HAR1: an insight into IncRNA genetic evolution

How to cite:
Waters, Ella; Pucci, Perla; Hirst, Mark; Chapman, Simon; Wang, Yuzhuo; Crea, Francesco and Heath, Christopher (2021). HAR1: an insight into IncRNA genetic evolution. Epigenomics, 13(22) pp. 1831–1843.

For guidance on citations see FAQs.

© 2021 The Authors.

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Version of Record

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.2217/epi-2021-0069

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.

oro.open.ac.uk
HAR1: an insight into IncRNA genetic evolution

Ella Waters*†,1, Perla Pucci†,1,2, Mark Hirst†,1, Simon Chapman†,1, Yuzhuo Wang†,3, Francesco Crea†,1 & Christopher J Heath†,1

1School of Life, Health & Chemical Sciences, The Open University, Milton Keynes, MK7 6AA, UK
2Division of Cellular & Molecular Pathology, Department of Pathology, University of Cambridge, Cambridge, CB2 0QQ, UK
3Department of Urologic Sciences, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

*Author for correspondence: ella.waters@open.ac.uk
†Authors contributed equally

Long noncoding RNAs (lncRNAs) have a wide range of functions in health and disease, but many remain uncharacterized because of their complex expression patterns and structures. The genetic loci encoding lncRNAs can be subject to accelerated evolutionary changes within the human lineage. HAR1 is a region that has a significantly altered sequence compared to other primates and is a component of two overlapping lncRNA loci, HAR1A and HAR1B. Although the functions of these lncRNAs are unknown, they have been associated with neurological disorders and cancer. Here, we explore the current state of understanding of evolution in human lncRNA genes, using the HAR1 locus as the case study.

First draft submitted: 22 February 2021; Accepted for publication: 1 October 2021; Published online: 22 October 2021

Keywords: DNA conservation • genetic evolution • HAR1 • lncRNAs • molecular structure

Introduction to long noncoding RNAs

Genomics has advanced considerably within the last decade. Substantial research has shown that approximately 99% of the transcriptome consists of nonprotein-coding regions [1,2]. An abundant class of these noncoding RNAs (ncRNAs) are long noncoding RNAs (lncRNAs) [3]. LncRNA transcripts are defined over 200 nucleotides in length and usually lack significant open reading frames but can be capped at the 5’ end, spliced and polyadenylated [4]. The DNA sequences from which ncRNAs, including lncRNAs, are transcribed were originally believed to be ‘junk’ DNA [5], but lncRNAs have a wide range of biological functions and have been found in some cases to be differentially expressed between healthy and diseased cells [6–9].

LncRNA research is rapidly producing novel developments and knowledge with a relevant impact in molecular biology and medicine. To simplify this new knowledge and aid the understanding of lncRNA functions and characterization, lncRNAs are classified depending on their genomic location, subcellular location, structure, and function.

LncRNAs can be broadly split into five categories depending on their genomic location, relative to immediately neighboring protein-coding genes: sense, antisense, bidirectional, intronic and intergenic [2,10] (Figure 1). Sense lncRNAs are transcript variants of a known protein-coding gene that overlap with their sense sequence, whereas antisense lncRNAs are transcribed from, and overlap with, the antisense strand of a protein-coding gene. Bidirectional lncRNA transcripts are head-to-head with a protein-coding gene on the opposite strand of DNA, and intronic lncRNAs overlap within the introns of a protein-coding gene. Intergenic lncRNAs are located between protein-coding genes.

The identification of alterations in lncRNA expression has prompted the characterization of their roles in important biological processes. A peculiarity of human lncRNAs is that they are often cell-, tissue- and disease-specific [11]. Because of their specificity within the cell, their functions can be indicated by their subcellular location. Nuclear lncRNA functions include transcriptional regulation such as enhancer lncRNAs, epigenetic and chromatin structure regulation [12] and alternative splicing [13]. Cytoplasmic lncRNAs function as miRNA sponges, to be involved in modifying mRNA stability [14], to promote proteasomal degradation [15], and cytoskeletal...
arrangement [16,17]. In addition to these examples of functions, some lncRNAs have also recently been found to encode for biologically active, short peptides called micropeptides [18,19], despite their noncoding name.

LncRNAs can also be characterized depending on their structure and function, interacting with other types of RNA, DNA and proteins:

- Signals regulate transcription in response to a stimulus;
- Decoys present binding sites to regulate the availability of regulatory factors, such as transcription factors and miRNAs;
- Scaffolds allow the assembly of multiple-component complexes including a RNP by providing structural domains, which can result in transcriptional activation or repression;
- Guides direct RNPs to their target genes;
- Enhancers are regions which can interact with promoters by recruiting RNA polymerase, thereby influencing the expression of transcription factors tethered to this region by other lncRNAs [20].

Depending on the site of these functions, lncRNAs can be described as acting in cis or trans [21,22]. Cis acting lncRNAs function at their loci and often regulate the expression of nearby genes. Trans acting lncRNAs function away from their site of transcription, such as regulating the expression of genes with loci on different chromosomes.

Despite the increasing volume of evidence showing that lncRNAs are involved in the regulation of gene expression and are implicated in numerous different human diseases, to date, only a relatively small percentage of lncRNA structures and functions have been elucidated [23,24]. This is likely due to the multifaceted expression patterns of lncRNAs, their complex molecular structures and relatively lower cellular abundance in physiological conditions [25], although there are specific patterns of lncRNA upregulation in cancer and other diseases.

Understanding the evolutionary constraints of lncRNAs and their degree of sequence conservation in humans may provide insights into their molecular interactions and functional roles, especially for those lncRNAs with genetic regions that have undergone accelerated evolution in the human lineage [26]. This review explores some of the current understanding of lncRNA evolution, using examples of key lncRNAs within the HAR1 locus to further analyze the relationship between their evolution and biological functions in health and disease.
**LncRNA origin**

LncRNA loci are often evolutionarily poorly conserved in their sequence among closely related organisms that share a common ancestor with humans. This phenomenon suggests a high gene turnover rate within the human lineage. This sequence turnover can refer to the LncRNA sequences diverging through mutation but retaining functionality, or to the process of functional LncRNAs originating from nonfunctional LncRNAs [27].

Novel LncRNAs can occur via de novo formation [28], although emerging evidence shows that LncRNAs may also be formed via LncRNA duplication or via the genetic rearrangements of protein-coding gene [29–31]. De novo evolution of LncRNAs can occur in regions not previously transcriptionally active, that become transcribed, such as transposable element activity. Transposable elements could influence the tissue specificity of LncRNAs [32]. The authors of the study also showed that a promoter region, which was originated from the transposable element family L1PA2, could act as a promoter for LncRNAs that are specifically expressed in the placenta [32]. This suggested that functional LncRNAs can originate from a nonfunctional transposable element, with functional LncRNAs contributing to a selectable trait persisting.

The phenomenon of LncRNAs formation via gene duplication of other LncRNAs is relatively understudied in comparison to protein-coding genes, although the duplication of a gene is the main process by which new protein-coding genes are created [33]. LncRNAs can form multiple LncRNA genes in this way, for example, a family of long intergenic ncRNAs (lincRNAs) originated from another lincRNA sequence FAM230C by segmental duplication [29]. LncRNAs can also emerge by assembly from the material of protein-coding sequences that are in the process of pseudogenization [34]. A relatively small number of LncRNAs originate from protein-coding sequences in this way [31]. The authors of the study also reported that despite losing their protein-coding abilities, these functional LncRNAs had elements that were conserved for millions of years.

**LncRNA evolution & conservation**

Conserved regions are sequences that remain unchanged or very similar in either different species or the same species, throughout generations in evolution. In LncRNAs, conserved sequences can vary in length, and have been shown in several algorithms to be of functional importance [35]. In contrast to protein-coding genes, the primary structures of multiple LncRNAs have poor evolutionary sequence conservation overall [36–38].

Large evolutionary studies of LncRNAs are instrumental in developing our understanding of regulatory gene networks, and therefore determining LncRNA functions [36]. This RNA-sequencing study of eight organs in 11 different species uncovered key information, such as the strong conservation of LncRNA promoter sequences for younger and older LncRNAs, comparable to protein-coding gene promoters [36]. This suggests stronger selective constraints at the transcriptional level. Furthermore, LncRNAs older than 90 million years (Myr) have a higher exonic sequence conservation, similar to that of protein-coding exons. Older LncRNAs may have more detectable homologous sequences within evolution, whereas the newer sequences may have evolved more rapidly, therefore evading detection with the current algorithms. The authors also confirmed that LncRNA transcription has diverged more rapidly over time compared with protein-coding genes, but tissue specificity is maintained.

A key advance in this area was made by the development of a software package called slncky, which identifies LncRNAs from RNA-sequencing data, and so improves the accuracy of LncRNA analysis [37]. Previous approaches have had difficulty in classifying conserved LncRNAs, as these were often mistaken for protein-coding genes or pseudogenes [37]. Owing to its increased sensitivity, slncky enables a more specific classification of LncRNAs and facilitates the study of LncRNA evolution, conservation and potential functions. The software identifies a syntenic region for a LncRNA, a region with the gene locus on the same chromosomal location in different species, the sequence and transcript conservation are then characterized. Slncky can also identify evolutionary properties, as shown by the software discovering specific selection on two distinct classes of ancestral intergenic LncRNAs. 232 LncRNAs were analyzed, finding that about 20% of these only have constraint on the act of transcription, with limited conservation on the actual transcript sequence. Approximately 80% of the LncRNAs analyzed have a strong purifying selection, therefore preserving and stabilizing the transcript in evolution. The latter 80% of LncRNAs could be considered as being part of a different class of lincRNAs. Further developing and utilizing software such as slncky to reveal more information on LncRNAs by evolutionary analysis, and further characterize LncRNAs, will assist in clarifying more of their functional roles.

More recently, comparative transcriptomics has been used to analyze LncRNAs at different developmental stages of mouse, rat and chicken, using tissues from the brain, testes, liver and kidney [38]. The study reported that the LncRNAs that are expressed during embryonic development, particularly in the brain and kidney, had higher levels...
of functional constraint. As the most conserved sequences were located in the promoters of IncRNAs, rather than their exons, these IncRNAs were suggested to have RNA sequence-independent biological functions. This means the process of transcription influences gene function, rather than the IncRNA product being functional itself. Many of IncRNAs that act with this kind of crosstalk are cis regulators [39]. This type of IncRNA functionality along with their specific expression patterns are significant findings and are a potential focus for future studies of IncRNAs.

With several previous studies being focussed on protein-coding gene or shorter RNA characterization, these studies represent some of the IncRNA-specific research over the past decade. Although these studies indicate that it may be difficult to predict the functions of IncRNAs based on their sequence similarities alone, as IncRNAs have species-specific expression patterns, their syntenic conservation and expression conservation within regulatory regions are the main types of conservation aiding the understanding of IncRNA functional regions. These studies and those alike are essential for characterizing understudied IncRNAs and identifying regions of interest and likely functional regions within their sequences.

**Rapid evolution of noncoding regions**

While increasing the understanding of IncRNA conservation and evolution, it has also been discovered that there are multiple regions within the human genome that have undergone rapid sequence and structural changes within their recent evolution. As previously mentioned, in contrast to protein-coding genes, the primary structures of multiple IncRNAs have poor evolutionary sequence conservation, although they may have conserved genomic locations [40]. As IncRNAs are transcripts with a long sequence length, they can fold into complex secondary and tertiary structures. This folding can increase their structural stability and conservation [41], which may explain why IncRNAs with rapidly evolving sequences still have a functional role, although few IncRNA structures have been mapped and characterized yet. Understanding the high-order structure of a IncRNA may also enable the identification of binding sites and specific motifs that are important for predicting and understanding function, more effectively than considering the primary structure in isolation [42].

Rapid structural changes and structural versatility are likely the largest contributors to the wide range of human specific IncRNA functions that have evolved, in addition to their evolutionary origin. Examples of IncRNA structural motifs influencing their functions include conserved pseudoknots, helical secondary structures connected by loops, forming in the IncRNA MEG3 tertiary structure allowing p53 interaction [43]; G-quadruplex secondary structure, rich in guanine, in the IncRNA GSEC being important for colon cancer cell migration [44] and conserved protein binding elements and domains essential for the mechanism of action of the IncRNA HOTAIR predicted from its secondary structure [45]. These few examples demonstrate the importance of understanding the structure of a IncRNA.

**Case study: the human accelerated region HAR1**

Human accelerated regions (HARs) are short (~260 base pairs) noncoding genomic regions that have undergone an accelerated rate of nucleotide substitutions and deletions, specifically in the human lineage, compared to their chimpanzee orthologous sequences [46]. 49 HARs with significantly high substitution rates were originally identified [46]. Of these, the most accelerated evolution was observed at the HAR1 locus, which has undergone 18 substitutions within its 118 base pairs in the human genome, compared to the expected 0.27 substitutions, as the divergence of the Homo and Pan gena [47]. This noncoding HAR1 sequence resides in a shared region within two divergently transcribed IncRNAs named after HAR1, HAR1A and HAR1B, in chromosome region 20q13.33 (Figure 2).

The high-order structure of the short HAR1 region has been studied in detail. The secondary structure of HAR1 was modeled to be a stable cloverleaf [48,49]. In these studies, the human HAR1 structure was compared to the orthologous HAR1 sequence in chimpanzees; the latter was predicted to fold into an unstable and extended hairpin. The structural stability of the HAR1 helix in humans likely arises due to the length of the region, the transition from weaker AT to stronger GC base-pairing occurring through its 18 substitutions, increasing the base stacking energy within the helix. A recent study has further analyzed the secondary structure of human HAR1, identifying a bulge within the structure as a potential binding site for molecular interaction [50].

There is a close similarity between human and chimpanzee genomes because of the relatively short amount of time available for mutations to accumulate; an approximate 5–7 million-year history of separate evolution. The HAR1 region is more ancient, dating back at least 310 Myr to a shared ancestor with chicken [47]. In contrast to the 18 altered bases between human and chimpanzee HAR1 genes, the chimpanzee and chicken genes differ by
Figure 2. Simplified schematic of the HAR1 loci. The long noncoding RNA HAR1A is on the sense strand, and the long noncoding RNA HAR1B is on the antisense strand on chromosome 20. These divergently transcribed long noncoding RNAs have a region of overlap, containing the ‘accelerated’ 118bp HAR1. There is also a miscellaneous RNA overlapping with HAR1A.

<table>
<thead>
<tr>
<th>Mutation position</th>
<th>Substitutions Pan → Homo</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>U → C</td>
</tr>
<tr>
<td>15</td>
<td>A → C</td>
</tr>
<tr>
<td>16</td>
<td>A → G</td>
</tr>
<tr>
<td>26</td>
<td>U → C</td>
</tr>
<tr>
<td>27</td>
<td>U → G</td>
</tr>
<tr>
<td>29</td>
<td>A → G</td>
</tr>
<tr>
<td>33</td>
<td>A → G</td>
</tr>
<tr>
<td>41</td>
<td>U → G</td>
</tr>
<tr>
<td>44</td>
<td>A → G</td>
</tr>
<tr>
<td>47</td>
<td>U → C</td>
</tr>
<tr>
<td>54</td>
<td>A → G</td>
</tr>
<tr>
<td>57</td>
<td>U → C</td>
</tr>
<tr>
<td>64</td>
<td>A → G</td>
</tr>
<tr>
<td>66</td>
<td>U → C</td>
</tr>
<tr>
<td>73</td>
<td>A → G</td>
</tr>
<tr>
<td>88</td>
<td>U → G</td>
</tr>
<tr>
<td>94</td>
<td>A → G</td>
</tr>
<tr>
<td>113</td>
<td>A → G</td>
</tr>
</tbody>
</table>

Figure 3. The human HAR1 secondary structure, including the mutations (numbers in blue) that have rapidly occurred between chimpanzee and human.

Figure created in BioRender.com.

only two bases out of its 118 (Figure 3). Furthermore, it has been suggested that the human-specific substitutions in HAR1 may have occurred in the human lineage within the last 1 Myr [47].

A computational model was developed to understand how the human HAR1 secondary structure might have evolved, comparing the chimpanzee (ancestral), the Denisovan and the modern human HAR1 secondary structures [51]. The computational model reconstructed the statistically most likely order of the 18 substitutions. The likely final mutation in HAR1 human evolution thus far is the variant resulting in the stabilization of the lowest stem. This stem has been recreated in the modern human, as it was weakened in the evolution of ancestral HAR1 to Denisovan [51]. These analyses, along with the predicted stability of the HAR1 helix structure having rapidly increased in humans, suggest that the structure of HAR1 is highly relevant to its function.

Although the secondary structure and nucleotide changes within the 118 base pair region HAR1 region has been mapped, little is known about the overall HAR1A and HAR1B structures, and any other functionally relevant sequences within these lncRNAs. As previously explained, identifying other short sequences, mapping
their structures and validating in the laboratory is important for understanding how a lncRNA is functional, as it has been recently shown that short syntenic sequences within lncRNAs are the functional elements, rather than the whole lncRNA sequence. A tool which may be relevant for HAR1A and HAR1B is the lncLOOM algorithm [35], which is designed to identify biologically relevant short sequences, 6–12 nucleotides long, that are deeply conserved within lncRNAs and other elements. These short sequences were compared across several species in lncLOOM, including humans, mice, opossums, chickens and zebrafish, with 18 different species compared for their analysis of the Cyanana lncRNA. These short sequences were identified in syntenic regions as having been conserved over large evolutionary distances because the order of nucleotides was identical across different species. This synteny may be important for function; therefore, these highly conserved motifs are likely the functional elements within the lncRNA. Many of these motifs were found to be regions of the lncRNA that contribute to the binding sites for RNA binding proteins and smaller RNAs [35].

Another study has found two lncRNAs, APOLO and UPAT, one in plants and the other in humans, respectively, have very different sequences but are involved in the same pathways and protein interactions [52]. As these lncRNAs are seemingly evolutionary unrelated in terms of sequence, it would be interesting to identify whether they share common short motifs or if their 3D structures are similar. This study demonstrates how lncRNA evolution is seemingly unique, and further study is required to increase the understanding of lncRNA functionality. As lncLOOM, or similar tools, can be used to identify functionally relevant short sequences within lncRNAs, it would be interesting to identify whether the HAR1 sequence includes one of these elements, with this being highly evolutionary accelerated, or if there are functionally conserved elements elsewhere in HAR1A and HAR1B that are more relevant to their functions in health and disease.

Identifying the tertiary interactions indicated by the secondary structure will help in establishing the mechanism of action of the lncRNAs overlapping with this accelerated region: HAR1A and HAR1B. Furthermore, it is necessary to identify whether short syntenic elements within the HAR1A and HAR1B sequences are responsible for a function or can be linked to a mechanism of action. These studies can, in turn, further identify the possible roles of HAR1A and HAR1B in health and disease.

HAR1 lncRNA loci in human health & disease

As explored in this review, the extent of a lncRNA’s evolutionary conservation is determined by several different factors, including the mechanism of its origin, its age, genomic environment and the structure and sequences of the lncRNA itself. Many of these factors, especially the selection of functionally conserved regions through long evolutionary distances, and the folding of the lncRNA structures, may contribute to these genes have functional roles in multiple diseases. To corroborate these models, we discuss recent evidence, showing the highly evolved HAR1 loci may have a functional role in neurodevelopment, neurological disorders and cancer.

HAR1 loci in neurodevelopment

HAR1A expression was identified specifically in Cajal-Retzius cells in the developing human neocortex and the dorsal telencephalon, between the 7th and 9th gestational weeks, and in the subpial granular layer up to 19 gestational weeks [47]. HAR1B expression was shown to be more widespread in the brain. The main function of Cajal-Retzius cells is the organization of neurons during neocortical development [53]. Cortical lamination is the organization of cells into six excitatory layers (I–VI). This cell lamination is achieved by the release of signals, such as the extracellular glycoprotein RELN [54]. As the expression of HAR1A overlaps with the expression of RELN in Cajal-Retzius cells between the critical development period of 17th–19th gestational weeks, it is necessary to understand their possible functional relationship, and to discern the role of HAR1A and HAR1B in nervous system development.

HAR1 loci in neurological & psychiatric disorders

Owing to the preferential expression of HAR1 in the human cortex [47], it is not surprising that emerging evidence has linked this locus to neurological and psychiatric disorders. A study has shown that the HAR1 locus may have a functional role in Huntington’s disease [55]. Huntington’s disease is an autosomal-dominant neurological disorder, with typical symptoms including involuntary movements, depression and dementia [56]. It is caused by a CAG trinucleotide expansion in the HTT gene [57], that results in the relocation of the transcriptional repressor, REST. The REST gene represses thousands of genes, many which are neuron specific [58].
The **HAR1** locus was shown to be transcriptionally regulated by REST protein in Huntington’s disease in cell lines and human tissue [55]. Firstly, the authors identified three REST binding sites in the vicinity of **HAR1**. Two of these contain canonical RE1 binding motifs that are specifically found in promoter regions [59]. Consistent with a transcriptional silencing effect of REST, quantitative PCR analysis demonstrated increased levels of **HAR1A** and **HAR1B** mRNAs when REST was silenced in vitro [55]. The authors also analyzed human brain tissue, detecting a significantly lower level of **HAR1A** and **HAR1B** transcripts in the striatum of Huntington’s disease patient tissue, compared to normal brain. The striatum has a significant role in motor control, and it is the main region causing neurodegeneration in Huntington’s disease [60]. The normal mechanisms of action of the **HAR1A** and **HAR1B** IncRNAs, likely regulated by REST, are poorly understood, but their dysregulation in Huntington’s disease striatum suggests they may contribute to abnormal cellular function and phenotype.

**HAR1A** and **HAR1B** have also been associated with Alzheimer’s disease. A machine-learning based model called Laplacian Regularized Least Squares for LncRNA–Disease Association (LRLSLDA) predicts novel IncRNA–disease associations, where the IncRNA may have a functional role [61]. LRLSLDA associated **HAR1A** and **HAR1B** to Alzheimer’s disease [62]. This model is based on the assumption that similar diseases have associations with functionally similar IncRNAs. These parameters may be a limitation of the model, and in vitro confirmation of this prediction is required.

The role of **HAR1A** was investigated in schizophrenia, however, a significant association was not identified within the 285-patient sample [63]. Nevertheless, the authors did find that a CCCCCG haplotype combined of six single nucleotide polymorphisms covered the **HARIA** region completely and was significantly associated with auditory hallucinations in 221 psychiatric regions [63]. As hallucinations can be a result of erroneous neuronal connections forming in brain development, the activity of **HAR1A** may be altered, increasing the chance of psychosis, but this requires further validation.

These studies indicate that **HAR1A** and **HAR1B** may have a role in neurological and psychiatric disorders, but more evidence is required. However, there is increasing recognition for the important role that IncRNAs have in Huntington’s disease and other neurological disorders [64–66].

**HAR1** loci in cancer

Similar to neurological and psychiatric disorders, the **HAR1** locus has been implicated in multiple cancers, including gliomas. Bioinformatics analysis showed significant downregulation of **HAR1A** in diffuse glioma samples [67]. The authors found that lower **HAR1A** expression resulted in significantly worse survival of diffuse glioma patients. Upregulation of **HAR1A** increased the survival rates of patients who underwent radiotherapy and chemotherapy, demonstrating a tumor suppressor role. Another study showed that **HARIA** downregulation also resulted in worse survival of diffuse glioma patients, specifically with the prognostic marker IDH mutant [68]. These results suggest that the downregulation of **HAR1A** may be a prognostic biomarker for diffuse gliomas, though the mechanisms of action and cellular function of this IncRNA are still unknown.

In another study using microarray analysis, **HAR1B** was one of nine IncRNAs in a signature pattern of gene expression, which were found to be significantly associated with glioblastoma survival [69]. The other eight IncRNAs are **AC078883.3**, **AC104653.1**, **RP11-944L7.4**, **RP4-635E18.7**, **RP5-1172N10.2**, **TP73-AS1**, **SAPCD1-AS1**, and **HOTAIR**. These IncRNAs together could be a potential biomarker for diagnosis and prediction of survival. Further research into the IncRNAs biological functions and mechanism of action is also required.

In addition to gliomas, **HAR1A** and **HAR1B** are dysregulated in other types of cancer. A recent study identified **HAR1B** as a predictive biomarker in bone and soft-tissue sarcoma cell lines [70]. **HAR1B** has differential expression in these cell lines, and siRNA silencing of **HAR1B** resulted in an increased resistance to a therapeutic inhibitor, pazopanib. **HAR1B** was upregulated in clear cell renal cell carcinoma within a signature containing **MIR155HG**, **PVT1**, and **TCL6**, and was correlated with poor overall survival [71]. **HAR1B** may also have a role in human parathyroid tumors, as silencing of the tumor suppressor MEN1 upregulated **HAR1B** [72]. The study reported that **HAR1B** silencing increased **SOX2** and **NANOG** levels in primary cell lines. Overexpression of these genes can result in cancer hallmarks such as metastasis. The extent of the role of **HAR1B** in this cancer requires further investigation. A study in breast cancer showed that the upregulation of **HARIA** within a signature containing eight other IncRNAs (**LINC00310**, **LINC00323**, **LINC00574**, **LINC00704**, **LINC00705**, **ARRDC1-AS1**, **FAM74A3** and **UMODL1-AS1**) predicted cancer reoccurrence [73]. For six of these IncRNAs together, including **HAR1A**, the alteration was significantly higher within the advanced breast cancer group. **HAR1A** has also been associated with recurrence free survival in papillary thyroid cancer [74]. A different study found that low expression of **HAR1A** and
HAR1A in oral cancer cell lines, demonstrating HAR1A as a tumor suppressor. When HAR1A is upregulated, ALPK1 levels reduce. This results in increased levels of BRD7 and myosin IIA, resulting in tumor suppression. Decreased levels of TNF-α and CCL2 with HAR1A upregulation suggest HAR1A has an anti-inflammatory role.

HAR1B in patients with hepatocellular carcinoma, compared to liver cirrhosis and chronic hepatitis B cases, was significantly associated with poor prognosis, advanced histological grade and progressive TNM stage [75].

Another recent publication has analyzed the downstream mechanism of action and functional role of HAR1A in oral cancer progression [76]. HAR1A was found to bind to the oncogene ALPK1 in the nucleus of an oral squamous cell carcinoma cell line, SAS, with HAR1A and ALPK1 having an inverse correlation. The discovery of HAR1A being primarily localized in the nucleus is important because of IncRNAs functional specificity depending on their subcellular localization. HAR1A knockdown increased the protein levels of pro-inflammatory cytokines TNF-α and CCL2, with ALPK1 knockdown decreasing these levels protein levels. Therefore, it is proposed that HAR1A may suppress a ALPK1/BRD7/myosin IIA pathway (Figure 4), as HAR1A knockdown may cause the ALPK1 protein to translocate into the nucleus, bind and downregulate BRD7, resulting in inflammation and oral cancer progression [76]. Further to a possibly anti-inflammatory role, the function of HAR1A was also analyzed in oral cancer cell lines. HAR1A silencing increases oral cancer cell proliferation, migration and promote apoptosis, further implicating HAR1A as a tumor suppressor. HAR1A silencing may also increase metastasis, investigated using epithelial-mesenchymal transition (EMT) markers. Mesenchymal markers N-cadherin, fibronectin, vimentin and slug were upregulated with HAR1A silencing, and this did not upregulate the epithelial marker E-cadherin. These results further implicate HAR1A as having a tumor suppressor role [76]. Although this study is one of the first identifying HAR1A downstream interacting partners, IncRNAs are often cell and disease specific, therefore further investigation into this mechanism of action and other interacting partners are needed in different cancers and in vivo studies.

These cancer studies show that HAR1A and HAR1B expression can be oncogenic in some cancers [71,73] and tumor suppressors in other cancers [67,75,76], and may be used as a prognostic biomarker to indicate the stage of tumor progression. They also suggest a functional role of these IncRNAs in human cancers, especially gliomas and oral carcinomas, but this requires further research in different types of cancer. As with neurological disorders, there is also an increasing focus for identifying the functional role of IncRNAs in cancer [77–79].

**Conclusion**

HAR1A and HAR1B are IncRNAs at the HAR1 loci which have been implicated in health and disease. As explained within this review, the functions of HAR1A and HAR1B in disease and the upstream mechanisms of action regulating these IncRNAs is currently unknown. An investigation into REST regulating HAR1A and HAR1B in diseases other than Huntington’s disease is required. Furthering our knowledge of evolutionary accelerated regions, through investigating their sequences and structures, will enhance our understanding of HAR1 loci functions in human disease. However, functionally conserved regions elsewhere in the HAR1A and HAR1B sequences may also contribute to their IncRNA function.

**Future perspective**

IncRNA regions comprise a significant part of our genome, owing to the abundance of studies identifying novel IncRNAs over the last decade. The more that IncRNA structures and sequences are analyzed, the more is shown that they have a wide range of functions and expression patterns within a cell, tissue, disease or species. Owing to their specificity, IncRNAs are recently emerging as promising biomarkers and therapeutic targets [80].

As IncRNAs typically have low sequence conservation and no protein-coding ability, their functionality was previously questioned. Studies are now emerging to further understand IncRNAs roles in health and disease. Many IncRNAs have been found to contain regions of high sequence conservation which may be responsible for their
function and interactions with other molecules [81]. Other lncRNA regions, such as HAR1A and HAR1B in the
HAR1 loci, have undergone rapid evolution, specifically within the human lineage. This high rate of sequence
turnover is possibly the reason the structure and function of these lncRNAs are likely unique in humans. The
in silico development of lncRNA-specific algorithms has enabled more data to be gathered regarding the degree of
lncRNA sequence conservation and their structures [35–38].

LncRNAs that are expressed in the developing brain and kidney, during embryonic development, had higher
levels of functional constraint and therefore may function in cis from the process of transcription rather than
from the lncRNA itself [38]. This is a significant finding and a potential focus for future studies of lncRNAs. For
example, if HAR1A, which was previously shown to be expressed early in brain development [47], conforms to these
findings, HAR1A may have evolved to have higher levels of functional constraints within humans, and therefore
may have a possible active function within neocortical development. Using algorithms such as IncLOOM may
allow evolutionary conserved motifs to be identified within the lncRNA sequence that are responsible for the
function [35], and tools may also identify other regions that have undergone rapid sequence changes, like HAR1.

As discussed in this review, the lncRNAs HAR1A and HAR1B have been shown to have significantly altered
expression and association in various diseases including neurodevelopment, Huntington’s disease [55], and cancers
including gliomas [67] and oral cancer [76], although their functional roles and upstream mechanisms of action are
unknown, so their novelty presents a wide scope for study. As HAR1A is co-expressed with RELN, and RELN is
responsible for the lamination of cortical neurons [54], HAR1A may have a functional role in migration and
possibly regulates RELN. It has previously been demonstrated that cortical layers III, V and VI of Huntington’s
disease patients experience loss of neurons, especially as the disease progresses [82]. Downregulation of HAR1A in
Huntington’s disease, transcriptionally regulated by REST, may contribute to the architecture of layering neurons
and this loss of cells. Further to this, tumor cell invasion was identified in deeper layers (IV–VI) of the cerebral cortex
in malignant gliomas [83]. HAR1A was identified to be downregulated in more aggressive gliomas [67], meaning HAR1A
may have a role in cancer cell invasion and migration here. The extent of their regulatory relationship
with REST, likely located upstream of HAR1A and HAR1B, will be interesting to further establish. This potential
migratory role requires laboratory characterization in vitro and in vivo relevant to these diseases, using techniques
such as transwell migration assays, and Boyden chamber assays.

HAR1A was identified to also have a tumor suppressor role in oral cancer [76]. This publication demonstrated that
HAR1A and the protein kinase ALPK1 may bind in monocytes, and HAR1A, possibly indirectly, inhibits this gene
in oral cancer cell lines. HAR1A knockdown resulted in increased cell viability, increased migration, and reduced
cell apoptosis [76]. HAR1A silencing resulted in oral cancer progression and inflammation. The domains of HAR1A
binding ALPK1 have not yet being identified. It would be interesting to understand whether the HAR1 evolutionary
accelerated region is included in this, or whether there are unknown functional motifs. Another key result was the
identification of HAR1A in the nucleus [76]. In oral cancer, HAR1A may function in cis to epigenetically regulate ALPK1, possibly resulting in inflammation-promoted tumorigenesis. These findings suggest that HAR1A
may have a number of different biological roles within health and disease including migration, inflammation and
tumorigenesis, dependent on the cell and tissue type.

As HAR1B is less characterized in comparison to HAR1A, HAR1B requires further laboratory characterization to
analyze the subcellular location in the appropriate cell line, identify a phenotypic effect and lncRNA function
through gain-of-function and loss-of-function experiments comparing cells with/without lncRNA expression, and
find possible binding partners and interacting-macromolecules to discover a mechanism of action. Direct binding
partners can be investigated using techniques such as RNA pull-down assays and mass spectrometry [84], followed
by luciferase reporter assays and chromatin precipitation [85]. HAR1A also requires further characterization in other
cancers, such as gliomas, and neurodegenerative disorders. The secondary structures of these lncRNAs also require
mapping. As well as in vitro and in vivo characterization of these, understanding the tertiary structure of HAR1A
will aid in locating binding sites and other motifs of functional interest. The RNA structure of the lncRNA Xist
was found to vary, likely depending on the experiment [86]. Structural and wet lab characterization of IncRNAs
should be a consistent set of experiments to reduce variability between laboratories.

As shown by the HAR1 loci, further knowledge of the evolutionary history of IncRNAs, including their evolved
or evolving structures and degree of sequence conservation within the human lineage, can guide their molecular
and functional characterization, with the final aim to identify their roles in health and disease. Methods to further
understand the roles of IncRNAs such as HAR1A and HAR1B that have aberrant expression in health and disease
will provide insights into improving diagnostics and therapeutic options.
Executive summary

Introduction to long noncoding RNAs
• Long noncoding RNAs (lncRNAs) are nonprotein coding regions over 200 nucleotides in length. LncRNAs are often cell- and disease-specific, and their functions can be inferred by their subcellular location.
• LncRNAs are involved in several human diseases, but the majority of lncRNAs are still structurally and functionally uncharacterized.

LncRNA origin, evolution & conservation
• LncRNA evolutionary conservation can be determined by the mechanism of origin, age, genomic environment and structure.
• Numerous studies have indicated that many lncRNAs have poor evolutionary sequence conservation; therefore, functionality may be conserved via specific sequence motifs or their folded structures.

Rapid evolution of lncRNAs
• The genetic loci encoding lncRNAs can be subject to an accelerated rate of nucleotide substitutions within the human lineage.

Case study: HAR1
• An example of a region that has undergone accelerated evolution is the HAR1. The HAR1 sequence resides within two overlapping but divergently transcribed lncRNAs, HAR1A and HAR1B.
• The HAR1 secondary structure has been mapped in Homo and Pan. This structure stabilizes in humans, suggesting the structure is relevant for HAR1 function.

HAR1 lncRNA loci in human health & disease
• HAR1A is expressed in the brain during neurodevelopment. The function of HAR1A in neurodevelopment is currently unknown.
• The HAR1 locus is involved in multiple neurological disorders. In Huntington’s disease, HAR1A and HAR1B are transcriptionally regulated by REST.
• The HAR1 locus is also implicated in multiple cancers, including tumor suppressor roles in gliomas and oral cancer. HAR1A is present in the nucleus. HAR1A was also shown to have an anti-inflammatory role.

Future perspective
• Further studies are required to identify the genes regulating the HAR1 loci in cancer and neurological disorders.
• Understanding lncRNA structures and their evolutionary conservation will aid in identifying more lncRNA functions and their mechanisms of action in health and disease, and therefore improve therapeutic options.

Financial & competing interests disclosure
The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References
Papers of special note have been highlighted as: ● of interest; ●● of considerable interest


34. Liu WH, Tsai ZTY, Tsai HK. Comparative genomic analyses highlight the contribution of pseudogenized protein-coding genes to human lincRNAs. *BMC Genomics* 18 (2017).

- Discusses a new algorithm identifying IncRNA functional regions.
  • This was the first study to identify the HARI loci.
  • Investigates the structure of the HARI region.
  • Identified REST and HARI expression in Huntington’s disease as a possible mechanism of action.


- Identified low HAR1A expression results in worse patient survival for diffuse glioma patients.


- Recent paper identifying the mechanism of action and functional roles of HAR1A in oral cancer, and the location of HAR1A in the cell nucleus.


