THE DARK MATTER OF THE GENOME
SOME INSIGHTS AND CLINICAL APPLICATIONS

ALFRED GRECH & MICHAEL BALZAN

ABSTRACT

Only approximately 1.5% of the human genome encodes protein sequence; the rest is 'dark matter'. Research on these noncoding regions shows that they play roles in cellular homeostasis, development, differentiation and metabolism. Cancer, cardiovascular, developmental, and neurological diseases are characterised by aberrant expression of these regions. Exploring their clinical utility as biomarkers and molecular targets in medical theranostics is a very promising way forward.

INTRODUCTION

It is now well known that only approximately 1.5% of the human genome encodes protein sequence.[1] However, comparative analyses with mammalian genomes have shown that at least 5% is under selective constraint and thus probably functional, of which approximately 3.5% consists of noncoding elements with apparent regulatory roles.[2] Collectively, this created an aura of mystery, leading to the label of 'dark matter', in a manner analogous to the 'dark matter' of the universe, which we can neither easily detect nor understand, but that nonetheless exists and is open to experimental queries. Ongoing research on these noncoding regions, which form a major part of this once proverbial genomic 'dark matter', shows that they play vital biological roles in cellular homeostasis, development, differentiation and metabolism. Indeed, their aberrant expression is being found in a variety of human diseases, including cancer, cardiovascular, developmental, and neurological diseases. Consequently, translational research is exploring the clinical utility of these noncoding RNAs (ncRNAs) as biomarkers and molecular targets in medical theranostics.

THE DARK MATTER IN THE CLINIC

ncRNAs represent a significant portion of the human transcriptome. Based on their size, ncRNAs are grouped into two major classes, namely, small ncRNA and long ncRNA (lncRNA). microRNAs (miRNAs, approximately 22 nucleotides long) and transcription initiation RNAs (tiRNAs, 18 nucleotides long) are two examples of the first class. In contrast, lncRNAs, which resemble mRNA transcripts, range from 200 nucleotides to approximately 100 kilobases.[3] In humans, lncRNAs have been identified to be transcribed from four chromosomal regions, termed the *Hox* gene loci. These four *Hox* loci (*Hoxa*, *Hoxb*, *Hoxc* and *Hoxd*) include dozens of genes that are involved in a variety of biological processes, including embryonic development, cell differentiation and tumorigenesis.[4]

Several lncRNAs are coded from regions between the genes in these *Hox* clusters, hence their other name being long intergenic non-coding RNA, or lincRNA. Increasing numbers of lncRNAs are being identified and their functions investigated. In fact, an emerging function is their role in genome modification, where they associate with Polycomb proteins to epigenetically silence genes. Specifically, this can occur through histone tail post-translational modifications, with methylation of histone H3 lysine 9 (H3K9me), lysine 27 (H3K27me), and histone H4 lysine 20 (H4K20me) being associated with regions of the genome that are transcriptionally inactive. Such silencing of genes through histone methylation is thought to be mediated by chromatin modelling complexes such as the Polycomb repressive complexes (PRC), PRC1 and PRC2. In this review, we will focus on what are perhaps the three most valued Polycomb-related lncRNAs in the clinical setting, i.e. ANRIL, HOTAIR, and XIST.

1. ANRIL

Spanning 126.3 kilobases in the genome, ANRIL is an antisense ncRNA in the INK4 locus. The INK4b (p15)–ARF (p14)–INK4a (p16) locus, which is found on chromosome 9p21, is said to be an essential regulator of cellular senescence. INK4 carries out this regulatory role by coding for three tumour suppressors i.e. p14 which increases p53 signalling, and p15 and p16, which (a) promote the function of the retinoblastoma protein pRB, and also, (b) inhibit cyclin-dependent kinases therefore causing cell cycle arrest. Regulation of the INK4 locus is governed by the Polycomb repressive complexes PRC1 and PRC2, where PRC2 initially trimethylates H3K27 in the transcriptionally silent heterochromatin, and then PCR1 recognises the methylated H3K27 as a sign to maintain the heterochromatin. Both *cis*- and *trans*-acting lncRNAs recruit Polycomb complexes to establish the heterochromatin. In this case, PRC1 and PRC2 are recruited to the INK4 locus by the lncRNA ANRIL, which is expressed antisense to the p14 and the p15 tumour suppressors.

It has been suggested that both Polycomb repressive complexes are recruited in *cis* to the INK4 locus gene through association with nascent ANRIL transcripts. Such a suggestion was made following results from a study showing that ANRIL knockdown leads to the upregulation of p15 and p16. Furthermore, the transcriptional state of the locus, which is often deleted or silenced in cancer, appears to be affected by changes in ANRIL expression.[5] Upregulation of ANRIL is seen in prostate cancer tissues for instance,[6] and in heart disease, type 2 diabetes, and risk-associated single-nucleotide polymorphisms (SNPs) for cancers overlapping with the ANRIL region.[7] One SNP in the 9p21 gene desert was also shown to be associated with coronary artery disease; this DNA variant disrupts the binding site for the STAT1 transcription factor which is known to represses the expression of ANRIL. Therefore, by stopping STAT1 from binding, it leads to the upregulation of ANRIL, and the cause behind coronary artery disease might well be the ANRIL-mediated silencing of p15.[8] Similar to ANRIL is the lncRNA HEIH which was also found to regulate the INK4 locus, where by recruiting PRC2 to tumour suppressors, it facilitates hepatocellular carcinoma tumorigenesis.[9]

2. HOTAIR

HOTAIR is one of the recently identified lncRNAs. It is a 2,158-nucleotide-long, spliced and polyadenylated lncRNA, encoded by a 6,232 base pair gene, located in the *Hoxc* cluster on chromosome 12 (specifically at 12q13). Only one strand of HOTAIR, which is antisense to the canonical *Hoxc* genes, is transcribed; hence its name, standing for *Hox* Antisense Intergenic RNA.[10] Unlike other documented lncRNAs that act strictly in *cis* (such as XIST), HOTAIR is the first lncRNA that is said to function in *trans*, because it is transcribed by one chromosome (chromosome 12), but regulates chromatin domains on another chromosome.[11] HOTAIR exists only in mammals, has been highly conserved in primates throughout evolution, and has evolved faster than nearby *HoxC* genes. Poorly conserved sequences are present in its six exons, except for a 239 base pair domain in exon 6, which is particularly conserved.[12]

Presently, the proposed functional mechanism of HOTAIR is to act as a scaffold for the recruitment and binding of the polycomb complex PRC2 and lysine-specific demethylase 1 (LSD1). PRC2 and LSD1 are multisubunit protein complexes that epigenetically modify chromatin. HOTAIR is believed to recruit these two complexes to regions of the genome so as to bring about gene silencing. For this reason, HOTAIR is emerging as an important player in tumorigenesis. It was found that high levels of HOTAIR are linked with metastatic spread and poor survival rate in breast cancer.[13] Specifically, HOTAIR was shown to be highly upregulated in primary and metastatic breast tumours, even up to two-thousandfold over normal breast tissue. HOTAIR expression levels were also found to correlate with metastasis in colorectal cancer,[14] gastrointestinal stromal tumours,[15] hepatocellular carcinoma,[16-17] and pancreatic cancer.[18]

THE MAIN CHALLENGE IN INTRODUCING ncRNA-BASED THERAPEUTICS INTO CLINICAL PRACTICE IS THE DELIVERY AND THE OFF-TARGET EFFECTS

3. XIST

XIST, or X inactive specific transcript, is a mammalian lncRNA located in the X chromosome inactivation centre. Its gene product is first transcribed from the inactive X chromosome, and then, it spreads along the same X chromosome from which it was transcribed. In mammals, silencing of one of the two X chromosomes is necessary to achieve dosage compensation. The lncRNA XIST triggers X chromosome inactivation (XCI) in cells of the early embryo and in hematopoietic progenitors where silencing factors are present. XIST is not however required for the maintenance of XCI. XIST is also found to be expressed in adult females, and for this reason, it is suggested that the loss of XIST in adults could lead to the reactivation of inactive X genes. Having said this, the exact molecular mechanism by which XIST inactivates the X chromosomes remains unclear.

Nonetheless, surmounting evidence suggests that XIST has a role in the differentiation and proliferation of human cells. In fact, the dysregulated expression of XIST may play a pathologic role in cancer, which could be related to changes in gene expression, from the alterations to the stability of heterochromatin. It is possible that cancer cells produce silencing factors that allow for the inactivation of the X chromosome outside of the context of embryonic development. SATB1 (or special AT-rich sequence-binding protein-1), for instance, has been identified as a factor related to XIST-mediated chromosome silencing, and its aberrant expression was shown to promote breast, hepatocellular, prostate and other types of cancer.[19] XIST silencing has also been reported in transgenic male fibrosarcoma cell lines, again suggesting a special context whereby X chromosome inactivation through XIST can occur in cancer cells.[20]

CONCLUSION

ANRIL, HOTAIR and XIST are merely three of the ncRNAs that are currently being investigated. To mention but a few, others include Dleu2, EGO, lncRNA-a7, lncRNA-P21, and MEG3, each with an equal potential for being the missing piece of the puzzle. It is not therefore impossible to envisage a therapeutic world based on ncRNAs. Presently, however, the main challenge in introducing ncRNA-based therapeutics into clinical practice is the delivery and the off-target effects. Breakthroughs in both of these areas will pave the way forward for the future of medicine.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
2. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478(7370):476-82.
3. Hung T, Chang HY. Long noncoding RNA in genome regulation Prospects and mechanisms. Rna Biol. 2010;7(5):582-5.
4. Yan DS, He DD, He SM, Chen XY, Fan Z, Chen RS. Identification and Analysis of Intermediate Size Noncoding RNAs in the Human Fetal Brain. Plos One. 2011;6(7).
5. Kim WY, Sharpless NE. The regulation of INK4/ARF in cancer and aging. Cell. 2006;127(2):265-75.
6. Yap KL, Li SD, Munoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, et al. Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. Mol Cell. 2010;38(5):662-74.
7. Pasmant E, Sabbagh A, Vidaud M, Bieche I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. Faseb J. 2011;25(2):444-8.
8. Harismendy O, Notani D, Song XY, Rahim NG, Tanasa B, Heintzman N, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature. 2011;470(7333):264-+.
9. Yang F, Zhang L, Huo XS, Yuan JH, Xu D, Yuan SX, et al. Long Noncoding RNA High Expression in Hepatocellular Carcinoma Facilitates Tumor Growth Through Enhancer of Zeste Homolog 2 in Humans. Hepatology. 2011;54(5):1679-89.
10. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by Noncoding RNAs. Cell. 2007;129(7):1311-23.
11. Wang XQ, Crutchley JL, Dostie J. Shaping the Genome with Non-Coding RNAs. Current genomics. 2011;12(5):307-21.
12. He S, Liu SP, Zhu H. The sequence, structure and evolutionary features of HOTAIR in mammals. Bmc Evol Biol. 2011;11.
13. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464(7291):1071-6.
14. Kogo R, Shimamura T, Mimori K, Kawahara K, Imoto S, Sudo T, et al. Long Noncoding RNA HOTAIR Regulates Polycomb-Dependent Chromatin Modification and Is Associated with Poor Prognosis in Colorectal Cancers. Cancer Res. 2011;71(20):6320-6.
15. Niinuma T, Suzuki H, Nojima M, Nosho K, Yamamoto H, Takamaru H, et al. Upregulation of miR-196a and HOTAIR Drive Malignant Character in Gastrointestinal Stromal Tumors. Cancer Res. 2012;72(5):1126-36.
16. Geng YJ, Xie SL, Li Q, Ma J, Wang GY. Large Intervening Non-coding RNA HOTAIR is Associated with Hepatocellular Carcinoma Progression. J Int Med Res. 2011;39(6):2119-28.
17. Yang Z, Zhou L, Wu LM, Lai MC, Xie HY, Zhang F, et al. Overexpression of Long Non-coding RNA HOTAIR Predicts Tumor Recurrence in Hepatocellular Carcinoma Patients Following Liver Transplantation. Ann Surg Oncol. 2011;18(5):1243-50.
18. Kim K, Jutooru I, Chadalapaka G, Johnson G, Frank J, Burghardt R, et al. HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. Oncogene. 2013;32(13):1616-25.
19. Mir R, Pradhan SJ, Galande S. Chromatin organizer SATB1 as a novel molecular target for cancer therapy. Current drug targets. 2012;13(13):1603-15.
20. Hall LL, Byron M, Sakai K, Carrel L, Willard HF, Lawrence JB. An ectopic human XIST gene can induce chromosome inactivation in postdifferentiation human HT-1080 cells. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(13):8677-82.