CellPress

Review

Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs

Bethany Signal,[1,2] Brian S. Gloss,[1,2,‡] and Marcel E. Dinger[1,2,‡,*]

Although a considerable portion of eukaryotic genomes is transcribed as long noncoding RNAs (lncRNAs), the vast majority are functionally uncharacterised. The rapidly expanding catalogue of mechanistically investigated lncRNAs has provided evidence for distinct functional subclasses, which are now ripe for exploitation as a general model to predict functions for uncharacterised lncRNAs. By utilising publicly-available genome-wide datasets and computational methods, we present several developed and emerging *in silico* approaches to characterise and predict the functions of lncRNAs. We propose that the application of these techniques provides valuable functional and mechanistic insight into lncRNAs, and is a crucial step for informing subsequent functional studies.

Trends

lncRNAs represent a large proportion of the transcriptome that is currently sparsely annotated.

Expression-based experiments often yield a large number of lncRNAs cosegregating with the biological system being studied.

The ability to effectively enrich candidate pools for lncRNAs most likely to be involved in the phenotype under study is crucial.

Powerful computational methods for investigating lncRNA function and biology from experimental and sequence information are emerging.

Combining several computational methods is an effective approach to maximise research findings and effectively deploy laboratory resources.

The Emerging Need for Computational Methodologies to Discern Functional lncRNAs

Over the past decade advances in sequencing methodologies have revealed the transcriptional complexity of the genome. Early use of genome tiling arrays and CAGE-sequencing led to the elucidation that a much greater portion of the genome is transcribed than previously expected, with the majority of transcription producing non-protein coding RNAs [1,2]. Initially hampered by characteristic low expression, biological specificity, and lack of sequence conservation [3], the functions of the group of long (>200 nt) noncoding RNAs (lncRNAs) typically remained overlooked in biological systems. Development of high-throughput RNA-sequencing methods allowed the unbiased analysis of transcription with a much greater depth and dynamic range than preceding technologies – such as expression microarrays and cDNA cloning libraries. Following these technological advancements, the catalogue of transcribed lncRNAs grew, and now outnumbers protein-coding genes in humans [4]. Despite such an extensive catalogue, only a small – but continually growing – proportion have functionally characterised roles [5]. lncRNAs were initially thought to be involved primarily in the epigenetic regulation of transcription [6], presumably due to the well-characterised examples XIST [7], H19 [8], and HOTAIR [9]. More recent examples illustrate their potential to regulate gene and protein expression at seemingly every step in the process, but also perform non-regulatory roles (Table 1). Classification of such a diverse group of molecules into distinct classes – much like the classification of proteins – is challenging because relatively few have been extensively functionally characterised [5], although some systems have been proposed [10,11].

Ultimately, lncRNA functionality should be tested and investigated using experimental approaches. However, classical methods such as gene knockdown, overexpression, or editing are often not suited to the analysis of what is typically an extensive pool of candidates.

[1]Garvan Institute of Medical Research, Sydney, Australia
[2]St Vincent's Clinical School, University of New South Wales, Sydney, Australia
[‡]These authors contributed equally to this manuscript.

*Correspondence:
m.dinger@garvan.org.au (M.E. Dinger).

 http://dx.doi.org/10.1016/j.tig.2016.08.004

Employment of large-scale RNAi screens has been successful in investigating hundreds of lncRNA, and can in part overcome the problem of scale in experimental approaches [12–14]. In addition, interpretation of the direct effects of these perturbations without further experimentation is difficult. Now evident from several established examples, lncRNAs can exert their functions in measurably different ways (Box 1). Broadly applicable and measurable features of functional lncRNAs are valuable in prioritising candidates for further functional examination and in guiding the design of such experiments. Recently, several *in silico* methods have been developed that anticipate roles, characteristics, or biological importance of lncRNAs through these features. These approaches typically do not require additional experimental data to be generated, instead relying on the extensive amount of publicly-available data, databases, algorithms, and predictive tools developed for lncRNA research. Application of several of these approaches to multiple transcripts is therefore a viable analytical approach, and has been successfully utilised for genome-scale annotations of lncRNAs [15] and particular subclasses of lncRNAs [16–18]. Only a subset of the extensive range of computational approaches can provide evidence for functional importance or probe mechanistic properties. We present several methods that can be used to investigate uncharacterised lncRNAs, and relate the outputs to probable biological function. We propose that systematic application of these techniques should become a fundamental step in investigating functionality before experimental design.

Computational Techniques to Impute lncRNA Function

Core features of functional lncRNAs can be probed via an array of computational methods strengthened by publicly-available datasets. Gene expression information is commonly utilised to detect potential regulatory targets – currently the most common mechanism of action – or involvement in biological processes. Expression-based approaches depend on experimental data, which continues to be generated across a diverse repertoire of biological contexts and made publicly available through repositories such as ENA and SRA (Table 2). In addition, several methods based on binding and sequence features can be applied to build evidence for function and point towards particular mechanisms. While currently less-developed compared to approaches using experimental datasets, predictive algorithms are beginning to show promise in interrogating lncRNA functional properties. Because predictive tools can work on lowly and specifically expressed transcripts, it is foreseeable that their continued development will enable functional characterisation of a much wider pool of lncRNAs. We propose that, by employing a combination of the techniques outlined, evidence of function and/or functional class can be used to guide effective experimental characterisation of lncRNAs.

Box 1. Measuring Evidence of lncRNA Functional Mechanisms

The number of mechanistically investigated lncRNAs, while small, can provide insight into how to probe potential features of mechanistic functionality using computational techniques.

The majority of currently characterised lncRNAs function through regulating gene or protein expression (Figure 2) by coordinating epigenetic, transcriptional, or post-transcriptional processes. The immediate effect of regulatory lncRNA expression is to repress or induce target molecule expression, measurable by positive or negative correlations of expression profiles. RNA-Seq and microarrays are widely utilised in understanding biological processes, and the resulting global expression information provides a rich resource for identifying potential regulatory relationships at the RNA level.

Regulatory and non-regulatory relationships encompass a diversity of functional mechanisms, most of which employ the ability of lncRNAs to bind in a sequence-specific manner to DNA, RNA, and protein. Most characterised lncRNAs appear to act through a combination of binding interactions, such as recruiting chromatin-modifying complexes to targeted DNA regions (Table 1), that drive specific regulatory changes to their targets. DNA binding implies a functional mechanism that acts at the epigenetic or transcriptional regulation level. RNA binding is the basis underlying targeted post-transcriptional regulatory mechanisms. While binding to DNA and RNA is typically described as being responsible for the specific targeted functions of lncRNAs, protein binding is likely to be responsible for the specific mechanism through which a lncRNA acts. As well as the effect of the lncRNA, these specific interactions offer insight into the mechanism by which it is functioning.

Glossary

***Cis*-regulatory:** a type of regulatory relationship defined by the close genomic proximity between the regulator and target genes.
Enhancer RNA (eRNA): a type of lncRNA transcribed from a genomic region possessing chromatin modifications typical of enhancer DNA. Enhancer lncRNAs may be non-functional, with the DNA being responsible for enhancer activity.
Functional enrichment: process by which functional annotations of groups of genes are tested for statistical enrichment in particular groups above background level.
G-quadruplex (GQ): a structure formed within guanine-rich nucleic acid sequences. Four guanine bases form a square guanine tetrad through Hoogsteen pairing, and these tetrads stack to form a quadruplex.
Machine learning: a method of data analysis where algorithms are applied to generate models of a feature of interest, and predict this feature in new datasets.
Product-independent transcription: mechanism by which transcription of a gene locus alone is sufficient for function. The process of transcription results in changes in chromatin structure or occupancy, thereby affecting expression of nearby loci.
Promoter lncRNA (plncRNA): a type of lncRNA with chromatin modifications at the promoter similar to protein-coding genes.
R-loop: hybrid structure composed of two DNA strands and one RNA strand. RNA forms a duplex with one strand of DNA, displacing the other strand.
Triplex: a structure of three strands of DNA and/or RNA bound together. Multiple base triples are formed between a duplex and a third strand by Watson–Crick and Hoogsteen pairing.

Table 1. Functional Mechanisms of lncRNAs

Stage of Action	Primary Mechanism of Action	lncRNA	Detailed Mechanism of Action	Predominant Localisation	Relationship with Target (*Cis*/*Trans*)	Relationship with Target Expression
Epigenetic Regulation	Chromatin remodelling	*fbp1* upstream RNAs [98]	Transcription of loci required for opening of chromatin at the *fbp1* promoter. Likely to be independent of the RNA product.	Unknown	*Cis*	Positive
	Chromatin remodelling	lncTCF7 [99]	Recruits the SWI/SNF complex to TCF7 to induce expression via chromatin remodelling.	Nuclear	*Cis*	Positive
	Chromatin remodelling	VIM-AS1 [100]	Forms a R-loop nearby the *VIM* transcription start-site, promoting the opening of chromatin structure, and binding of NF-κB to the promoter, increasing *VIM* expression.	Nuclear	*Cis*	Positive
	Chromatin remodelling	Haunt [101]	Haunt RNA inhibits chromatin looping and accessibility of *HOXA* enhancers, repressing *HOXA* expression.	Nuclear and chromatin	*Cis*	Negative
	Chromatin remodelling	cga eRNA [102]	Facilitates looping between the enhancer and promoter of cga.	Unknown	*Cis*	Positive
	Positive chromatin modification	Oct4P4 [103]	Complexes with SUV39H1 HMTase to remove SUV39H1 and H3K3Me3 marks at the promoter of *Oct4*.	Nuclear	*Trans*	Positive
	Positive chromatin modification	HOTTIP [104]	Interacts with WDR5, targeting WDR5/MLL complexes to *HOXA* to drive gene expression.	Nuclear	*Cis*	Positive
	Negative chromatin modification	HOTAIR [105]	3′-Domain binds to the LSD1/CoREST/REST complex to mediate silencing of *HOXD*.	Nuclear [106]	*Trans*	Negative
	Negative chromatin modification	COOLAIR [107]	COOLAIR transcripts are induced by cold, and bind to the *FLC* locus to induce silencing by reducing levels of H3K36me3.	Nuclear and cytoplasmic	*Cis*	Negative
	Negative chromatin modification	Braveheart [108]	Acts as an upstream regulator of *MesP1* by interacting with the PRC2 complex.	Nuclear and cytoplasmic	*Trans*	NA (unknown target)
	DNA methylation	Dali [109]	Interacts with DMNT1 to methylate promoter regions of multiple target genes.	Chromatin (nuclear)	*Trans*	Negative
	DNA methylation	Dum [110]	Silences *Dppa2* in *cis* by recruiting Dnmt1, Dnmt3a and Dnmt3b.	Nuclear and cytoplasmic	*Cis*	Negative
	DNA methylation	TARID [111]	Binds to GADD45A to demethylate the promoter of *TCF21*.	Nuclear	*Cis*	Positive

Table 1. (continued)

Stage of Action	Primary Mechanism of Action	lncRNA	Detailed Mechanism of Action	Predominant Localisation	Relationship with Target (*Cis*/*Trans*)	Relationship with Target Expression
	DNA methylation/ chromatin remodelling	pRNA [112,113]	Forms a RNA:DNA triplex with rDNA promoters to promote epigenetic silencing. Interacts with NoRC and DNMT3b influencing nucleosome positioning and promoter methylation respectively.	Nuclear	*Trans*	Negative
	DNA methylation/ chromatin remodelling	Kcnq1ot1 [114,115]	Recruits EZH2, G9a and DNMT1 to epigenetically silence imprinted genes.	Nuclear	*Cis*	Negative
Transcriptional regulation	Transcriptional interference	Airn [116]	Transcription prevents Pol II binding to the *Igfr2* promoter.	Nuclear	*Cis*	Negative
	Transcription factor cofactor	DEANR1 [117]	Associates with SMAD2/3, directing it to the *FOXA2* promoter to induce expression.	Nuclear	*Cis*	Positive
Post-transcriptional regulation	RNA splicing/RNA translation	ZEB2 NAT [118]	Prevents the splicing of the Zeb2 5′-UTR[a], increasing its translation.	Unknown	*Cis*	Positive
	RNA processing	PCA3 [75]	Forms a duplex with PRUNE2 pre-mRNA. ADARs[b] bind to this structure to mediate RNA editing and reduce levels of PRUNE2.	Nuclear	*Cis*	Negative
	RNA stability	TINCR [74]	Binds to STAU1 to mediate the stabilisation of multiple mRNAs	Cytoplasmic	*Trans*	Positive
	RNA stability	½-sbsRNA1 [81]	Binds to SERPINE1 and FLJ21870 mRNA to activate STAU1-mediated decay.	Cytoplasmic	*Trans*	Negative
	RNA stability	Sirt1 AS [119]	Binds to Sirt1 3′-UTR[a], competing with miR-43a to increase Sirt1 stability.	Cytoplasmic	*Cis*	Positive
	miRNA binding	lncRNA-ATB [120]	Binds to miR-200s, resulting in upregulation of ZEB1 and ZEB2.	Cytoplasmic	*Trans*	Positive
	miRNA binding	lincRNA-RoR [121]	Binds to miR-145 as a competing endogenous RNA, upregulating several pluripotency associated genes.	Cytoplasmic	*Trans*	Positive
	miRNA binding	CDR1as [122–124]	Forms a stable circular RNA that harbours multiple binding sites for miR-7, preventing the destabilisation of CDR1.	Cytoplasmic	*Trans*	Positive
	RNA localisation	ASBEL [125]	Forms a duplex with ANA/BTG3 mRNA, preventing its export to the cytoplasm.	Nuclear	*Cis*	Negative (protein)

Table 1. (continued)

Stage of Action	Primary Mechanism of Action	lncRNA	Detailed Mechanism of Action	Predominant Localisation	Relationship with Target (*Cis*/*Trans*)	Relationship with Target Expression
	RNA translation	PU.1 AS [126–128]	Binds to PU.1 mRNA, inhibiting its translation.	Unknown	*Cis*	Negative (protein)
	RNA translation	lincRNA-p21 [129]	Associates with CTNNB1 and JUNB mRNAs to prevent translation.	Cytoplasmic [129]	*Trans*	Negative (protein)
	RNA translation	Antisense UCHL1 [62]	Upon rapamycin treatment, is transported to the cytoplasm, where it mediates the association of UCHL1 mRNA and polysomes.	Nuclear (cytoplasmic upon rapamycin treatment)	*Cis*	Positive (protein)
	Protein stability	AOC4P [130]	Binds to vimentin, promoting its degradation.	Unknown	*Trans*	Negative (protein)
Non-regulatory	Protein scaffold	NEAT1 [131,132]	Necessary for the formation of nuclear paraspeckles.	Nuclear (paraspeckles)	NA	NA
	Protein scaffold	LINK-A [133]	Mediates the recruitment of BRK to GPNMB for kinase activation.	Cytoplasmic	NA	NA
	Protein modification	NKILA [134]	Binds to the NF-κB:IκB complex, inhibiting IκB phosphorylation.	Cytoplasmic	NA	NA
	Mimic/Decoy	GAS5 [55]	Mimics the glucocorticoid response element and binds to the glucocorticoid receptor.	Cytoplasmic and nuclear (nuclear in presence of dexamethasone)	NA	NA
	Protein decoy	NORAD [135]	Binds to PUM proteins, inhibiting their ability to repress target mRNAs.	Cytoplasmic	NA	NA
	Protein decoy	PANDA [64]	Binds to the transcription factor NF-YA to prevent it binding to chromatin.	Nuclear	NA	NA
	Protein localisation	SNHGS [136]	Binds to MTA2 protein, preventing its export to the cytoplasm.	Cytoplasmic	NA	NA

[a]Untranslated region.
[b]Members of the adenosine deaminase acting on RNA protein family.

Table 2. Resources for Discerning lncRNA Functional Properties

Name	Link	Description
Algorithms		
BLASTn [137]	http://blast.ncbi.nlm.nih.gov/Blast.cgi	Web server for local sequence similarity searching by nucleotide sequence.
ChEA [70]	http://amp.pharm.mssm.edu/lib/chea.jsp	Web server for transcription factor enrichment analysis.
ClusterProfiler [138]	http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html	R package for GO and KEGG enrichment analysis.
Coding Potential Calculator [139]	http://cpc.cbi.pku.edu.cn/	Web server and software for alignment-based coding potential calculations.
CPAT [140]	http://lilab.research.bcm.edu/cpat/	Web server and software for alignment-free coding potential calculations.
CummeRbund::JSdist	http://compbio.mit.edu/cummeRbund/	R function to calculating Jenson–Shannon divergence for condition specific expression.
Gene Ontology: GO enrichment analysis [141]	http://geneontology.org/page/go-enrichment-analysis	Web-based application for calculating functional enrichment of GO terms.
GSEA [37]	http://www.broadinstitute.org/gsea/index.jsp	Software for gene set enrichment analysis.
HOMER [69]	http://homer.salk.edu/homer/	Software for motif discovery and analysis.
LAST [82]	http://last.cbrc.jp/	Software for genome-scale nucleotide sequence similarity searching.
MEME [68]	http://meme-suite.org/index.html	Software for motif-based analyses.
QmRLFS-finder [86]	http://rloop.bii.a-star.edu.sg/?pg=qmrlfs-finder	Software for prediction of potential R-loops in DNA and RNA.
ReactomePA [142]	http://bioconductor.org/packages/release/bioc/html/ReactomePA.html	R package for Reactome enrichment analysis.
ROKU/TSGA [143]	http://www.cab.zju.edu.cn/ics/faculty/zhuj/software/tsga/index.htm	R package built upon ROKU algorithm for detecting tissue-specific expression.
Specond [144]	http://www.bioconductor.org/packages/release/bioc/html/SpeCond.html	R package for detecting condition-specific expression.
TargetScan [145]	http://www.targetscan.org/	Software and database for miRNA binding site predictions.
Triplexator/Triplex-Inspector [146,147]	http://bioinformatics.org.au/tools/triplexator/inspector/	Intermolecular triplex formation predictor.
WGCNA [33]	https://cran.r-project.org/web/packages/WGCNA/index.html	R package for WGCNA.
Data Sources		
European Nucleotide Archive (ENA)	http://www.ebi.ac.uk/ena	Repository for raw and processed sequencing data.
ENCODE [148]	https://www.encodeproject.org/	ENCODE consortium datasets including expression profiling (RNA-Seq) in multiple tissues, cell lines, and subcellular fractions; and DNA binding (ChIP-Seq) of transcription factors and epigenetic marks in multiple species.
Epigenome Roadmap [149]	http://www.roadmapepigenomics.org/data/	Epigenome roadmap datasets covering genome-wide epigenetic modifications and transcriptional profiles in humans.
FANTOM [150]	http://fantom.gsc.riken.jp	Expression data and functional annotation analyses of RNA from mammalian cells.

Table 2. (continued)

Name	Link	Description
GTEx [151]	http://gtexportal.org/home/	Expression data from a range of human tissues.
Sequence Read Archive (SRA)	http://www.ncbi.nlm.nih.gov/sra	Repository for raw sequencing data.
Databases		
BioXpress [152]	https://hive.biochemistry.gwu.edu/tools/bioxpress/	Database of coding and noncoding gene expression in cancers.
ChIPBase [153]	http://deepbase.sysu.edu.cn/chipbase/	Transcription factor binding evidence from ChIP-Seq data.
Cancer RNA-Seq Nexus [154]	http://syslab4.nchu.edu.tw/	lncRNA expression in multiple cancer RNA-Seq datasets.
Co-LncRNA [155]	http://www.bio-bigdata.com/Co-LncRNA/	lncRNA annotation by enrichment of GO and KEGG terms in coexpressed genes.
dbSNP [156]	http://www.ncbi.nlm.nih.gov/SNP/	SNPs in genomes.
DIANA-LncBase [157]	www.microrna.gr/LncBase	miRNA–lncRNA interactions from CLIP-Seq experiments and *in silico* predictions.
JASPAR [67]	http://jaspar.genereg.net/	Transcription factor binding sites.
lincSNP [158]	http://bioinfo.hrbmu.edu.cn/LincSNP	Disease-associated SNPs in human lincRNAs.
Lnc2Cancer [159]	http://www.bio-bigdata.net/lnc2cancer/	Cancer-associated lncRNAs.
LncRNA2Function [160]	http://mlg.hit.edu.cn/lncrna2function/	Prediction of lncRNA function from coexpression and enriched functional terms.
lncRNA2Target [161]	http://mlg.hit.edu.cn/lncrna2target/	Database of targets genes of lncRNAs, using evidence from lncRNA knockdown or overexpression.
LncRNADisease [47]	http://cmbi.bjmu.edu.cn/lncrnadisease	Experimentally supported lncRNA–disease associations.
LNCipedia [162]	http://www.lncipedia.org/db/search	Database of >100 000 human annotated lncRNAs. In addition to basic transcript annotation, also includes information on secondary structure, protein-coding potential, conservation, and miRNA targets.
lncRNASNP [163]	http://bioinfo.life.hust.edu.cn/lncRNASNP/	SNPs in human and mouse lncRNAs.
NHGRI-EBI GWAS Catalog	http://www.ebi.ac.uk/gwas/	Curated collection of published GWAS.
NPInter [76]	http://www.bioinfo.org/NPInter/	Physical interactions between noncoding RNAs and other biomolecules (proteins, DNA, and RNA).
ORegAnno [66]	http://www.oregano.org	Experimentally identified DNA regulatory regions, transcription factor binding sites, and regulatory variants.
Predicted RNA–RNA interactions [164]	http://rtools.cbrc.jp/cgi-bin/RNARNA/index.pl	Predicted lncRNA and mRNA interactions.
StarBase [165]	http://starbase.sysu.edu.cn/index.php	miRNA–RNA interactions from CLIP-Seq experiments and *in silico* predictions.
UCSC [166]	http://genome.ucsc.edu/	Genome browser with multiple annotation layers and associated tools, including genetic variants, conservation, regulation [66], chromatin states [43], and gene annotations.

Coding Potential

De novo assembly of transcripts is often performed before functional analyses of lncRNAs. Owing to their highly-specific and often variable expression patterns, annotations of lncRNAs are likely still incomplete in many biological systems [15,19,20]. Therefore, when RNA sequencing-based data is available *de novo* assembly can provide a more complete and relevant annotation. Coding potential of transcripts should be determined before downstream functional analyses to exclude the possibility that novel assembled transcripts may encode proteins (Table 2).

Differential Expression

Expression patterns of transcripts are crucial in uncovering functions of novel genes. With a growing application of RNA sequencing and microarray expression-profiling experiments comes a plethora of expression data that can be leveraged to specify more accurate hypotheses of lncRNA functions. The most common method of inferring lncRNAs function in a system is through differential expression analysis. While this widely utilised and generally accepted method is adept at prioritising candidates for further examination, differential expression alone does not typically produce any functional insights. Alternative methods, such as guilt-by-association (see below), take advantage of the general characteristics of lncRNAs by exploiting other biological contexts.

Guilt-by-Association

Guilt-by-association is predicated on the idea that coexpressed transcripts are more likely to be coregulated, share similar functions, or are involved in similar biological processes [21]. Guilt-by-association, as the name suggests, assigns putative functions to transcripts based on those it is coexpressed with. Unlike differential expression, expression patterns from multiple related biological conditions can be used, enabling the identification of distinct relationships between transcripts. The confidence of the association is dependent on the number of conditions for which expression data are available. Time-dependent data can be particularly valuable because dynamic regulation of expression can be informative of the particular pathways in which a lncRNA functions [22]. The popularity and utility of this type of approach have given rise to several subtypes of analyses – including the use of ***cis*-regulatory** (see Glossary) relationships, whole-transcriptome correlations with a noncoding or coding candidate, and clustering- and network-based approaches.

Genome-wide clustering of expression profiles can be utilised to identify groups of transcripts that are coregulated and show **functional enrichment** for a process. lncRNAs in such groups are more likely to be involved in, or regulate, particular biological pathways associated with such groups [21]. Multiple methods for clustering exist, each with their own advantages and disadvantages. Three commonly used methods are hierarchical clustering [23], *k*-means clustering [18], and self-organising maps (SOMs) [24]. These methods all require some form of selection of cluster number or size, and therefore care should be taken in selecting and supporting these parameters to ensure that the clusters obtained as a result are informative.

Given the complexity of lncRNA function, several network-based approaches have been used to decipher them [25–27]. Similarly to clustering, modules produced by networks can be used to assign functional associations to lncRNA. However, because networking methods can utilise multiple layers of information in addition to expression patterns, more complex relationships can be discovered [28–32]. Network construction can be computationally difficult, and many algorithms and statistical approaches exist for this purpose. A more simplistic approach is to use weighted gene coexpression analysis (WGCNA) [33]; however, this method is limited to gene expression alone.

Groups of transcripts identified by clustering should be subjected to a functional enrichment step to interpret the biological processes these genes are involved in. Potential pathway and

functional enrichments can be determined using data from GO [34], KEGG [35], or Reactome [36]. Several web-based tools and software packages are available which perform statistical tests to determine if the enrichment of particular terms or pathways in a group of genes is greater than expected, and how significant this is (Table 2). Similarly, gene-set enrichment analysis (GSEA) [37] can be used to evaluate the enrichment of particular processes using a ranked list of genes. It should be noted that predictions based on these enrichments are only as reliable as the underlying annotations, and care should therefore be taken to ensure that an up-to-date database is used.

Regulatory lncRNAs can act in both *cis* and *trans*, influencing or interacting with nearby or distant genes. Several lncRNAs located nearby, overlapping, or within protein-coding genes can positively or negatively regulate the expression of these genes in *cis*. *Cis*-relationships can foreseeably arise through complementary sequence motifs, tethering, blocking, and **product-independent transcription** [38–40]. **Enhancer RNAs** (eRNAs) in particular can act in *cis* to positively regulate gene expression [41]. Therefore, the genomic relationship with other genes can be a useful initial filtering step to search for potential regulatory relationships. In the case of transcripts that function in *cis*, genomic location can be used as a guide for guilt-by-association analyses. Significant positive correlations may indicate positive regulatory relationships [41]. However, this correlation can foreseeably arise from the similar epigenetic profiles of proximal genes [42]. In this case, separation of genomic loci by insulators [16] or differing chromatin states [43] can be used as additional evidence to support positive *cis*-regulatory relationships. More confident predictions can arise from negative correlations because these would not be due to similar epigenetic states. Additional evidence of a relationship may be beneficial to improve the confidence of regulatory predictions.

To identify novel transcripts involved in similar processes to a known gene of interest, expression correlations can be used. In this case, a correlation matrix consisting of all transcripts versus all transcripts is produced. Any transcripts of unknown function that are significantly positively or negatively correlated can then be identified. This method can also be reversed – using an uncharacterised lncRNA to find a subset of characterised transcripts with correlated expression.

Condition-Specific Expression

lncRNAs appear to be under exquisite transcriptional control with specialised functions and therefore typically tend to be more lowly and specifically expressed than protein-coding transcripts [3,18]. Numerous lncRNAs show specific temporal and spatial expression patterns, which can direct us towards the biological context in which they are acting [44]. Multiple algorithms are available for the detection of condition-specific expression which can be used in place of differential expression testing in a larger number of conditions (Table 2). In this case, expression data from a broad range of conditions – such as tissue or cell types – can be interrogated. Although the simple method of counting the conditions for which a gene is expressed above a particular level can be used, more sophisticated approaches are available (Table 2). Compared to protein-coding genes, transcribed lncRNAs tend to have higher expression variability within the same condition, which can complicate annotation [19,45]. Low variability may be used as a potential indicator of transcript function in normal cell functions, whereas high variability may indicate environment- and disease-related function.

Disease Associations

Given the characteristic specificity of lncRNA expression, there is growing interest in the use of these molecules as disease biomarkers [46]. Disease associations of lncRNAs can be derived by a differential expression or condition-specific analysis of the disease/normal status. Hundreds of experimentally supported lncRNA–disease associations have been collected and curated within the lncRNADisease database [47]. SNPs identified by genome-wide association studies (GWAS)

and other genomic variations contained within or nearby lncRNAs can also point to functional roles in specific phenotypes. Locations of SNPs can be found using dbSNP or the variation track of UCSC Genome Browser (Table 2). Those that have been implicated with various phenotypes have been documented in multiple GWAS databases (Table 2). The alteration of activity of functional lncRNAs by SNPs has been observed for multiple lncRNAs [48–51], and has driven the creation of databases cataloguing lncRNA SNPs in human and mouse (Table 2). While these approaches may reveal biomarkers for the disease/condition of interest, it is important to note that expression alone cannot be taken as sufficient evidence of functionality. Several lncRNAs – such as ANRIL [52], HOTAIR [53], and GAS5 [54,55] – influence cancer-related phenotypes such as cell proliferation, apoptosis, and metastasis. However, diseases such as cancers can be highly epigenetically altered, and non-functional pervasive transcription may occur as a byproduct [56]. Therefore, additional supporting evidence for functionality should be sought.

Conservation

Conservation can be a powerful tool to elucidate the functional importance of particular sequences. The primary sequences of lncRNA exons, although less conserved than that of protein-coding genes, are more conserved than random intergenic and intronic regions. As lncRNAs are not under the same evolutionary constraints as protein-coding genes [57,58], owing to their entirely different structure–function relationship, use of primary sequence conservation scores (Table 2) to lncRNAs under the same parameters as protein-coding genes should be approached with caution. At the level of secondary structure, conservation can also point towards evolutionary importance [59]. However, distinct structure–function relationships are yet to be elucidated (Box 2). Orthologs of lncRNAs can be identified using an alignment-guided synteny approach. In this method, BLAST or genome-wide pairwise alignments from UCSC are used to identify short segments of sequences that align in two species, show evidence of transcription in both, and share the same flanking genes [57,58,60]. Evidence of conserved transcription implies that the sequence is functional, and likely acts through a product-dependent mechanism. Examination of the functional properties of orthologous transcripts can strengthen hypotheses of functional mechanisms, particularly when more data are available for these analyses in other species.

Cellular Localisation

lncRNAs can be subdivided according to the cellular compartment in which they exert their function [6,61]. Nuclear-localised RNAs are more likely to have roles in transcriptional regulation, post-transcriptional regulation of mRNAs before nuclear export, or as components of

Box 2. Functional Structural Motifs in lncRNAs

Similar to proteins, structure is likely a key functional property of lncRNAs. Intricate structures formed through intramolecular RNA binding enable the interaction with proteins and other cellular components. These are likely to be modular, with distinct 2D/3D and nucleic acid binding domains joining together targeted proteins and DNA/RNA, respectively. Several RNA structural prediction tools are available for application to lncRNA sequences (reviewed in [93,94]). However, beyond their use to identify structurally conserved regions, distinct structural domains and subdomains at present are not informative of functional mechanisms.

The disconnect between structure and function in lncRNAs arises from the current lack of both experimentally defined structures and accurate cataloguing of lncRNAs interacting with specific proteins. Structures have been experimentally characterised for a select number of well-studied lncRNAs, including most recently HOTAIR [95] and XIST [96]. These investigations confirmed that multiple distinct domains form, with specific domains being required for specific interactions. However, it remains unclear as to whether encoded interactions arise from the formation of subdomains – such as stem-loops – or from domains comprised of several structured subdomains. Several lncRNAs share mechanisms by interacting with the same protein, as is the case in epigenetic regulation through recruitment of the PRC2 complex. Binding assays have identified several lncRNAs that are specifically and non-specifically associated with PRC2 [97]. This so-called promiscuous binding – resulting in numerous lncRNAs being identified as interactors – generates noise, complicating studies of shared structures responsible for specific binding.

nuclear-localised structural complexes (Table 1). Cytoplasmic-localised RNAs with known function have roles in modifying mRNA stability and affecting translation. Several expression-profiling experiments have been carried out on different components of cells, usually chromatin, the nucleus, and the cytoplasm (Table 2). Localisation may help to determine the regulatory stage at which lncRNAs are functioning. Evidence of predominant expression in a particular cellular location can be used to strengthen hypotheses of mechanisms of action. However, these localisations are not always exclusive, and can be highly context- and target-dependent. Indeed, non-specific localisation cannot be considered sufficient evidence for non-functionality or product-independent functions. Owing to the apparent specificity of lncRNA expression, localisation may be dependent on function [62]. In this case, the biological contexts in which these experiments are performed should be considered when inferring function.

Epigenetic Status

Intergenic lncRNAs can be divided into promoter and enhancer lncRNAs by using chromatin marks. **Promoter lncRNAs** (plncRNAs), similar to protein-coding genes, show enrichment of H3K4me3 at the promoter and H3K36me3 along the transcribed region. Enhancer RNAs (elncRNAs/eRNAs) show enrichment of H3K4me1 over H3K4me3 within their promoter region. These two subclasses are likely to have distinct functions, with elncRNAs displaying a significantly higher correlation with their protein-coding neighbours than plncRNAs [63]. This suggests that using evidence of enhancer-like properties – either through interrogation of biologically-relevant histone modification ChIP-Seq data or use of the Broad ChromHMM annotations [43] (Table 2) – in combination with a positive correlation in expression can be used to infer enhancer-like function of lncRNAs.

lncRNA Locus–Protein Interactions

Interactions with transcription factors as cofactors or guides can increase the expression of target genes, whereas interactions as decoys can reduce expression. Promoter regions of lncRNAs may hold clues to their regulation. The shared use of a promoter region can point towards common regulatory mechanisms of expression [64]. Transcription factor binding site (TFBS) evidence or prediction can be used in the promoter regions of lncRNAs to find potential regulators of their expression, and therefore the regulatory networks in which they may be involved [65]. Experimental evidence of transcription factor binding to DNA can be determined by ChIP-Seq or may be found in the ORegAnno [66] database. Because lncRNA transcription is known to be highly spatiotemporally specific, evidence of binding may not be apparent in such experiments. TFBS prediction can supplement experimental evidence by evaluating possible interactions; however, these can have high false-positive rates. JASPAR provides a curated, up-to-date database with predefined transcription factor binding motifs/position weight matrices calculated from experimental evidence [67]. In addition, tools such as MEME [68], HOMER [69] and CHEA [70] can perform motif-enrichment analyses on sets of lncRNA promoters that may be coregulated. Transcription of lncRNA from canonical RNA polymerase II promoters implies that – like protein-coding genes – the product is functional [17]. Evidence of a canonical promoter through chromatin marks (Table 2) can therefore be used, similarly to conservation, as additional evidence of functionality.

lncRNA–Protein Interactions

Identification of proteins that interact with lncRNAs provides another opportunity to predict function or regulatory relationship. XIST, one of the best-studied lncRNAs, functions partially through binding the PRC2 complex to deposit repressive chromatin marks along the inactive X-chromosome in mammals [7]. Several transcripts identified more recently share this or a functionally similar mechanism, consisting of targeted recruitment of proteins to epigenetically regulate expression [71–73]. Post-transcriptional regulation can also be mediated through interactions with proteins to regulate stability [74] or modify RNA [75]. Evidence of binding to these

proteins can be found in the database Npinter [76], which uses high-throughput data and text mining as experimental evidence of interactions. **Machine-learning** methodologies [77–79] are yielding increasingly promising results for prediction of potential protein binding, and have been successfully applied in a range of genomic problems (reviewed in [80]). As a more complex parallel to TFBS prediction, it is foreseeable that such predictive tools will be applicable in the future.

Nucleic Acid Binding

The ability to bind to other nucleic acids in a specific manner is a core component of the functionality of regulatory lncRNAs. Evidence for potential intermolecular binding of lncRNAs can be used to identify target DNA or RNA sequences and strengthen evidence for regulatory relationships. In addition, specific interactions can provide a hypothesis for the mechanism of action of regulatory lncRNAs. Formation of RNA:RNA [81] duplexes allows the targeting of specific RNA molecules, potentially affecting RNA splicing, editing, stability, translation, or localisation (Table 1). Potential binding capability between a lncRNA and a target can be found using alignment methods such as BLASTn and LAST [82]. In cases where a lncRNA does not contain an overlapping antisense sequence, regions with high homology as a result of duplication or repeat content may be responsible for targeted binding [81]. Post-transcriptional regulatory function by binding miRNAs as a decoy also utilises RNA:RNA binding (Table 1). lncRNA:miRNA binding potential can be assessed through the use of miRNA target-site prediction algorithms (Table 2), though the high false-positive rates of such algorithms should be taken into consideration.

Complex structures can form through binding to DNA or RNA, such as **triplexes** [71,83], **G-quadruplexes** (GQs), and **R-loops**. In addition to potentially recruiting complexes to a target sequence, the formation of these structures can intrinsically regulate expression and translation [84]. Intermolecular triplex formation between lncRNA and a target can be predicted by the triplexator algorithm (Table 2). Current GQ structure prediction tools can only predict intramolecular interactions, highlighting the need for a tool that can identify intermolecular interactions. Experimental evidence for R-loop formation can be obtained from DRIP-Seq datasets. In addition, usage of R-loop prediction algorithms can identify potential formation sites [85,86].

Integrative Approaches for Functional Candidate Selection

Each of the approaches outlined above can extract potential candidates for characterisation from an initial pool of lncRNAs. However, isolated use produces insufficient evidence to generate detailed hypotheses of function. This is typical with differential expression – where a multitude of lncRNA can be differentially expressed between conditions – but this change tells us little about whether the expression pattern is functional, and if so, how. Instead, integrative approaches that test separate avenues of lncRNA functionality can produce a more refined candidate list, with some insight as to how they are exerting their function (Figure 1).

Initial experiments typically focus on a biological condition of interest through differential or condition-specific expression, which can produce a large pool of potential candidate lncRNAs for further study. Additional evidence for biological functionality can be gleaned from clustering- and network-based guilt-by-association approaches, followed by enrichment for biological terms relating to the system, as well as and through evidence of binding sites of key TFs. Biological importance can be determined through additional analyses, including conservation and/or potential disease association. Not all functional lncRNAs will show evidence through current conservation-based methods, and thus this avenue is less crucial when selecting candidates. Supplementing evidence of biological importance in a selected context by probing several mechanistic features can further filter down an initial pool to those that show evidence of function. Well-developed computational methods are currently biased towards the characterisation of regulatory lncRNAs (see Outstanding Questions) and, as such, integrative approaches are most effective for these transcripts. Archetypical regulatory lncRNAs exhibit highly correlated expression

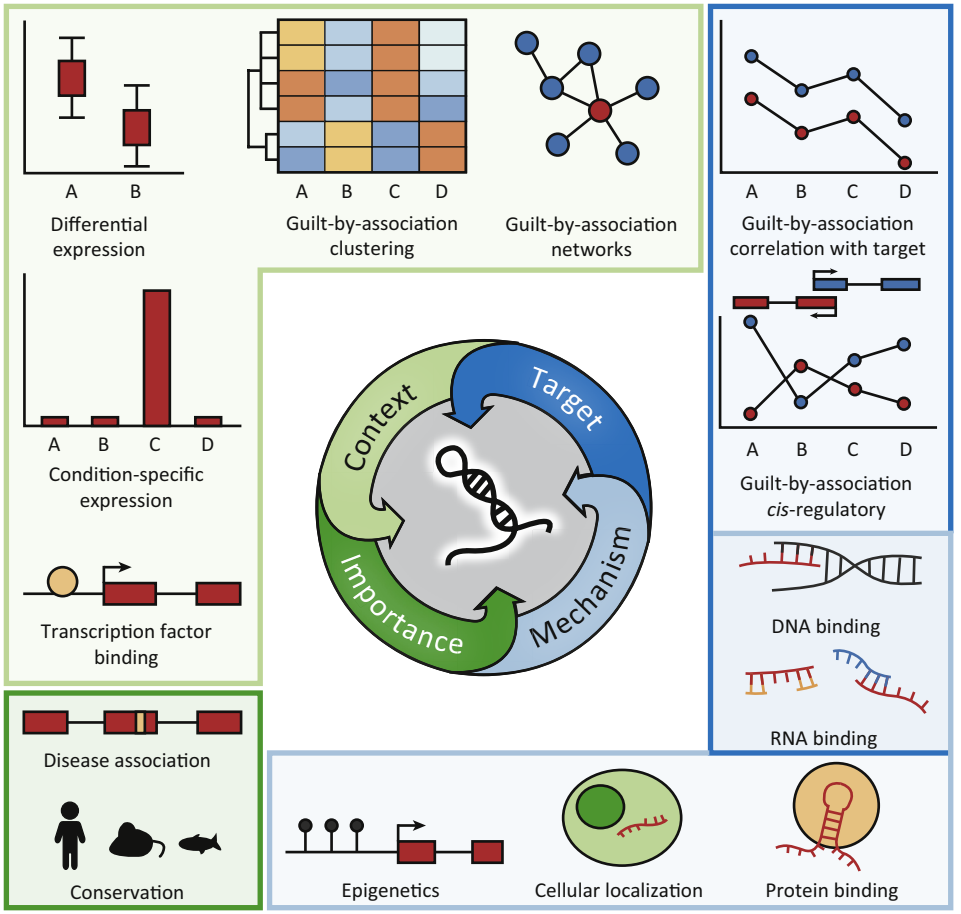

Figure 1. Summary of the Methods Presented. Each of the methods described in the main text probes a component of functional lncRNA biology. Those informative of the biological context are highlighted in light green, biological importance in dark green, regulatory target in dark blue, and functional mechanism in light blue. Using methods from each of these allows candidate lncRNA(s) to be identified, and further experiments targeting unknown biology to be designed.

profiles with that of their targets, evidenced by correlation through guilt-by-association. Although this can identify potential targets, further evidence of a direct relationship (particularly in the case of positive correlations) can be used to distinguish this from genes under the same regulatory control and incidental correlations. Mechanistic insights can be gained through epigenetic status, DNA- and RNA-binding capability, to identify eRNAs, transcriptional regulators, and post-transcriptional regulators, respectively. DNA and RNA binding can also be used as additional evidence of a functional relationship between lncRNAs and potential targets. Cellular localisation can further stratify candidates into functional subclasses – such as post-transcriptional regulators acting following nuclear export of targets.

By layering information from incorporation of publicly-available datasets, databases, and specialised computational tools, additional evidence of a functional link between a lncRNA and a target and mechanistic insights can be found to strengthen functional hypotheses. When such information is gleaned from initial analyses, it should be compared with expected features of mechanistic classes of characterised lncRNAs (Table 1). In doing so, the design of experiments can be informed by the association of a candidate lncRNA to a particular functional subclass.

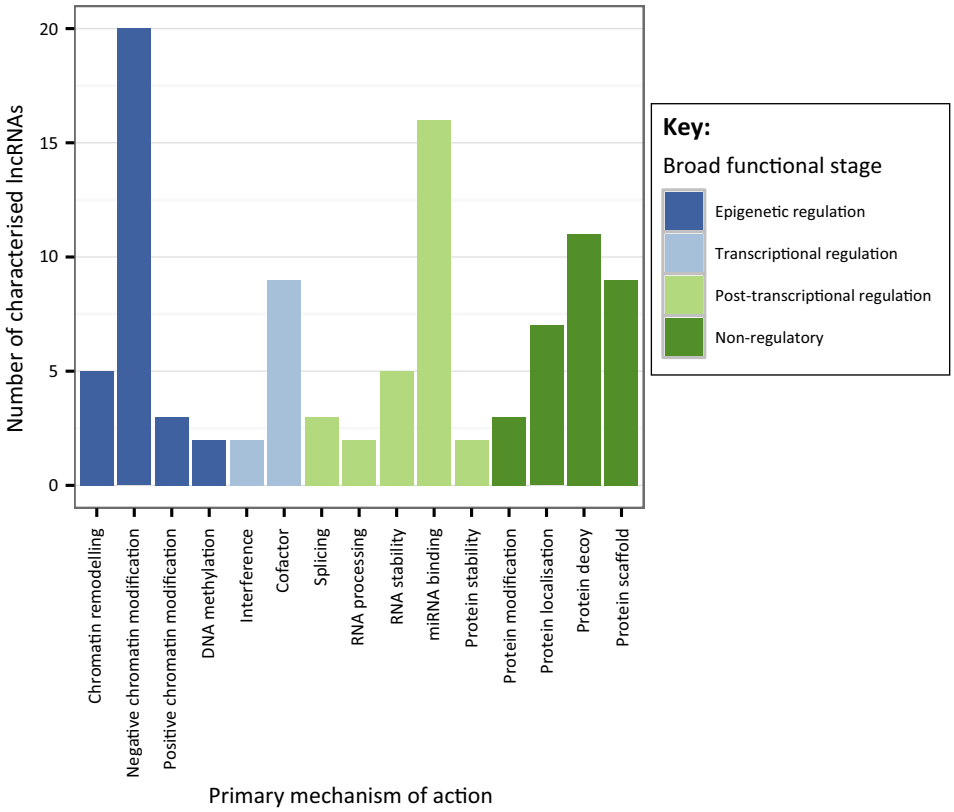

Figure 2. Experimentally Characterised Functional Mechanisms of lncRNAs Catalogued in lncRNAdb [5]. Examples of each subclass are detailed in Table 1.

Ilott *et al.* [87] demonstrate the use of integrating analysis of their initial expression data with publicly-available datasets and using computational analyses to identify candidate lncRNAs. RNA-Seq before and after lipopolysaccharide stimulation in human monocytes identified hundreds of lncRNAs and regions of bidirectional transcription that were differentially expressed. Several of these were then identified as potential regulators of monocytic inflammatory response genes in *cis* by using neighbouring gene expression correlations. Integration with ENCODE ChIP-Seq data enabled the identification of enhancer-like lncRNAs, providing further evidence that these may be transcriptional regulators. In addition, the ChIP-Seq binding profile of NF-κB – an inflammatory TF – showed that these lncRNAs are enriched in NF-κB binding sites, and are therefore potentially regulated in this manner. The combined usage of these techniques led to the selection of candidate lncRNAs surrounding IL-1β, which were then shown to have a regulatory role.

Testing lncRNA Functionality

Computational approaches provide a way to test multiple avenues of functionality on a large cohort of potentially important transcripts. Because of their relative ease in reuse on multiple candidates, computational analyses have been extensively used in identifying lncRNAs on a genome-wide scale [15,17,18,45]. While these approaches can provide insight into the general trends of lncRNA biology, the known specificity of lncRNA expression and function require that experimental methods (reviewed in [88,89]) are used to validate the functionality of lncRNAs in the biological system being tested.

To aid experimental characterisation, candidate lncRNAs should be selected on a basis of degree of evidence. Those with evidence of functionality covering multiple categories (Figure 1) and/or a refined mechanism (e.g., evidence of target RNA/DNA binding and specific protein binding) should be prioritized at this stage. For example, Bergmann *et al.* [90] used predominant nuclear expression and a high level of connectedness in OCT4 expression networks to identify potential regulatory lncRNAs involved in embryonic stem cell (ESC) maintenance. Understanding potential mechanisms allows these to be validated in a more targeted manner, and further refined in cases of incomplete evidence. If evidence points towards particular interactions (based on expression patterns), these should be verified through proximity or binding assays, particularly if evidence comes from predictions or a different cellular context.

Typically, the production of a distinct cellular phenotype by lncRNA loss- or gain-of-function experiments has been used as sufficient evidence of functionality [91]. However, effects such as product-independent transcription can complicate interpretation, and a strong phenotype may not be observable [92]. Instead, experimental characterisation may need to be more targeted with knowledge of potential mechanisms in mind. Once potential mechanism has been investigated, loss-of-function approaches can be employed to determine if these mechanisms and the effects resulting from them are altered. For example, in the case of a putative chromatin-remodelling lncRNA, experiments should be designed to assess chromatin looping as well as gene expression following lncRNA expression changes.

Outstanding Questions

Can product-dependent (driven by the transcript) and product-independent (driven by transcription) functions be distinguished?

Is apparent non-specific transcription truly biological 'noise', or are there aspects to transcription-guided regulatory control that remain to be elucidated?

The majority of currently characterised lncRNAs appear to regulate expression of transcripts. Has this arisen through extensive use of guilt-by-association, and other expression-based methods? Or is this reflective of a greater trend in lncRNA functions?

Structural motifs appear to have important roles in lncRNA biology. Do repetitive elements also have a similar role? How can these be linked to specific functions?

Concluding Remarks and Future Directions

Within the past decade, lncRNAs have emerged as important RNA species, capable of fulfilling previously unascertained biological roles. Through the increasing application of high-throughput sequencing methods, lncRNAs continue to be discovered. However, the gap between identified and functionally characterised molecules remains considerably larger than that of protein-coding genes. By utilising a spectrum of freely-available software and publicly-available datasets, it is possible to investigate and interpret functionality *in silico* and narrow the functional search space for many lncRNAs. Systematic application of these techniques is a fundamental step in the functional characterisation of lncRNAs, and is crucial for informing further specific experiments to characterise any individual or given set of lncRNAs.

Acknowledgements

The authors would like to thank Daniel Thompson for helpful discussion on earlier drafts of this manuscript.

References

1. Kapranov, P. *et al.* (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* 15, 987–997
2. Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563
3. Derrien, T. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789
4. Zhao, Y. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* 44, D203–D208
5. Quek, X.C. *et al.* (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* 43, D168–D173
6. Mercer, T.R. and Mattick, J.S. (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* 20, 300–307
7. Brockdorff, N. (2002) X-chromosome inactivation: closing in on proteins that bind Xist RNA. *Trends Genet.* 18, 352–358
8. Gabory, A. *et al.* (2010) The H19 locus: role of an imprinted non-coding RNA in growth and development. *Bioessays* 32, 473–480
9. Rinn, J.L. *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323
10. Ma, L. *et al.* (2013) On the classification of long non-coding RNAs. *RNA Biol.* 10, 924–933
11. Laurent, G.S. *et al.* (2015) The landscape of long noncoding RNA classification. *Trends Genet.* 31, 239–251
12. Zhang, Y. *et al.* (2016) Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat. Struct. Mol. Biol.* 23, 522–530
13. Chakraborty, D. *et al.* (2012) Combined RNAi and localization for functionally dissecting long noncoding RNAs. *Nat Methods.* 9, 360–362
14. Guttman, M. *et al.* (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477, 295–300
15. Iyer, M.K. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet.* 47, 199–208
16. Laurent, G.S. *et al.* (2016) Functional annotation of the vlinc class of non-coding RNAs using systems biology approach. *Nucleic Acids Res.* 44, 3233–3252

17. St Laurent, G. *et al.* (2013) VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol.* 14, R73
18. Cabili, M.N. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927
19. Kornienko, A.E. *et al.* (2016) Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol.* 17, 14
20. Bussotti, G. *et al.* (2016) Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Res.* 26, 705–716
21. Stuart, J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255
22. Dinger, M.E. *et al.* (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* 18, 1433–1445
23. Ramos, A.D. *et al.* (2013) Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. *Cell Stem Cell* 12, 616–628
24. Kim, D.H. *et al.* (2015) Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. *Cell Stem Cell* 16, 88–101
25. Xiao, Y. *et al.* (2015) Predicting the functions of long noncoding RNAs using RNA-seq based on Bayesian network. *Biomed Res. Int.* 2015, 839590
26. Zhou, M. *et al.* (2015) Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. *Mol. Biosyst.* 11, 760–769
27. Yao, P. *et al.* (2015) Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nat. Neurosci.* 18, 1168–1174
28. Guo, X. *et al.* (2013) Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res.* 41, e35
29. Liao, Q. *et al.* (2011) Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic Acids Res.* 39, 3864–3878
30. Sun, J. *et al.* (2014) Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network. *Mol Biosyst.* 10, 2074–2081
31. Yang, X. *et al.* (2014) A network based method for analysis of lncRNA–disease associations and prediction of lncRNAs implicated in diseases. *PLoS One* 9, e87797
32. Chen, X. *et al.* (2015) Constructing lncRNA functional similarity network based on lncRNA–disease associations and disease semantic similarity. *Sci. Rep.* 5, 11338
33. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559
34. Ashburner, M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29
35. Kanehisa, M. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205
36. Milacic, M. *et al.* (2012) Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers* 4, 1180–1211
37. Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550
38. Lee, J.T. (2009) Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev.* 23, 1831–1842
39. Guil, S. and Esteller, M. (2012) Cis-acting noncoding RNAs: friends and foes. *Nat. Struct. Mol. Biol.* 19, 1068–1075
40. Kornienko, A.E. *et al.* (2013) Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.* 11, 59
41. Orom, U.A. *et al.* (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143, 46–58
42. Ebisuya, M. *et al.* (2008) Ripples from neighbouring transcription. *Nat. Cell Biol.* 10, 1106–1113
43. Ernst, J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49
44. Mercer, T.R. *et al.* (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U.S.A.* 105, 716–721
45. St Laurent, G. *et al.* (2012) Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics* 13, 504
46. Du, Z. *et al.* (2013) Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* 20, 908–913
47. Chen, G. *et al.* (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986
48. Jendrzejewski, J. *et al.* (2012) The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8646–8651
49. Kumar, V. *et al.* (2013) Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.* 9, e1003201
50. Zhang, X. *et al.* (2014) The identification of an ESCC susceptibility SNP rs920778 that regulates the expression of lncRNA HOTAIR via a novel intronic enhancer. *Carcinogenesis* 35, 2062–2067
51. van Dijk, M. *et al.* (2015) Mutations within the LINC-HELLP non-coding RNA differentially bind ribosomal and RNA splicing complexes and negatively affect trophoblast differentiation. *Hum. Mol. Genet.* 24, 5475–5485
52. Yu, W. *et al.* (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 451, 202–206
53. Gupta, R.A. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076
54. Mourtada-Maarabouni, M. *et al.* (2009) GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* 28, 195–208
55. Kino, T. *et al.* (2010) Noncoding RNA Gas5 is a growth arrest and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal.* 3, ra8
56. Jones, P.A. and Baylin, S.B. (2007) The epigenomics of cancer. *Cell* 128, 683–692
57. Hezroni, H. *et al.* (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11, 1110–1122
58. Necsulea, A. *et al.* (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640
59. Smith, M.A. *et al.* (2013) Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* 41, 8220–8236
60. Washietl, S. *et al.* (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* 24, 616–628
61. Zhang, K. *et al.* (2014) The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene* 547, 1–9
62. Carrieri, C. *et al.* (2012) Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* 491, 454–457
63. Marques, A.C. *et al.* (2013) Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol.* 14, R131
64. Hung, T. *et al.* (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat. Genet.* 43, 621–629
65. Mohamed, J.S. *et al.* (2010) Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA* 16, 324–337
66. Lesurf, R. *et al.* (2016) ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.* 44, D126–D132

67. Mathelier, A. *et al.* (2015) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* gkv1176
68. Bailey, T.L. *et al.* (2015) The MEME Suite. *Nucleic Acids Res.* 43, W39–W49
69. Heinz, S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589
70. Lachmann, A. *et al.* (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 26, 2438–2444
71. O'Leary, V.B. *et al.* (2015) PARTICLE, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation. *Cell Rep.* 11, 474–485
72. Liu, G.Y. *et al.* (2016) The long noncoding RNA Gm15055 represses Hoxa gene expression by recruiting PRC2 to the gene cluster. *Nucleic Acids Res.* 44, 2613–2627
73. Huang, M.D. *et al.* (2015) Long non-coding RNA TUG1 is up-regulated in hepatocellular carcinoma and promotes cell growth and apoptosis by epigenetically silencing of KLF2. *Molecular cancer* 14, 165
74. Kretz, M. *et al.* (2013) Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 493, 231–235
75. Salameh, A. *et al.* (2015) PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8403–8408
76. Hao, Y. *et al.* (2016) NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. *Database* 2016, baw057
77. Glazko, G.V. *et al.* (2012) Computational prediction of polycomb-associated long non-coding RNAs. *PLoS One* 7, e44878
78. Ge, M. *et al.* (2016) A bipartite network-based method for prediction of long non-coding RNA–protein interactions. *Genomics Proteomics Bioinformatics* 14, 62–71
79. Muppirala, U.K. *et al.* (2011) Predicting RNA–protein interactions using only sequence information. *BMC Bioinformatics* 12, 1
80. Libbrecht, M.W. and Noble, W.S. (2015) Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332
81. Gong, C. and Maquat, L.E. (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* 470, 284–288
82. Kiełbasa, S.M. *et al.* (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493
83. Mondal, T. *et al.* (2015) MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA–DNA triplex structures. *Nat. Commun.* 6, 7743
84. Buske, F.A. *et al.* (2011) Potential in vivo roles of nucleic acid triple-helices. *RNA Biol.* 8, 427–439
85. Wongsurawat, T. *et al.* (2012) Quantitative model of R-loop forming structures reveals a novel level of RNA–DNA interactome complexity. *Nucleic Acids Res.* 40, e16-e16
86. Jenjaroenpun, P. *et al.* (2015) QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Res.* gkv344
87. Ilott, N.E. *et al.* (2014) Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nat. Commun.* 5, 3979
88. Yan, B. *et al.* (2012) The research strategies for probing the function of long noncoding RNAs. *Genomics* 99, 76–80
89. Kashi, K. *et al.* (2016) Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochim. Biophys. Acta* 1859, 3–15
90. Bergmann, J.H. *et al.* (2015) Regulation of the ESC transcriptome by nuclear long noncoding RNAs. *Genome Res.* 25, 1336–1346
91. Mudge, J.M. *et al.* (2013) Functional transcriptomics in the post-ENCODE era. *Genome Res.* 23, 1961–1973
92. Bassett, A.R. *et al.* (2014) Considerations when investigating lncRNA function in vivo. *Elife* 3, e03058
93. Yan, K. *et al.* (2016) Structure prediction: new insights into decrypting long noncoding RNAs. *Int. J. Mol. Sci.* 17, 132
94. Blythe, A.J. *et al.* (2016) The ins and outs of lncRNA structure: how, why and what comes next? *Biochim. Biophys. Acta* 1859, 46–58
95. Somarowthu, S. *et al.* (2015) HOTAIR forms an intricate and modular secondary structure. *Mol. Cell* 58, 353–361
96. Lu, Z. *et al.* (2016) RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* 165, 1267–1279
97. Davidovich, C. *et al.* (2015) Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. *Mol. Cell* 57, 552–558
98. Hirota, K. *et al.* (2008) Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* 456, 130–134
99. Wang, Y. *et al.* (2015) The long noncoding RNA lncTCF7 promotes self-renewal of human liver cancer stem cells through activation of Wnt signaling. *Cell Stem Cell* 16, 413–425
100. Boque-Sastre, R. *et al.* (2015) Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proc. Natl. Acad. Sci. U.S.A.* 112, 5785–5790
101. Yin, Y. *et al.* (2015) Opposing roles for the lncRNA Haunt and its genomic locus in regulating HOXA gene activation during embryonic stem cell differentiation. *Cell Stem Cell* 16, 504–516
102. Pnueli, L. *et al.* (2015) RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin alpha-subunit gene. *Proc. Natl. Acad. Sci. U.S.A.* 112, 4369–4374
103. Scarola, M. *et al.* (2015) Epigenetic silencing of Oct4 by a complex containing SUV39H1 and Oct4 pseudogene lncRNA. *Nat. Commun.* 6, 7631
104. Wang, K.C. *et al.* (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120–124
105. Tsai, M.-C. *et al.* (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689–693
106. Khalil, A.M. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11667–11672
107. Csorba, T. *et al.* (2014) Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. *Proc. Natl. Acad. Sci. U.S.A.* 111, 16160–16165
108. Klattenhoff, C.A. *et al.* (2013) Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* 152, 570–583
109. Chalei, V. *et al.* (2014) The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. *Elife* 3, e04530
110. Wang, L. *et al.* (2015) LncRNA Dum interacts with Dnmts to regulate Dppa2 expression during myogenic differentiation and muscle regeneration. *Cell Rep.* 25, 335–350
111. Arab, K. *et al.* (2014) Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A. *Mol. Cell* 55, 604–614
112. Manelyte, L. *et al.* (2014) Chromatin targeting signals, nucleosome positioning mechanism and non-coding RNA-mediated regulation of the chromatin remodeling complex NoRC. *PLoS Genet.* 10, e1004157
113. Schmitz, K.M. *et al.* (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.* 24, 2264–2269
114. Pandey, R.R. *et al.* (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* 32, 232–246
115. Mohammad, F. *et al.* (2010) Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development* 137, 2493–2499
116. Latos, P.A. *et al.* (2012) Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* 338, 1469–1472
117. Jiang, W. *et al.* (2015) The lncRNA DEANR1 facilitates human endoderm differentiation by activating FOXA2 expression. *Cell Rep.* 11, 137–148

118. Beltran, M. *et al.* (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial–mesenchymal transition. *Genes Dev.* 22, 756–769
119. Wang, G.-Q. *et al.* (2016) Sirt1 AS lncRNA interacts with its mRNA to inhibit muscle formation by attenuating function of miR-34a. *Sci. Rep.* 6, 21865
120. Yuan, J.-H. *et al.* (2014) A long noncoding RNA activated by TGF-β promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell* 25, 666–681
121. Wang, Y. *et al.* (2013) Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev. Cell* 25, 69–80
122. Memczak, S. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338
123. Hansen, T.B. *et al.* (2013) Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388
124. Hansen, T.B. *et al.* (2011) miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J.* 30, 4414–4422
125. Yanagida, S. *et al.* (2013) ASBEL, an ANA/BTG3 antisense transcript required for tumorigenicity of ovarian carcinoma. *Sci. Rep.* 3, 1305
126. Pang, W.J. *et al.* (2013) Knockdown of PU.1 AS lncRNA inhibits adipogenesis through enhancing PU.1 mRNA translation. *J. Cell Biochem.* 114, 2500–2512
127. Wei, N. *et al.* (2014) Knockdown of PU.1 mRNA and AS lncRNA regulates expression of immune-related genes in zebrafish *Danio rerio*. *Dev. Comp. Immunol.* 44, 315–319
128. Wei, N. *et al.* (2015) PU.1 antisense lncRNA against its mRNA translation promotes adipogenesis in porcine preadipocytes. *Anim. Genet.* 46, 133–140
129. Yoon, J.-H. *et al.* (2012) LincRNA-p21 suppresses target mRNA translation. *Mol. Cell* 47, 648–655
130. Wang, T.H. *et al.* (2015) Long non-coding RNA AOC4P suppresses hepatocellular carcinoma metastasis by enhancing vimentin degradation and inhibiting epithelial–mesenchymal transition. *Oncotarget* 6, 23342–23357
131. Clemson, C.M. *et al.* (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell* 33, 717–726
132. Sunwoo, H. *et al.* (2009) MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* 19, 347–359
133. Lin, A. *et al.* (2016) The LINK-A lncRNA activates normoxic HIF1alpha signalling in triple-negative breast cancer. *Nat. Cell Biol.* 18, 213–224
134. Liu, B. *et al.* (2015) A cytoplasmic NF-κB interacting long non-coding RNA blocks IκB phosphorylation and suppresses breast cancer metastasis. *Cancer Cell* 27, 370–381
135. Lee, S. *et al.* (2016) Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell* 164, 69–80
136. Zhao, L. *et al.* (2016) Long non-coding RNA SNHG5 suppresses gastric cancer progression by trapping MTA2 in the cytosol. *Oncogene* http://dx.doi.org/10.1038/onc.2016.110
137. Johnson, M. *et al.* (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9
138. Yu, G. *et al.* (2012) ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287
139. Kong, L. *et al.* (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349
140. Wang, L. *et al.* (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41, e74
141. Gene Ontology Consortium (2015) Gene ontology consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056
142. Yu, G. and He, Q.Y. (2016) ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* 12, 477–479
143. Kadota, K. *et al.* (2006) ROKU: a novel method for identification of tissue-specific genes. *BMC Bioinformatics* 7, 294
144. Cavalli, F.M. *et al.* (2011) SpeCond: a method to detect condition-specific gene expression. *Genome Biol.* 12, R101
145. Agarwal, V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4, e05005
146. Buske, F.A. *et al.* (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res.* 22, 1372–1381
147. Buske, F.A. *et al.* (2013) Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci. *Bioinformatics* 29, 1895–1897
148. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
149. Kundaje, A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330
150. Lizio, M. *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* 16, 22
151. Melé, M. *et al.* (2015) The human transcriptome across tissues and individuals. *Science* 348, 660–665
152. Wan, Q. *et al.* (2015) BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database 2015* bav019
153. Yang, J.-H. *et al.* (2013) ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Res.* 41, D177–D187
154. Li, J.R. *et al.* (2016) Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Res.* 44, D944–D951
155. Zhao, Z. *et al.* (2015) Co-LncRNA: investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. *Database 2015* bav082
156. Sherry, S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311
157. Paraskevopoulou, M.D. *et al.* (2013) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res.* 41, D239–D245
158. Ning, S. *et al.* (2014) LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs. *BMC Bioinformatics* 15, 152
159. Ning, S. *et al.* (2016) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* 44, D980–D985
160. Jiang, Q. *et al.* (2015) LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. *BMC Genomics* 16 (Suppl. 3), S2
161. Jiang, Q. *et al.* (2015) LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic Acids Res.* 43, D193–D196
162. Volders, P.J. *et al.* (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.* 43, 4363–4364
163. Gong, J. *et al.* (2015) lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Res.* 43, D181–D186
164. Terai, G. *et al.* (2016) Comprehensive prediction of lncRNA–RNA interactions in human transcriptome. *BMC Genomics* 17, 153
165. Li, J.-H. *et al.* (2014) StarBase v2. 0: decoding miRNA–ceRNA, miRNA–ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97
166. Speir, M.L. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* 44, D717–D725