JOURNAL OF
HUMAN
TRANSCRIPTOME

http://informahealthcare.com/jht
ISSN: 2332-4015 (electronic)

J Hum Transcriptome, 1(1): 2–9
DOI: 10.3109/23324015.2015.1077591

Taylor & Francis
Taylor & Francis Group

REVIEW

# Conservation and tissue-specific transcription patterns of long noncoding RNAs

Melanie Ward*, Callum McEwan*, James D Mills and Michael Janitz

*School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia*

**Abstract**

Over the past decade, the focus of molecular biology has shifted from being predominately DNA and protein-centric to having a greater appreciation of RNA. It is now accepted that the genome is pervasively transcribed in tissue- and cell-specific manner, to produce not only protein-coding RNAs, but also an array of noncoding RNAs (ncRNAs). Many of these ncRNAs have been found to interact with DNA, protein and other RNA molecules where they exert regulatory functions. Long ncRNAs (lncRNAs) are a subclass of ncRNAs that are particularly interesting due to their cell-specific and species-specific expression patterns and unique conservation patterns. Currently, individual lncRNAs have been classified functionally; however, for the vast majority the functional relevance is unknown. To better categorize lncRNAs, an understanding of their specific expression patterns and evolutionary constraints are needed.

**Key Words:**

lncRNAs, comparative genomics, gene regulation, transcriptome, RNA-Seq

**History**

Received 14 May 2015
Accepted 15 July 2015

## Introduction

Recent developments in RNA sequencing (RNA-Seq) technology have given scientists an in-depth view of the human transcriptome [1]. It is apparent that traditional views of RNA as merely an intermediary molecule between DNA and protein discredits the complexity of the human genome and ignores the pivotal role of noncoding RNA (ncRNA) as a regulatory molecule in essential life processes [2]. Despite merely a twofold increase in the number of protein-coding genes between the human genome and that of the common fruit fly, *Drosophila melanogaster*, these species exhibit dramatically differing levels of phenotypic complexity. To account for this disparity, there must exist a multi-level regulatory mechanism enabling such drastic diversity from a similar number of protein-coding genes.

There is a direct correlation between the proportion of ncRNAs in an organism's genome and its developmental complexity [3]. The largest subclass of ncRNAs is long noncoding RNAs (lncRNAs). These are mRNA-like transcripts arbitrarily defined as being greater than 200 nucleotides long, with no protein-coding capacity, which however undergo alternative splicing and post-transcriptional processing [4]. Initially dismissed as 'junk DNA' where any transcription was interpreted an artifact of transcriptional noise, it has recently been shown that far more of the genome is pervasively transcribed than first hypothesized [5]. While they do not code for a protein, lncRNAs have been strongly associated with the regulation of epigenetic processes and expression of protein-coding genes. lncRNAs can be arranged as intergenic/intervening, antisense, intronic, overlapping and bidirectional, in relation to their localization to protein-coding genomic loci (Figure 1) [6]. There is now a growing wealth of data to suggest that lncRNAs possess biological function [7-9].

The dysregulation of lncRNAs expression has been implicated in a number of diseases across different tissue types. Merely 7% of disease-associated single nucleotide polymorphism (SNPs) is located within protein-coding regions compared to the 93% of SNPs that are found in noncoding regions [10]. Despite this asymmetry in SNPs distribution, the determination of lncRNAs role in disease pathogenesis remains difficult due to a lack of functional information prohibiting domain and functional prediction that is possible with protein-coding genes.

lncRNAs have been shown to be expressed in a distinct pattern across a number of tissue types. A number of lncRNAs have also been shown to be expressed in discrete cell types and within distinct subcellular structures [11,12]. These findings coincide with notions of lncRNA as regulators of gene expression in specific cell types. Thus, the

Correspondence: Michael Janitz, PhD MD School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia. Tel: +61 2 938 58608. Fax: +61 2 938 51483. E-mail: m.janitz@unsw.edu.au
*These authors contributed equally to this work.

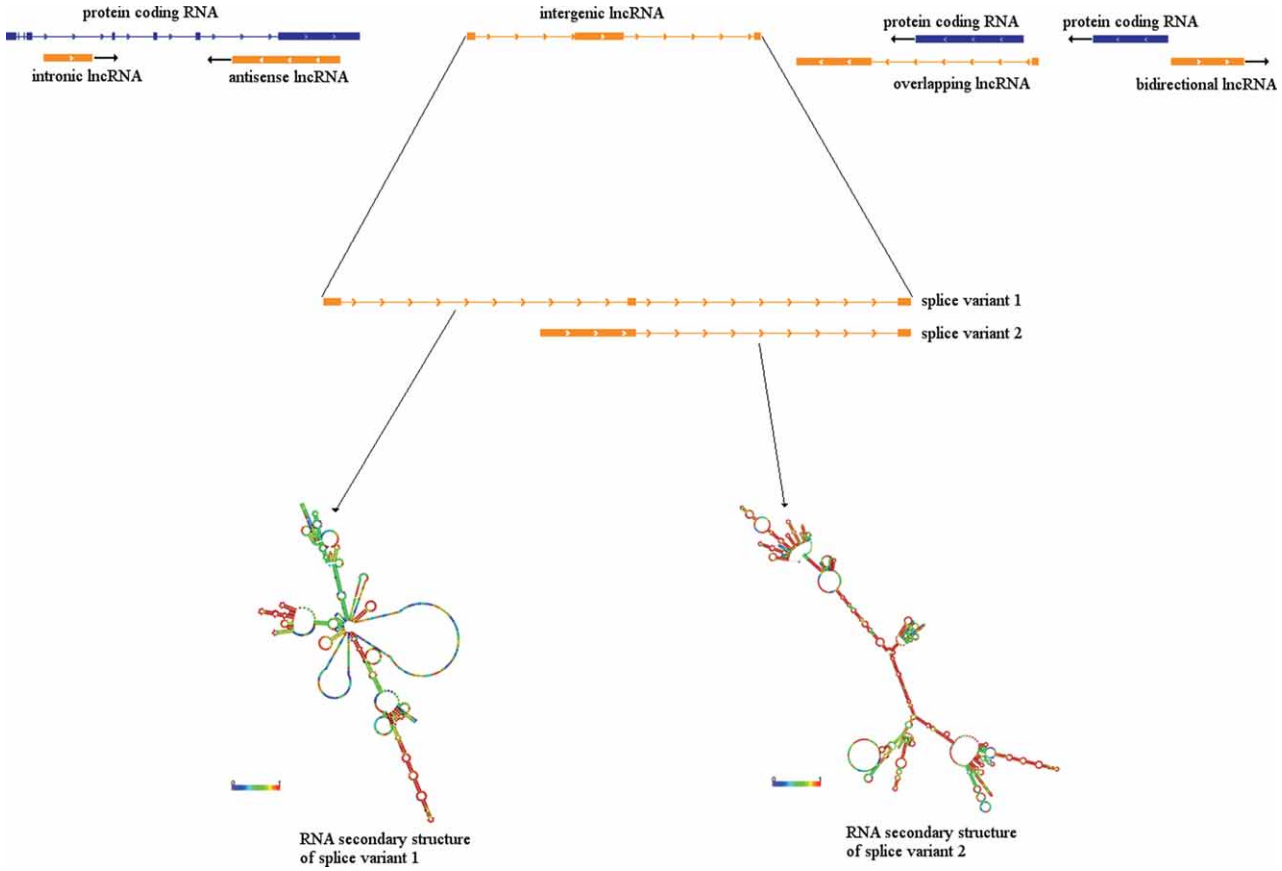


Figure 1. Genomic localization-based classes of lncRNAs. Upper panel: *Intronic lncRNA:* the lncRNA is transcribed from an intronic region of a protein-coding gene. *Antisense lncRNA:* the lncRNA is transcribed from the strand opposite to protein-coding gene, with partial or complete overlap of any intronic or exonic regions. *Intergenic/intervening lncRNA*: the lncRNA is transcribed from a region located between other genes. There is no overlap with any protein-coding genes. *Overlapping lncRNA:* The intron of the lncRNA encompasses a protein-coding gene. *Bidirectional lncRNA:* The lncRNA shares its transcription start site with a protein-coding gene on the opposite strand. Arrows indicate orientation of transcription. Lower panel: lncRNAs can be alternatively spliced to produce numerous splice variants. Here, the intervening lncRNA is spliced to produce two variants. Each of these variants produces RNAs with unique secondary structure. The unique RNA secondary structure can determine function of the lncRNA isoform.

identification and characterization of human lncRNAs with tissue-specific expression become essential in order to determine their relevant functions.

Another interesting property of lncRNAs is their rapid evolution across species. Previously, the conservation of sequence was thought to be evidence of functionality but lncRNAs have proved that this is not always the case. The tissue-specific expression patterns of lncRNAs, coupled with their distinctive conservation patterns, make lncRNAs a unique transcriptional element that warrants further investigation.

## Tissue-specific expression of lncRNAs

LncRNAs exhibit notably higher degree of tissue specificity when compared to protein-coding genes [13,14]. The apparent specificity of lncRNAs throughout various tissue and cell types has been repeatedly highlighted and is indicative of specific regulatory roles within essential cellular processes [11,15,16]. Indeed, if lncRNAs were merely the result of transcriptional noise we would expect little variation in expression levels between tissues [2]. A comparative study investigating tissue specificity of lncRNAs across 11 tissue types found that the majority of lncRNA expression was restricted to discrete tissue types with 67% of lncRNAs demonstrating a tissue-specific expression pattern and with 29% found to be expressed in only one discrete tissue type [17]. This widespread consistency of specific lncRNAs across different tissue types is suggestive of their specific biological function within the individual tissue. Despite this, little work has been done, as of now, in characterizing the expression profiles of tissue-specific lncRNAs beyond possible roles in disease pathogenesis, and in particular cancer [18].

### *Brain*

The brain is the most complex tissue in the human body. Beyond its billions of neurons, the brain comprises a number of other cell types, such as oligodendrocytes, astrocytes and microglia with heterogeneous distribution across anatomical subregions. Due to this complexity in terms of both structure and function, the brain requires a similarly complex regulatory system and as a result is the richest source of lncRNAs in the body [2]. LncRNAs play an essential role in the brain in terms of development, neuronal maintenance and function and have been linked to a number of neurodegenerative diseases [19]. When addressing the brain physiology of humans, the lncRNA repertoire is the greatest point of differentiation from other primates and other vertebrate species entirely due

to the increased developmental complexity of the human brain [20]. Despite a high level of sequence similarity of protein-coding genes between humans and other primates we see far less conformity in the noncoding portion of the genome that is transcribed. Indeed the number of lncRNAs, particularly brain-specific, have been shown to directly increase in correlation with developmental complexity even as the number of protein-coding genes remains relatively unchanged [2].

There is a growing amount of data on the highly tissue-specific lncRNA transcripts located between protein-coding gene loci, known as long intervening ncRNAs (lincRNAs), and their role in regulation of fundamental cellular processes [14]. The lincRNAs are generally expressed at lower levels than protein-coding genes (~10-fold lower) [14]; however, the brain transcriptome contains many unique lincRNA transcripts that are expressed at significantly higher levels than many protein-coding genes, such as the oligodendrocyte maturation-associated lincRNA (*OLMALINC*) [15]. *OLMALINC* is a primate-specific transcript that has been shown to play an essential role in the regulation of genes responsible for human oligodendrocyte maturation [15]. *OLMALINC* is highly expressed in the white matter of the human frontal cortex with expression levels of 71.5 fragments per kilobase of exon per million fragments mapped (fpkm) as determined by RNA-Seq. Such high level of expression indicates a strong regulatory role in oligodendrocytes, which comprise the majority of white matter. The differential expression of *OLMALINC* in gray and white matter (16.2 and 71.5 fpkm, respectively) demonstrates the dynamic nature of the brain transcriptome and the tissue specificity of lincRNAs.

A recent profiling of the transcriptome patterns of gray and white matter highlighted the tissue-specific nature of lincRNAs in a healthy brain [21]. The expression of lincRNAs differs significantly between gray and white matter and this is believed to be largely due to the nonconformity in cell populations between the tissues. Thus in each tissue type there exists divergent transcriptome profiles indicative of discrete roles in brain function for the different tissue types and provides evidence that lincRNAs function in a cell type-specific manner [21].

There is a growing need for the development of more comprehensive expression profiles of lncRNAs for all regions of the human brain [2]. Recent transcriptome analyses of the hippocampus and pre-frontal cortex of the adult mouse brain found highly specific lncRNA expression signatures within subregions of the brain and distinct neuronal populations [22]. A total of 2759 lncRNAs were found to be expressed in the hippocampus, 2561 in the pre-frontal cortex and of these 2390 lncRNAs were expressed in both regions while 24 were differentially regulated. The expression levels of the six highest differentially expressed lncRNAs were then analyzed in the cerebellum and striatum, and compared to that of the hippocampus and the pre-frontal cortex. The majority of these lncRNAs were found to be differentially regulated across all of the brain subregions. A further study using the Allen Brain Atlas showed lncRNAs to be expressed not only in specific subregions of the mouse brain but also in specific cell types and subcellular compartments [23]. The specific localization of lncRNAs supports the premise of their functionality.

The purported functional relationship between lncRNAs, as *cis*-regulatory elements, and adjacent protein-coding genes has also been observed in the human brain [9,24]. It has been found that many of these adjacent protein-coding genes have neurodevelopmental functions and the expression levels of adjacent lncRNAs consistently impact the transcription of the protein-coding genes. Despite this, no consistent pattern has emerged linking the transcription of lncRNAs and adjacent protein-coding genes [25]. These lncRNAs have also been indicated to have an integral role in the regulation of cellular differentiation of neuronal and glial cells, particularly during development [25].

Despite an incomplete annotation of the long noncoding transcriptome and a general lack of functional information, the dysregulation of tissue-specific lncRNAs has been strongly associated with a number of diseases [19]. The differential expression of lncRNAs in healthy and diseased states is shown through comparisons of the transcriptome profiles, which differ significantly in neurodegenerative diseases such as multiple system atrophy (MSA) [26] and Parkinson's disease [27]. Despite consistent association between the dysregulation of brain-specific lncRNAs and neurological disorders [19,28], further research is required to individually categorize and ascertain the functions and molecular mechanisms of action of the dysregulated lncRNAs in order to determine their role in disease progression.

### *Testis*

Testis is a rich source of many unique lncRNA transcripts; however, very little is known about lncRNAs expressed solely in this organ. In-depth analyses of the testis transcriptome using RNA-Seq data have shown a widespread and diverse transcription of both protein-coding and ncRNAs [29]. The testis has two key functions: the secretion of sex hormones and spermatogenesis. The production of spermatozoa is a complex biological process involving multiple stages controlled by epigenetic and molecular mechanisms at both transcriptional and post-transcriptional levels [30]. The need for such regulation has been suggested as a reason for the diversity of the testis transcriptome with specific lncRNAs predicted to play key regulatory roles [14].

A comparative study investigating the five most common cell types involved in spermatogenesis found that in addition to expressing a greater palette of lncRNAs transcripts than cells of the brain or liver, the expression of unique lncRNAs differed significantly between the cells of the testis producing highly specific expression patterns [29]. This was particularly pronounced in spermatids and spermatocytes, which exhibited the highest levels of lncRNA transcription [29].

Currently there are limited studies into human testis-specific lncRNAs expression and as a result we must rely on animal models. A recent study produced lncRNA expression profiles for the testis of a neo-natal and adult mouse [31]. This study identified over 3000 differentially expressed lncRNAs between the neo-natal and adult mice [31]. These dramatic differences in lncRNA expression could indicate a

significant biological role for lncRNA during the testis postnatal development. Furthermore, lncRNAs were found to exhibit a greater spatial and temporal specificity than protein-coding genes consistent with previous studies and supportive of a cell type-specific regulatory role.

*Liver*

The role and function of lncRNAs in the liver is largely unknown but the dysregulation of specific transcripts has been associated with liver diseases such as hepatocellular carcinoma [32] and nonalcoholic steatohepatitis [33]. Liver-specific lncRNAs have also been implicated in the regulation of processes such as lipid metabolism. Liver-specific triglyceride regulator (*lncLSTR*) was found to regulate the clearance of triglyceride and help maintain systemic lipid homeostasis through a novel lncRNA signaling pathway. Its apparently key role in this crucial metabolic process highlights the potential physiological importance of lncRNAs in the liver.

*Heart*

Little is known about the role of lncRNAs in the heart; however, a heart-specific lncRNAs has been found to be involved in cardiac development. FOXF1 adjacent noncoding developmental regulatory RNA (*Fendrr*) is a lateral mesoderm-specific lncRNA that is essential for the development of the heart wall in mouse and was shown to have an orthologous transcript in humans [34]. *Fendrr* was found to modulate chromatin signatures that define gene activity by binding directly to the histone-modifying complexes Polycomb repressive complex 2 (PRC2) and histone–lysine N-methyltransferase 2A (KMT2A), which play a central role in the activation of genes responsible for cell differentiation and lineage commitment. PRC2 and KMT2A act as a repressor and activator of cellular proliferation, respectively, in the heart during embryonic development. The knockdown of *Fendrr* in mice was shown to be lethal to the embryos due to heart wall deficits and significantly impaired heart function demonstrating its importance for normal heart function.

*Skeletal muscle*

Long intergenic ncRNA, muscle differentiation 1 (*Linc-MD1*) has been identified to have a significant role in myogenesis through its control of muscle differentiation [35]. *Linc-MD1* expression is temporally dynamic in order to control the progression through the stages of muscle differentiation where it functions as a competing endogenous RNA for the binding of the microRNAs (miRNAs) *miR-133* and *miR-135*. The two miRNAs regulate the binding of the transcription factors that promote muscle differentiation. Hence, *Linc-MD1* plays a crucial role in the regulation of muscle terminal differentiation through its action as part of a network of regulatory interactions.

*Retina*

Several retina-specific lncRNAs in mice have been identified and determined to be of functional importance in retinal cell development and differentiation. Six3 opposite strand transcript (*Six3OS*) is promoter-associated lncRNA found to play a role in the regulation of retinal cell differentiation through knockdown and overexpression studies [36]. *Six3OS* was also shown to modulate the expression of associated protein-coding genes through the recruitment of histone modification enzymes. *Six3OS* acts as a molecular scaffold that leads to the recruitment of histone modification enzymes. A retina-specific lncRNA ventral anterior homeobox 2, opposite strand (*Vax2os*) was also shown to regulate the cell cycle during mammalian retina development [37]. Overexpression of this transcript during the early stages of development was associated with a reduced rate of retinal cell proliferation. *Vax2os* is so far the only example of a cell type-specific lncRNA regulating the cell cycle during mammalian development.

## Rapid evolution of lncRNAs

LncRNAs show very little conservation in their sequence and they evolve rapidly [38-40]. The predicted amount of shared functional sequence decreases dramatically as the divergence between mammalian species increases, suggesting a very high rate of sequence turnover [41]. The rate of nucleotide substitution in protein-coding sequences is ~ 10%, whereas noncoding sequences have a substitution rate of 90%.

The rapid evolution of lncRNAs originally led to the assumption that they were nonfunctional. Nonfunctional sequences tend to display a similar rate of sequence change when compared to evolutionarily neutral sequences [42]. However, lncRNAs have demonstrated more constraint than random intergenic regions [43]. Ancient lncRNAs (minimum of 90 Myr) show higher levels of long-term exonic sequence conservation than untranslated regions, with the oldest presenting similar levels of constraint with protein-coding exons. In comparison, young lincRNAs (under 25 Myr) show lower levels of exonic sequence conservation than random intergenic regions [39]. This may be due to the fact that young genes demonstrate rapid evolution [44]. Young genes are more susceptible to variable selection pressures than well-established genes [45]. Interestingly, lncRNAs with multiple exons appear to demonstrate greater evolutionary constraints within exons [46].

## Conservation beyond the primary sequence

The sequence of RNAs can differ whilst their secondary structure can be conserved [47,48]. Many lncRNAs showed a number of correlated positions that could be the result of conserved secondary structures (Derrien *et al.* 2012). One of the well-characterized lncRNAs, HOX transcript antisense RNA (*HOTAIR*), is believed to have conserved structures but divergent sequences across species [44]. RNAs can form a variety of structures such as tetraloops [49], GU base pair motifs [50], adenosine platforms, helixes and tandem repeats [51]. These motifs have demonstrated sequence conservation, for example the hairpin loop and the tRNA-like structure in the lncRNA metastasis-associated lung adenocarcinoma transcript 1 (*Malat1*) [52]. The majority of the helixes appear to be

conserved across a variety of species, in comparison to the base paired regions, which are not so well conserved [53]. This theory is supported by the fact that many lncRNAs with differing sequences are able to bind to the same protein [54,55].

The functional role of the lncRNA may also be conserved. One established lncRNA is X-inactive-specific transcript (*Xist*), which is involved in X-chromosome inactivation. The function of *Xist* is conserved across mammals, even though the sequence is evolving at a high rate [56]. In addition mouse and zebra fish lncRNAs, involved in embryonic development, did not have conserved sequences, whereas the function appears to be conserved [57]. If the functional roles of lncRNAs are conserved across species, then it is most likely that their loci will also be conserved [38]. Indeed, studies have found that lncRNAs have conserved synteny across a range of species [39,58].

## LncRNA evolution in primates

King and Wilson first proposed that the major biological differences between humans and chimpanzees are due to gene regulation, not differences in sequence [59]. There are probably too few changes in the amino acid sequence of proteins to result in the phenotypical differences between humans and chimpanzees [20]. In fact, a larger number of protein-coding genes are conserved for primates when compared to lncRNAs; 92% of human intergenic lncRNAs are expressed in chimpanzee or bonobo and ~ 72% are expressed in the macaque. In comparison > 98% of protein-coding genes is conserved for all primates [39].

It is believed that human brain evolution has occurred through changes in noncoding parts of the genome [60]. The human brain is in fact a rich source of lincRNAs, further supporting this theory [21,26]. The majority of gene expression differences between the brains of humans and nonhuman primates involved upregulation of gene expression in humans [61]. While this may be due to higher levels of neuronal activity, it has been found that genes critical for neural development are upregulated across mammals [62].

Brain growth patterns vary across primate species [63] and humans show a unique pattern of expression [61]. The expression pattern of genes in the chimpanzee brain cortex is more similar to gene expression patterns in macaques than humans [64]. This indicates an increase in the rate of evolution in gene regulation in the human lineage [64]. The expression of human-specific genes was greater in the frontal lobe in comparison to the hippocampus and caudate. This suggests that the majority of evolutionary change in the human brain was focused in the frontal lobe [20]. Genes in the frontal lobe that are associated with neuron projections, neurotransmitter transport, synapses, axons and dendrites, as well as genes implicated in schizophrenia showed increased connectivity in the human brain when compared to chimpanzees and macaques [20].

One example of a noncoding gene that is thought to have evolved a unique function in humans is the human accelerated region 1 (*HAR1*). It has been suggested that *HAR1* has been evolutionarily selected for increased stability [65]. It is believed that A/T to C/G substitutions led to a more stable secondary structure in *HAR1* [62]. Forkhead box protein P2 (*FOXP2*) and abnormal spindle-like microcephaly associated protein (*ASPM*), which are involved in speech production and brain size respectively, have undergone the same kind of evolutionary change [66,67].

## Methods of detecting lncRNAs

RNA-Seq is a high-throughput next-generation sequencing technique that is capable of measuring RNA expression levels and providing an accurate picture of the transcriptome [68]. RNA-Seq has numerous advantages over other transcriptome profiling techniques such as microarrays. RNA-Seq has a higher resolution, lower levels of background noise, lower requirement of input RNA and can detect a greater range of expression levels [69]. The most important aspect of RNA-Seq is that it can be used to assemble transcriptomes *de novo*; this allows for the discovery of un-annotated transcripts and novel splicing events [69]. This ability makes RNA-Seq an ideal tool for the identification species- and tissue-specific lncRNAs, many of which have not been previously annotated.

More recently, slight modifications of the template preparation stage of RNA-Seq have allowed for the strand of origin from which an RNA molecule is transcribed from to be tracked, thus allowing for the identification of antisense transcription. These techniques are known as strand-specific RNA-Seq [70,71]. While a multitude of different strand-specific RNA-Seq exist, currently the most widely used is the dUTP second-strand marking method [70]. Strand-specific RNA-Seq techniques allow for the identification of antisense transcripts and this feature is particularly relevant to lncRNAs. Examples of antisense lncRNAs include TSIX transcript, XIST antisense RNA (*TSIX*) [72] and the beta-site APP-cleaving enzyme 1 antisense RNA (*BACE1-AS*) [73]. It is estimated that between 20–30% of human transcripts have an antisense partner [74,75]. Further, the amount of antisense transcription will vary from cell type to cell type [76]. Another important technical advance concerning RNA-Seq is the use of ribosomal depletion to select the RNA fraction for sequencing rather than selecting only those transcripts that are polyadenylated. Ribosomal depletion removes ribosomal RNA from the samples, allowing for the selection of both polyadenylated positive (poly(A)+) and polyadenylated negative (poly(A)−) fractions for sequencing [77]. This is important as large amounts of transcriptional output in eukaryotic cells is poly(A)− [78]. As ribosomal depleted strand-specific RNA-Seq becomes the standard for all transcriptome-profiling experiments, it is expected that more lncRNAs will be found throughout different tissue types in the human body.

Raw RNA-Seq data needs to be processed and analyzed to answer all sorts of bioinformatics enquires, including investigation of gene/transcript expression levels, detection of alternative splicing events and identification of unannotated genes/transcripts. In brief to analyze RNA-Seq data, first the reads must be mapped to the reference genome, next

transcripts are assembled followed by a differential expression analysis. A common workflow currently used by researchers is known as the Tuxedo suite, which utilizes the software packages Tophat, Cufflinks and Cuffdiff [79]. This workflow is ideal for the identification of lncRNAs as it has the ability to identify novel splicing events and un-annotated transcripts down to the resolution of a single base. It also takes advantage of data generated by ribosomal strand-specific RNA-Seq to locate antisense transcripts.

A typical RNA-Seq experiment will produce vast amounts of data. Generally it is not feasible to analyze data on a personal computer due to limitations in storage size and raw processing power. These problems can be overcome through the use of high-performance computing (HPC) clusters. A HPC cluster consists of multiple nodes, with each node containing one or more central processing units (CPUs), each with numerous cores. HPC clusters are normally a resource shared across a major institute such as a university or hospital. Another alternative is to take advantage of cloud computing services such as Amazon Web Services (AWS) (http://aws.amazon.com/). AWS allows researchers to dynamically adjust the computing power and storage requirements based on current requirements and has potential computing power much larger than any HPC cluster.

## Concluding remarks and future directions

Only recently has technology been able to identify lncRNAs using high-throughput methods such as RNA-seq. Questions still remain as to how many of the proposed lncRNAs are functional, what that function is and the role that they have played in evolution. More knockdown and overexpression studies are necessary to explore the diverse roles that lncRNAs possess. For example, the overexpression of the 3'UTR region of the phosphatase and tensin homolog pseudogene 1 (*PTENP1*), through retroviral vectors, revealed its role in the regulation of the phosphatase and tensin homolog (*PTEN*) [80]. RNAi of *OLMALINC* in human oligodendrocytes [15] revealed the perturbation of the expression of genes involved in the maturation and myelination of oligodendrocytes. A systematic approach is needed to attempt to elucidate the function of various lncRNAs, which could prove difficult due to the species and tissue specificity of many lncRNAs.

In order to better determine lncRNA role as part of a regulatory network it is essential to produce comprehensive, functional annotations for lncRNAs similar to those that exist for protein-coding genes. This is especially relevant for those novel lncRNAs associated with human diseases. As a result of advances into ncRNA research, there are several public databases of annotated lncRNAs; however complete functional characterization of all lncRNAs is needed beyond merely basic sequence and transcript information [81]. A number of lncRNA databases currently exist, each with different focuses which determines their utility. This includes LNCipedia with broad coverage of a high number of lncRNAs, lncRNA database (lncRNAdb) providing in-depth annotation of a variety of different lncRNAs and GermlncRNA with a tissue-specific catalogue of lncRNAs.

The lncRNAdb (http://www.lncrnadb.org/) provides a summary of known eukaryotic lncRNAs. lncRNAdb differs from many other databases as entries must be supported by literature and they do not pull their data from unconfirmed sources to ensure validity [82]. Thus, lncRNAdb serves as a reliable resource for exploration of eukaryotic lncRNAs; however it represents only a small fraction of currently annotated lncRNAs. The database currently contains 287 eukaryotic lncRNAs that have been manually curated and described independent of scientific literature [83]. It provides information on lncRNA function, sequences, expression data and relevant supportive literature. Of these, 100 lncRNAs have had function determined through direct *in vitro* and/or *in vivo* experiments.

GermlncRNA (http://germlncrna.cbiit.cuhk.edu.hk/) is a web-based lncRNA catalog containing annotations of male germ-cell specific lncRNAs [84]. This catalog currently contains 110476 annotated lncRNAs and 2790 novel lncRNAs, the latter classified as novel as they were unannotated in any of the public genomic databases. The database was created through the integration of male germ transcriptome profiles from microarray, RNA-Seq and GermSAGE studies [84]. A tissue-specific focus allows for more comprehensive gene coverage, especially important for the testes, which are a rich source of lncRNAs.

LNCipedia (http://www.lncipedia.org/) is a comprehensive database for annotated human lncRNAs generated through the incorporation of data obtained from a number of different sources. This allowed for a rapid increase of the gene entries from 21488 annotated lncRNAs in LNCipedia v.1.0 to 111685 annotated lncRNAs in LNCipedia v.3.1. [85]. Along with sequence/transcript information, secondary structure and protein-coding potential are explored in detail for many of the cataloged lncRNAs [86]. A strategy to detect lncRNAs with protein-coding potential has been integrated within the database, which reanalyzes the mass spectrometry data publicly available from the PRIDE database. The wide scale of LNCipedia allows for incorporation of its content into large genomic projects, including development of customized microarrays allowing genome-wide surveys of lncRNA expression.

These databases were created through the integration of pre-existing public resources. While this allows for large amounts of information to be shared and combined, it also led to lncRNA predictions that greatly vary between individual repositories. This is due to differences in methodology, classification and assembly algorithms, which result in many lncRNAs to be missed or improperly categorized [81]. Constant verification is required to ensure the validity of the database, which is particularly difficult to achieve in large lncRNA databases. This remains an issue with the number of lncRNAs being annotated constantly increasing but experimental functional characterization lagging behind. Before function can be determined for all annotated lncRNAs, for example utilizing knockdown and overexpression approaches, a complete and comprehensive catalog of evolutionary conservation and tissue-specific expression for these transcripts must firstly be produced.

**Acknowledgements**

The authors would like to thank Cathy and Travis Hore and their family and friends for generous donations to the UNSW MSA Research Fund.

**Declarations of interest**

The authors report no declarations of interest.

## References

[1] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63.
[2] Mattick JS. The central role of RNA in human development and cognition. FEBS Lett 2011;585:1600–16.
[3] Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 2007; 29:288–99.
[4] Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem 2012;81:145–66.
[5] Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. Science 2005;309:1559–63.
[6] Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. Nat Struct Mol Biol 2015;22:5–7.
[7] Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 2006;22:1–5.
[8] Guttman M, Garber M, Levin JZ, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 2010;28:503–10.
[9] Ponjavic J, Oliver PL, Lunter G, Ponting CP. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet 2009;5:e1000617.
[10] Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009;106:9362–7.
[11] Sone M, Hayashi T, Tarui H, et al. The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. J Cell Sci 2007;120:2498–506.
[12] Bond CS, Fox AH. Paraspeckles: nuclear bodies built on long noncoding RNA. J Cell Biol 2009;186:637–44.
[13] Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. Nature 2012;489:101–8.
[14] Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011;25:1915–27.
[15] Mills JD, Kavanagh T, Kim WS, et al. High expression of long intervening non-coding RNA OLMALINC in the human cortical white matter is associated with regulation of oligodendrocyte maturation. Mol Brain 2015;8:2.
[16] Wu SC, Kallin EM, Zhang Y. Role of H3K27 methylation in the regulation of lncRNA expression. Cell Res 2010;20:1109–16.
[17] Sasaki YT, Sano M, Ideue T, et al. Identification and characterization of human non-coding RNAs with tissue-specific expression. Biochem Biophys Res Commun 2007;357:991–6.
[18] Quagliata L, Terracciano LM. Liver diseases and long non-coding RNAs: new insight and perspective. Front Med 2014;1:35.
[19] Wu P, Zuo X, Deng H, et al. Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. Brain Res Bull 2013;97:69–80.
[20] Konopka G, Friedrich T, Davis-Turak J, et al. Human-specific transcriptional networks in the brain. Neuron 2012;75:601–17.
[21] Mills JD, Kavanagh T, Kim WS, et al. Unique transcriptome patterns of the white and grey matter corroborate structural and functional heterogeneity in the human frontal lobe. PLoS One 2013;8:e78480.
[22] Kadakkuzha BM, Liu XA, McCrate J, et al. Transcriptome analyses of adult mouse brain reveal enrichment of lncRNAs in specific brain regions and neuronal populations. Front Cell Neurosci 2015;9:63.
[23] Mercer TR, Dinger ME, Sunkin SM, et al. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA 2008;105:716–21.
[24] Orom UA, Derrien T, Beringer M, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell 2010;143:46–58.
[25] Mercer TR, Qureshi IA, Gokhan S, et al. Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. BMC Neurosci 2010;11:14.
[26] Mills JD, Kim WS, Halliday GM, Janitz M. Transcriptome analysis of grey and white matter cortical tissue in multiple system atrophy. Neurogenetics 2015;16:107–22.
[27] Soreq L, Guffanti A, Salomonis N, et al. Long non-coding RNA and alternative splicing modulations in Parkinson's leukocytes identified by RNA sequencing. PLoS Comput Biol 2014;10: e1003517.
[28] Barry G, Briggs JA, Vanichkina DP, et al. The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. Mol Psychiatry 2014;19:486–94.
[29] Soumillon M, Necsulea A, Weier M, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Reports 2013;3:2179–90.
[30] Bao J, Wu J, Schuster AS, et al. Expression profiling reveals developmentally regulated lncRNA repertoire in the mouse male germline. Biol Reprod 2013;89:107.
[31] Sun J, Lin Y, Wu J. Long non-coding RNA expression profiling of mouse testis during postnatal development. PloS One 2013;8: e75750.
[32] Huang JF, Guo YJ, Zhao CX, et al. Hepatitis B virus X protein (HBx)-related long noncoding RNA (lncRNA) down-regulated expression by HBx (Dreh) inhibits hepatocellular carcinoma metastasis by targeting the intermediate filament protein vimentin. Hepatology 2013;57:1882–92.
[33] Takahashi K, Yan I, Haga H, Patel T. Long noncoding RNA in liver diseases. Hepatology 2014;60:744–53.
[34] Grote P, Wittler L, Hendrix D, et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev Cell 2013;24:206–14.
[35] Cesana M, Cacchiarelli D, Legnini I, et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell 2011;147:358–69.
[36] Rapicavoli NA, Poth EM, Zhu H, Blackshaw S. The long noncoding RNA Six3OS acts in trans to regulate retinal development by modulating Six3 activity. Neural Deve 2011;6:32.
[37] Meola N, Pizzo M, Alfano G, et al. The long noncoding RNA Vax2os1 controls the cell cycle progression of photoreceptor progenitors in the mouse retina. RNA 2012;18:111–23.
[38] Kutter C, Watt S, Stefflova K, et al. Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. PLoS Genet 2012;8:e1002841.
[39] Necsulea A, Soumillon M, Warnefors M, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 2014;505:635–40.
[40] Ulitsky I, Shkumatava A, Jan CH, et al. Conserved Function of lincRNAs in Vertebrate Embryonic Development Despite Rapid Sequence Evolution. Cell 2011;147:1537–50.
[41] Meader S, Ponting CP, Lunter G. Massive turnover of functional sequence in human and other mammalian genomes. Genome Res 2010;20:1335–43.
[42] Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell 2009;136:629–41.
[43] Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009;458:223–7.
[44] He S, Liu SP, Zhu H. The sequence, structure and evolutionary features of HOTAIR in mammals. BMC Evol Biol 2011;11:14.
[45] Vishnoi A, Kryazhimskiy S, Bazykin GA, et al. Young proteins experience more variable selection pressures than old proteins. Genome Res 2010;20:1574–81.
[46] Chodroff RA, Goodstadt L, Sirey TM, et al. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol 2010;11:R72.

[47] Lindgreen S, Gardner PP, Krogh A. Measuring covariation in RNA alignments: physical realism improves information measures. Bioinformatics 2006;22:2988–95.
[48] Torarinsson E, Sawera M, Havgaard JH, et al. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Res 2006;16:885–9.
[49] Woese CR, Winker S, Gutell RR. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". Proc Natl Acad Sci USA 1990;87:8467–71.
[50] Gautheret D, Konings D, Gutell RR. G.U base pairing motifs in ribosomal RNA. RNA 1995;1:807–14.
[51] Gautheret D, Konings D, Gutell RR. A major family of motifs involving G? A mismatches in ribosomal RNA. J Mol Biol 1994;242:1–8.
[52] Smith MA, Gesell T, Stadler PF, Mattick JS. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res 2013;41:8220–36.
[53] Novikova IV, Hennelly SP, Sanbonmatsu KY. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Res 2012;40:5034–51.
[54] Guttman M, Donaghey J, Carey BW, et al. LincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 2011;477:295–U60.
[55] Khalil AM, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA 2009;106:11667–72.
[56] Nesterova TB, Slobodyanyuk SY, Elisaphenko EA, et al. Characterization of the genomic xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. Genome Res 2001;11:833–49.
[57] Ulitsky I, Bartel DP. LincRNAs: genomics, evolution, and mechanisms. Cell 2013;154:26–46.
[58] Consortium TF, Carninci P, Kasukawa T, et al. The transcriptional landscape of the mammalian genome. Science 2005;309:1559–63.
[59] King M, Wilson A. Evolution at two levels in humans and chimpanzees. Science 1975;188:107–16.
[60] Haygood R, Babbitt CC, Fedrigo O, Wray GA. Contrasts between adaptive coding and noncoding changes during human evolution. Proc Natl Acad Sci USA 2010;107:7853–7.
[61] Caceres M, Lachuer J, Zapala MA, et al. Elevated gene expression levels distinguish human from non-human primate brains. Proc Natl Acad Sci USA 2003;100:13030–5.
[62] Lambert N, Lambot MA, Bilheu A, et al. Genes expressed in specific areas of the human fetal cerebral cortex display distinct patterns of evolution. PLoS One 2011;6:e17753.
[63] Leigh SR. Brain growth, life history, and cognition in primate and human evolution. Am J Primatol 2004;62:139–64.
[64] Enard W, Khaitovich P, Klose J, et al. Intra- and interspecific variation in primate gene expression patterns. Science 2002;296:340–3.
[65] Pollard KS, Salama SR, Lambert N, et al. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 2006;443:167–72.
[66] Enard W, Przeworski M, Fisher SE, et al. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 2002;418:869–72.
[67] Evans PD, Anderson JR, Vallender EJ, et al. Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. Hum Mol Genet 2004;13:489–94.
[68] Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol 2010;2010:853916.
[69] Courtney E, Kornfeld S, Janitz K, Janitz M. Transcriptome profiling in neurodegenerative disease. J Neurosci Methods 2010;193:189–202.
[70] Levin JZ, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 2010;7:709–15.
[71] Mills JD, Kawahara Y, Janitz M, Strand-specific RNA-Seq provides greater resolution of transcriptome profiling. Curr Genomics 2013;14:173–81.
[72] Lee JT, Davidow LS, Warshawsky D. Tsix, a gene antisense to Xist at the X-inactivation centre. Nat Genet 1999;21:400–4.
[73] Engstrom PG, Suzuki H, Ninomiya N, et al. Complex loci in human and mouse genomes. PLoS Genet 2006;2:e47.
[74] Chen J, Sun M, Kent WJ, et al. Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res 2004;32:4812–20.
[75] Ozsolak F, Kapranov P, Foissac S, et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell 2010;143:1018–29.
[76] He Y, Vogelstein B, Velculescu VE, et al. The antisense transcriptomes of human cells. Science 2008;322:1855–7.
[77] Chen Z, Duan X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. Methods Mol Biol 2011;733:93–103.
[78] Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 2005;308:1149–54.
[79] Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with tophat and cufflinks. Nat Protoc 2012;7:562–78.
[80] Poliseno L, Salmena L, Zhang J, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 2010;465:1033–8.
[81] Fritah S, Niclou SP, Azuaje F. Databases for lncRNAs: a comparative evaluation of emerging tools. RNA 2014;20:1655–65.
[82] Amaral PP, Clark MB, Gascoigne DK, et al. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res 2011;39:D146–51.
[83] Quek XC, Thomson DW, Maag JL, et al. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res 2015;43:D168–73.
[84] Luk AC, Gao H, Xiao S, et al. GermlncRNA: a unique catalogue of long non-coding RNAs and associated regulations in male germ cell development. Database (Oxford) 2015;2015:bav044.
[85] Volders PJ, Verheggen K, Menschaert G, et al. An update on LNCipedia: a database for annotated human lncRNA sequences. Nucleic Acids Res 2015;43:D174–80.
[86] Volders PJ, Helsens K, Wang X, et al. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res 2013;41:D246–51.