing of lin-4 RNA with lin-14 mRNA was proposed as a molecular "post-transcriptional gene silencing" mechanism, thus decreasing the levels of LIN-14 protein in the initial larval stages of nematode development (Wightman *et al.*, 1993). Michael Wassenegger observed a similar phenomenon in plants that he described as "homology-dependent gene silencing". This process is mediated by the incorporation of viral RNA that induces the methylation of viral cDNA and the silencing of the gene expression (Wassenegger *et al.*, 1994). The whole process of "gene silencing through RNA" was put forward in 1988 by Andrew Fire and Craig Mello in similar experiments with the *C. elegans* unc-22 gene (Fire *et al.*, 1998).

In 2000, let-7, another essential miRNA, was identified in *C. elegans* and as homologs in several other organisms including humans (Ameres and Zamore, 2013; He and Hannon, 2004). Biogenesis, as well as the molecular

mechanisms of regulation through miRNA were largely characterized by Thomas Tuschl. His team showed that in *C. elegans,* a long double-stranded RNA is transformed into shorter 21-25 nts fragments. Next, it was shown that the premature transcripts in the nucleus are transformed into hairpin-shaped RNA by the microprocessor complex containing the endonuclease protein, Drosha, before being exported to the cytoplasm where they are cleaved into small double strands by the type III RNAse protein, Dicer. One of the strands passes through the RISC complex and is then directed through complementarity to an mRNA molecule, thus inducing a repression in translation (He and Hannon, 2004). This simplified schematic (Figure 1.4) constitutes the basis of the interference mechanism of RNA (RNAi) and actually combines all the gene silencing phenomena mediated by small ncRNA including miRNA, small interfering RNA (siRNA) and Piwi-interacting RNA (piRNA) at the transcriptional (TGS) and post-transcriptional levels (PTGS) (Castel and Martienssen, 2013; Montgomery, 2004).

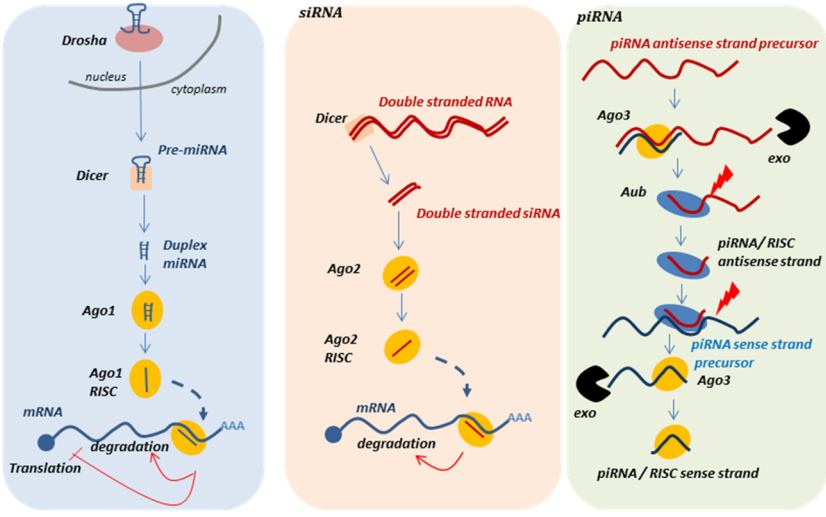


Figure 1.4. *Model of RNA interference. For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*

NOTES ON FIGURE 1.4.– Three modes of maturation of RNA miRNA, siRNA and piRNA. For the miRNA, in the nucleus, Drosha cleaves the stem-loop of pre-miRNA RNA which is then exported into the cytoplasm where DICER matures the loop and generates a duplex miRNA taken up by AGO1 that unravels it into a single strand RNA. The AGO1/RISC complex pairs with a target mRNA to prevent its translation and/or initiates its degradation. For siRNA, the double-stranded RNA is cleaved by DICER, generating small double-stranded RNA which is taken up by AGO1 to unravel it into a single strand. The AGO2/RISC complex targets an mRNA and initiates its degradation. For piRNA, a single strand precursor is recognized by an antisense strand carrying AGO3 which protects the double strand from degradation with an exonuclease. AUB cleaves the remaining single strand and forms the piRNA/RISC precursor. This targets an antisense RNA for cleavage while AGO3 protects the single strand from complete degradation through an exonuclease and forms the piRNA/RISC sense strand. This return phenomenon is called "ping-pong".

Although RNA interference resulted in a breakthrough for modern biology and biotechnology and the emergence of the concept of non-coding RNA as a regulator, the pertinence of the role of lncRNA remains largely unknown. However, some lncRNAs such as H19 and Xist were studied in the late 1980s and led to milestones in dosage compensation mechanisms in mammals. Dosage compensation, also called genomic imprinting or parent-of-origin specific expression, is fundamental to the diversity of gene expression in a lineage. Essentially, the parental alleles of the chromosomes inherited from the male or female are not always expressed in the same way. Certain domains inherited from the mother or father are switched off or overactive depending on a marker left on the DNA in the region inherited from the

parent. The fundamental discovery of Xist and H19 put lncRNA where it is now, as a key factor in the establishment and maintenance of the parental imprint.

1.2. The discovery of long non-coding RNA in the pre-genomic era

In the 1980s, scientists used differential hybridization methods of cDNA libraries to clone and study non-coding genes. Later, an *a posteriori* approach was adopted without regard for the coding potential of RNA. Thanks to this approach, H19 became the first non-coding gene discovered even though at the time, it was initially classed as mRNA (Pachnis *et al.*, 1984).

1.2.1. *H19: the first in the history of long non-coding RNA*

The detailing of the genomic imprinting phenomenon, of parent-of-origin specific gene expression, forms a part of the dosage compensation mechanisms that work by sequestering genes in an epigenetic manner (see glossary). Independently, two "imprinted" genes were identified: Igf2r coding for the paternally expressed protein and H19, expressed by the mother (Figure 1.5). The two genes were located on chromosome 7 in mice, near the H19/IGF2 cluster (Barlow *et al.*, 1991; Bartolomei *et al.*, 1991). What makes H19 unusual is the absence of its translation into a protein, even though the gene contains small open reading frames. H19 is conserved in mammals and the transcript, in abundance, presents mRNA characteristics: transcribed by polymerase II RNA, spliced, polyadenylated at the 3' end and exported to the cytoplasm (Brannan *et al.*, 1990). The heterologous expression of H19 in transgenic mice proved to be fatal at prenatal stages, which not only suggested that the dosage of this lncRNA was strictly controlled, but that it also played an important role in embryonic development. However, the function of H19 as an RNA molecule remained a mystery

until the functional characterization of another lncRNA involved in the dosage compensation of the mammalian X chromosome, Xist, was defined. Since then, H19 has been the subject of extensive study and represents the prototype of an lncRNA with multiple functions.

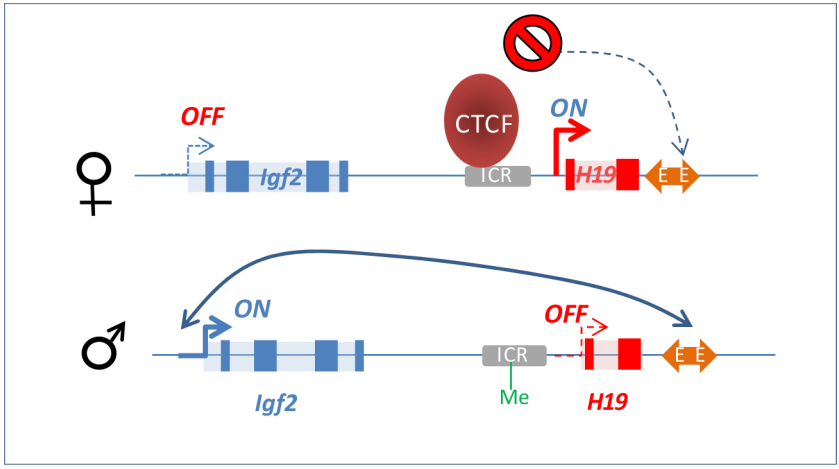


Figure 1.5. *Regulation of parental imprinting and expression of H19. For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*

NOTES ON FIGURE 1.5.– On the maternal allele, binding of the CTCF repressor on the ICR region prevents the Igf2 activator sequence enhancers (E) from touching the IGF2 promoter. The maternal gene is thus switched off and the lncRNA H19 is activated. The opposite is seen on the paternal allele where the methylation of the ICR DNA (Me), inherited from paternal chromosomes, prevents the CTCF factor from binding and thus allows the activation of the IGF2 gene by enhancer sequences. The methylation of ICR represses H19. Paternal IGF2 is therefore switched ON.

1.2.2. *Inactivation of X, the eXISTence of XIST*

In living organisms, gender can be determined in many ways and is defined in mammals by the X and Y chromosomes: whereas males have a single copy of the X and the Y chromosomes, females have two X chromosomes in their karyotype. The X chromosome, although involved in male and female gender differentiation, carries many genes whose functions are not a part of gender determination. As a consequence, there is a need for dosage compensation between males and females to enable these genes to be expressed in an equivalent manner.

In Drosophila, the dosage compensation mechanism consists of doubling the transcription of the single X chromosome in males. In mammals, one of the two X chromosomes in females is inactivated. This phenomenon, called X Chromosome Inactivation (XCI), was first discovered in mice by Mary Lyon in 1961 (Lyon, 1961) and later generalized to other mammals. XCI is established in the first stages of development and initiated by a single locus, the X inactivation center (Xic).

In the early 1990s, it was shown that this locus produced a long non-coding RNA, XIST (X-inactive-specific-transcript), weakly expressed in both the male and female undifferentiated embryonic stem cells (ES). During differentiation, the expression of XIST is randomly activated by a single allele, in female cells, by the future inactivated X (Xi). Retained in the nucleus, XIST sets off gene silencing in *cis* (Figure 1.6), physically placing itself and spreads widely on the future Xi (Borsani *et al.*, 1991; Brockdorff *et al.*, 1991; Brown *et al.*, 1991).

Unlike H19 and other lncRNA involved in dosage compensation, XIST is very unusual because it leads to silencing in the whole chromosome. The spread of XIST along Xi, called “coating”, requires the enveloping of RNA

around the X and the recruitment of multiple factors, including the repressor complexes Polycomb 1 and 2 (PRC1 and PRC2). These trigger a cascade of changes in the chromatin and a spatial reorganization of the entire Xi before finally resulting in the stable repression of nearly all the genes linked to Xi throughout the course of development and adult life (Gendrel and Heard, 2014). While the expression of XIST is crucial for the initiation of XCI, in somatic cells XIST and Xic are not indispensable for the maintenance of gene silencing in mice (Brown and Willard, 1994). XIST does not only play a role in stable inactivation (Heard *et al.*, 1999). Xic was first defined in mice as the minimum region on the X chromosome that contained all the necessary and sufficient sequences for the initiation of XCI. Xic is spread over a million base pairs and transcriptomic studies have revealed that this region contains several coding and non-coding genes, including Linx, Ftx and others. Some non-coding genes in Xic show poor sequence conservation between man and mouse, including the XIST sequence itself (Chureau *et al.*, 2002). Tsix in particular, characterized in mice, is an antisense transcript that overlaps the whole XIST gene (Figure 1.6).

In humans, the main Tsix regulator elements are truncated and the transcript only overlaps XIST at the 3' end. These differences remove Tsix's role in the transcriptional repression of XIST on the future X chromosome that is active in humans (Lee and Lu, 1999; Migeon *et al.*, 2002). Recently, another lncRNA, Xact, has been discovered in human ES cells. This gene is located in the intergenic region, outside of Xic, and is not conserved in mice. In human female stem cells, Xact is expressed and covers both X chromosomes. This lncRNA appears specific to pluripotent cells and may ensure a specific control of XCI in humans (Vallot *et al.*, 2013). Over the last few years, the biogenesis of XIST, its structure and the molecular mechanism of XCI have been at the heart of many

development studies in different mammals (Gendrel and Heard, 2014).

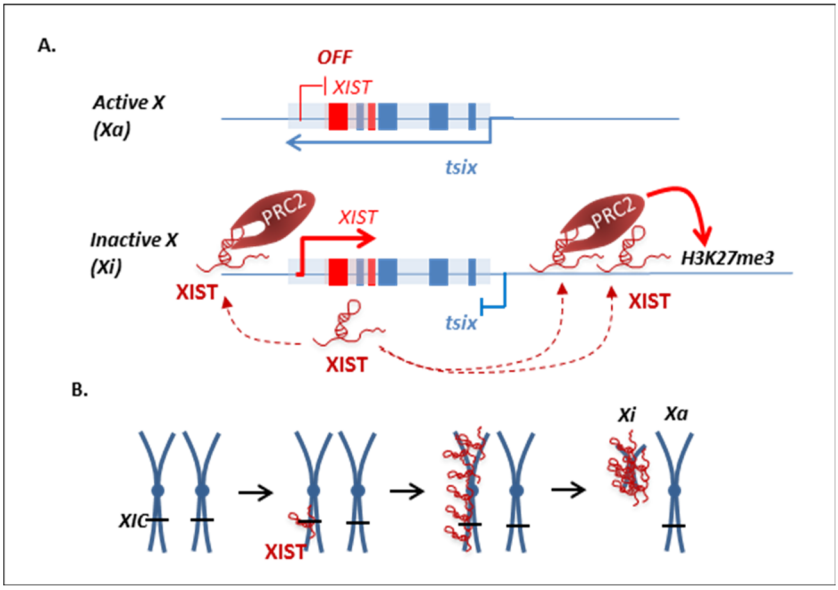


Figure 1.6. *Regulation of the inactivation of the X chromosome by XIST. For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*

NOTES ON FIGURE 1.6.–

A. The active X chromosome (Xa) transcribes the XIST antisense (Tsix) which represses the expression of XIST. The inactive X (Xi) shows a weak rate of transcription of Tsix and XIST is activated. It spreads along the length of the X chromosome, recruits/activates PRC2 which deposits the H3 lysine 27 trimethyl histone (H3K27me3), a marker of chromatin repression. The large majority of the inactive X chromosome genes are thus repressed..

B. Stages of X chromosome inactivation. XIST is expressed on the X inactivation center (XIC) on one of the 2 chromosomes, spreads and inactivated the whole chromosome (Xi).

These pioneering studies on H19 and XIST revolutionized our vision of non-coding RNA functions and the biological pertinence of lncRNA in general. It has been suggested in particular that non-coding RNA could provide a link between the genetic code and the epigenetic code to modulate genome accessibility in order to finely regulate its expression throughout both its cellular life and development (Figure 1.7). These examples have shown the complexity and versatility of regulation circuits orchestrated by a single lncRNA. They have also stimulated further discoveries and suggestions of potential mechanisms for other non-coding transcripts that remain uncharacterized. A global effort in the identification and characterization of lncRNA was launched in the 2000s with the complete sequencing of the human genome and through the ENCODE (Encyclopedia of DNA Elements) project (Djebali *et al.*, 2012).

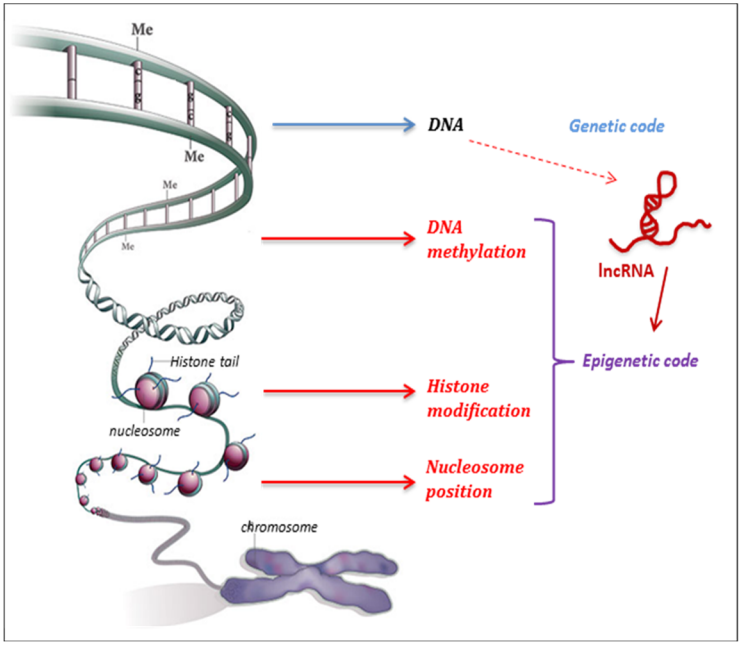


Figure 1.7. *Non-coding RNA, the link between the genetic and epigenetic code. For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*

NOTES ON FIGURE 1.7.– Through sequence complementarity, non-coding RNA can place epigenetic markers on the DNA sequence by modulating, targeting, and activating DNA modification complexes (Methylation, histone position or post-transcriptional modification).

1.3. From the non-coding genome to the non-coding transcriptome, the advent of the genomic era

Our modern vision of eukaryotic transcriptomes was preceded by in-depth studies of genomic DNA and the discovery that, besides protein coding (PC) sequences and the essential regulator elements for the transcription of PC genes (PCG), a large majority of the genome contained sequences considered as evolutionary fossils that are, therefore, useless. For about 20 years, to differentiate these sequences from PC sequences, these regions of DNA were called non-coding and considered as "selfish" DNA (often truncated repeat elements) or "undesirable" (Orgel and Crick, 1980).

1.3.1. *The human genome project: genomic DNA is essentially non-coding*

In 1978, Frederick Sanger generated the very first complete genomic sequence : that of the viral genome of the ϕX174 bacteriophage (Sanger *et al.*, 1978). Since then, Sanger sequencing has been widely used across the world. Its discovery and development earned Sanger and Walter Gilbert the Nobel Prize for chemistry. Throughout the following years, several viral genomes were sequenced and in the late 1990s, a complete human genome sequencing project, the Human Genome Project (HGP), was launched by the National Health Institute (NIH, United States). In parallel, the American biochemist and entrepreneur Craig Venter founded his own company to reach the same goal and

thus the race to unravel the human genome began. The first bacterial genome was published in 1995 (Fleischmann *et al.*, 1995), followed in 1999 by the sequencing of the euchromatic part of the human chromosome (Dunham *et al.*, 1999) which covers around 65% of what is now known as chromosome 22. Finally, a first copy of the complete human genome was published in *Nature* in 2001, covering 96% of the euchromatin (Lander *et al.*, 2001b), followed the very next day by Craig Venter's publication in *Science* of the whole genome sequence (Venter *et al.*, 2001). Since 2003, regular updates have completed these initial sequences. At the same time, the genomes of several other organisms were obtained, notably that of yeast, the *pufferfish* fish, the *C. elegans* worm, the *D. melanogaster* fly and the mouse (Adams *et al.*, 2000; Goffeau *et al.*, 1996; Roest Crollius *et al.*, 2000; Waterston and Sulston, 1995; Waterston *et al.*, 2002), allowing for advanced comparative studies to be carried out.

The first surprise to come out of this complete genomic sequencing effort was the number of rather weak PCG compared with what was expected. Initial studies predicted 70,000 to 80,000 coding genes in the human genome (Antequera and Bird, 1993), and a figure closer to 100,000 genes was suggested in the mid-1980s. The whole sequencing results were predicted to be around 31,000 PCG in 2001, which was then decreased to 22,287 PCG in 2004 (consortium, 2004; Lander *et al.*, 2001b) and 19,600 is suggested in the current Gencode V25 version (Wright *et al.*, 2016). Overall, the results are astounding: gene coding exons represent only 1.2% of the human genome, while 24% and 75% are thought to be introns or non-coding intergenic DNA (Figure 1.8). This, however, is not the only surprise. By comparing the proportions of coding and non-coding genome, it became apparent that the percentage of non-coding evolves spectacularly with the "complexity" of the organism, even if the notion of complexity itself remains to be defined (Figure 1.9). Either way, it is striking to consider that 25 to 50% of bacterial or unicellular eukaryotic genomes are non-coding

whereas in vertebrates this figure lies at over 95%. The idea that the non-coding genome is somehow involved in specialization and cell identity therefore begins to pave its way.

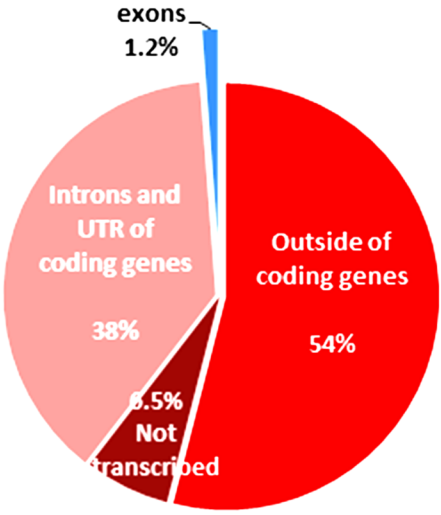


Figure 1.8. *Non-coding genome. Proportion of transcribed and non-transcribed regions in genes coding for proteins (exons) and non-coding sequences (introns, UTRs and outside of genes) in the human genome, according to ENCODE. For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*

1.3.2. *Permissive transcription and the hidden face of the genome*

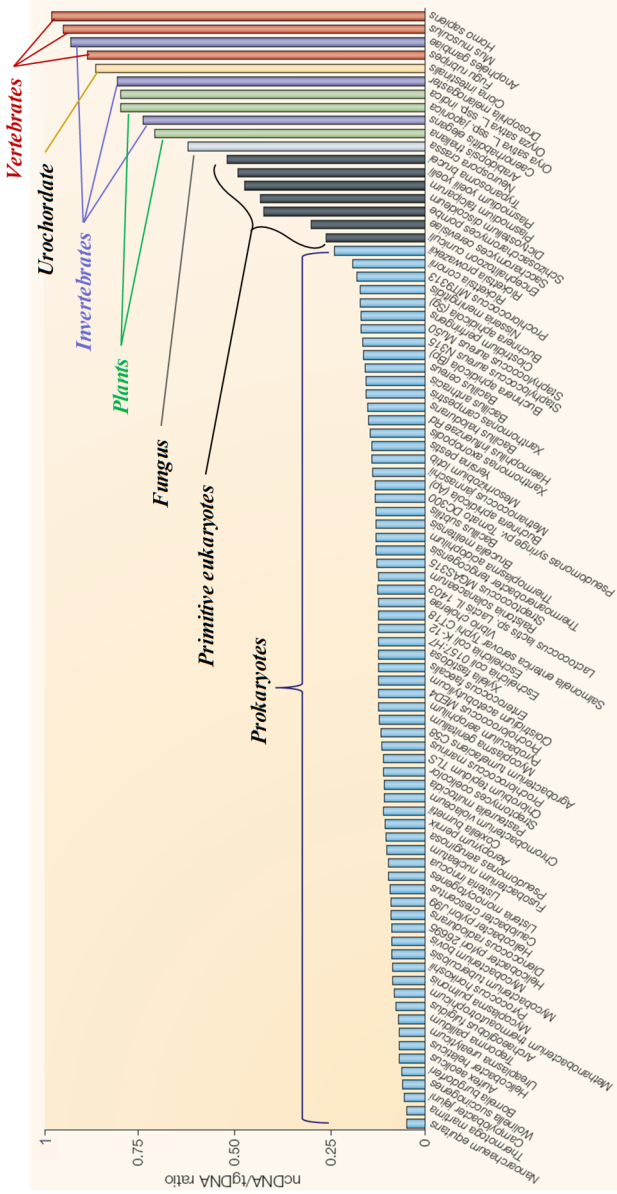
Although the idea of genome fossilization (through inert strata of non-coding genes) continued to be discussed, the notion was shattered when the HGP revealed that the majority of the genome is transcribed, whether protein-coding or not. The use of oligonucleotide sensors covering human chromosomes 21 and 22 had previously revealed that 90% of the cytosolic polyadenylated transcripts detected

corresponded to non-coding genomic regions and not to exons (Kapranov *et al.*, 2002). Similar results were revealed by the FANTOM and RIKEN consortia following the analysis of human and mouse transcriptomes (Carninci *et al.*, 2005; Okazaki *et al.*, 2002). More than 60,000 complete mouse cDNA were sequenced to generate precise maps of the 5' and 3' ends of all transcripts, thus defining the transcription start sites (TSS) and termination sites (TTS). Remarkably, more than 23,000 ncRNA resulting from the transcription of both sense and antisense genes represented around two thirds of the mouse genome (Katayama *et al.*, 2005). For the first time, antisense transcription was suggested as a contributor in gene expression regulation at the transcription level in mammals (Katayama *et al.*, 2005).

These results were later confirmed by further in-depth studies on human genomes carried out by the ENCODE (Encyclopedia of DNA Elements) Consortium. In its pilot phase, this project involved more than 200 experiments (Birney *et al.*, 2007) and subsequently up to 1,640 databases from 147 different cell lines were defined (consortium, 2012). Thanks to various sequencing techniques, the landscape of DNase I hypersensitive sites, histone modification sites, transcription factor binding sites and collection of transcriptomes has been defined. These studies concluded that around 93% of the human genome is actively transcribed and associated with at least one primary transcript (coding and non-coding exons and introns). Of these transcripts, around 39% of the genome represent PCGs (of the poly-A promoter), 1% are coding exons while the remaining 54% are not PCG (Figure 1.8). More recently, the mouse counterpart of the ENCODE Consortium confirmed previous studies by publishing a similar analysis showing that 4 to 6% of the mouse genome produces mRNA whereas at least 87% of its genome is transcribed (Stamatoyannopoulos *et al.*, 2012; Yue *et al.*, 2014).

Many studies that aim to characterize non-coding transcription have also been carried out on other eukaryotic organisms, including *Saccharomyces cerevisiae*. Even in this primitive eukaryotic unicellular organism, around 85% of the genome is transcribed (David *et al.*, 2006). This trait of transcribing the majority of the genome, a phenomenon often called "permissive transcription", is spread throughout eukaryotes and an increasing number of publications are discussing its potential function(s) (Berretta and Morillon, 2009; Dinger *et al.*, 2009). The identification and characterization of non-coding transcripts has expanded the old definition of a "gene" beyond its protein-coding function. Furthermore, the discovery of both the non-coding genome and the permissive transcriptome has generated heated debates within the scientific community concerning the biological significance and functional pertinence of this DNA and these non-coding RNA, still perceived as dark matter (Clark *et al.*, 2013; Dinger *et al.*, 2009; Mattick, 2003). These debates have challenged Watson and Crick's central dogma of molecular biology by promoting ncRNA at the epicenter of cell processes, as a dynamic motor of biological complexity (Figure 1.9).

The analysis of coding and non-coding RNA expression in the 14 different ENCODE cell lines (Djebali *et al.*, 2012) highlighted the exceptional specificity of the lncRNA for each cell type. While more than 50% of the RNA have a comparable level of expression throughout the 14 cell lines (Figure 1.10), more than 50% of the lncRNA are only detectable in a single cell line. These striking results suggest that lncRNA might have a determining role in cell identity.


Vertebrates
Urochordate
Invertebrates
Plants
Fungus
Primitive eukaryotes
Prokaryotes
ncDNA/tgDNA ratio
1
0.75
0.50
0.25
0
Homo sapiens
Mus musculus
Amphioxus gamboes
Fugu rubripes
Ciona intestinalis
Drosophila melanogaster
Oryza sativa L. ssp. indica
Caenorhabditis elegans
Arabidopsis thaliana
Neurospora crassa
Trichomonas vaginalis
Plasmodium falciparum
Dictyostelium discoideum
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Encephalitozoon cuniculi
Pichia stipitis
Candida albicans
Ashbya gossypii
Kluyveromyces lactis
Neosartorya fumigatus
Phycomyces blakesleeanus
Coprinus cinereus
Ustilago maydis
Cryptococcus neoformans
Aspergillus nidulans
Neurospora crassa
Fusarium graminearum
Magnaporthe grisea
Botrytis cinerea
Sclerotinia sclerotiorum
Alternaria alternata
Cladosporium fulvum
Pyrenophora tritici-repentis
Phaeosphaeria nodorum
Ustilago hordei
Cryptococcus neoformans
Malassezia globosa
Tremella mesenterica
Pleurotus ostreatus
Lentinula edodes
Agaricus bisporus
Phanerochaete chrysosporium
Coprinopsis cinerea
Schizophyllum commune
Trichoderma reesei
Aspergillus niger
Aspergillus fumigatus
Aspergillus oryzae
Aspergillus terreus
Aspergillus flavus
Aspergillus nidulans
Penicillium chrysogenum
Penicillium marneffei
Penicillium chrysogenum
Penicillium marneffei
Penicillium chrysogenum
Penicillium marneffei

NOTES ON FIGURE 1.9.– Graph representing the proportion of the genome that is non-coding (non-exonic) versus the coding genome (exons) for the sequenced genome of each species (on the abscissa). In light blue: the bacteria; in black: simple eukaryotes (unicellular); in pastel blue: a fungus; in green: plants; in dark blue: invertebrates; in yellow: a urochordate and in red are the vertebrates. Adapted from (Mattick, 2004).

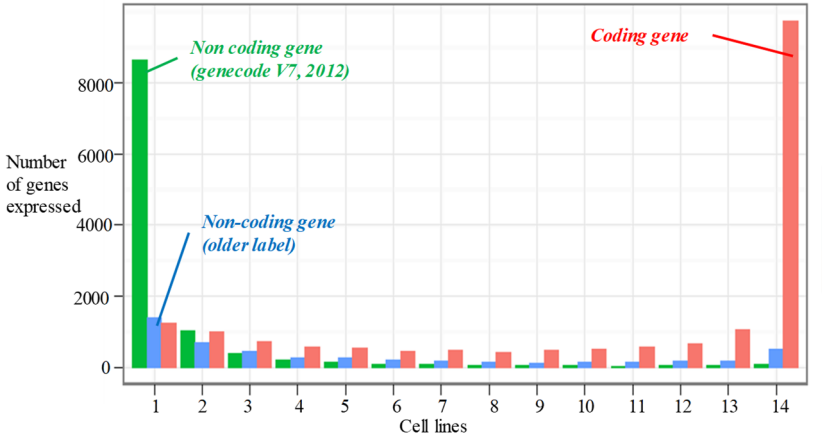


Figure 1.10. *Cell specificity of non-coding gene expression. The number of non-coding genes (coding genes in red; non-coding in green and blue with the 2012 ENCODE label and older label) of which one RNA was detected in one or more cell lines sequenced in the ENCODE project (Djebali et al., 2012). For a color version of this figure, see www.iste.co.uk/morillon/RNA.zip*