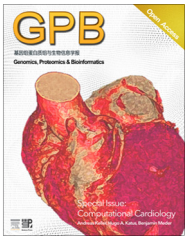
Genomics Proteomics Bioinformatics 14 (2016) 191–199

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



RESOURCE REVIEW

Long non-coding RNA Databases in Cardiovascular Research

Frank Rühle [1,a], **Monika Stoll** [1,2,*,b]

[1] *Institute of Human Genetics, Genetic Epidemiology, University of Muenster, 48149 Muenster, Germany*
[2] *Cardiovascular Research Institute Maastricht (CARIM), Genetic Epidemiology and Statistical Genetics, Maastricht Center for Systems Biology (MaCSBio), Maastricht University, 6211 LK Maastricht, The Netherlands*

Received 18 January 2016; revised 16 March 2016; accepted 17 March 2016
Available online 2 April 2016

Handled by Andreas Keller

KEYWORDS

Database;
Non-coding;
lncRNA;
Gene regulation;
Cardiovascular disease;
ANRIL

Abstract With the rising interest in the regulatory functions of long **non-coding** RNAs (**lncRNAs**) in complex human diseases such as **cardiovascular diseases**, there is an increasing need in public **databases** offering comprehensive and integrative data for all aspects of these versatile molecules. Recently, a variety of public data repositories that specialized in lncRNAs have been developed, which make use of huge high-throughput data particularly from next-generation sequencing (NGS) approaches. Here, we provide an overview of current lncRNA databases covering basic and functional annotation, lncRNA expression and regulation, interactions with other biomolecules, and genomic variants influencing the structure and function of lncRNAs. The prominent lncRNA antisense noncoding RNA in the INK4 locus (***ANRIL***), which has been unequivocally associated with coronary artery disease through genome-wide association studies (GWAS), serves as an example to demonstrate the features of each individual database.

Introduction

Although substantial genetic heritability is estimated for complex cardiovascular diseases, *e.g.*, 40% in coronary artery disease (CAD) [1], and extraordinary efforts have been made in genome-wide association studies (GWAS) and meta-analyses to identify genetic variants leading to CAD, only a small fraction of genetic variance of CAD of ~10% can be explained by genetic variants in protein-coding genes [2]. Additionally, the high proportion of GWAS associations in non-coding genome regions contradicts the simple view of potentially-deleterious protein mutations and indicates a complex regulatory network driven by non-coding RNAs (ncRNAs) [3,4]. Since only 1% of the mammalian genome is translated into proteins, but approximately 85% of the genome is transcribed into RNA, ncRNAs potentially represent an additional layer of epigenetic regulation. Especially long ncRNAs (lncRNAs, RNA >200 nucleotides in length) provide a wide range of regulatory functions including interactions with DNA, RNAs, and proteins [5,6].

* Corresponding author.
E-mail: mstoll@uni-muenster.de (Stoll M).
[a] ORCID: 0000-0001-7688-2394.
[b] ORCID: 0000-0002-2711-4281.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2016.03.001

For instance, the lncRNA X-inactive-specific transcript (*XIST*) directly binds to the polycomb repressive complex 2 (PRC2) and thereby downregulates the entire chromosome during X-chromosome inactivation [7]. Other lncRNAs influence gene activity by RNA-directed chromatin remodeling [8], RNA-directed DNA methylation [9] or as activator or repressor molecules for transcription factors (TFs) [10,11]. By recruiting splicing factors or by masking splice junctions of mRNAs, lncRNAs can influence alternative splicing of coding genes [12]. Various lncRNA interactions with microRNAs (miRNAs) impact mRNA stability by masking miRNA-responsive elements or by competing for miRNA binding in competing endogenous RNA (ceRNA) networks [13,14]. Additionally, discrimination between coding and non-coding genes is sometimes ambiguous, because functional lncRNA transcripts containing open reading frames may also be translated to (small) proteins [15].

Dysregulated expression or function of lncRNAs has been recognized to contribute to heart development and complex cardiovascular diseases [6]. For instance, transcript levels of the antisense noncoding RNA in the INK4 locus (*ANRIL*, alias *CDKN2B-AS1*) lncRNA, which is encoded on chromosome 9p21 at the strongest genetic susceptibility locus for CAD, are directly correlated with the severity of atherosclerosis [16]. The locus at chromosome 5q31 carrying the non-coding steroid receptor RNA activator (*SRA1*) as well as human leukocyte antigen (HLA) complex group 22 (*HCG22*) at chromosome 6p21 have been significantly associated with dilated cardiomyopathy (DCM) [17,18]. The myocardial infarction (MI)-associated transcript (*MIAT*) encoded on chromosome 22q12 is implicated to play a role in MI [19].

To discover potentially harmful lncRNA functions, it is important to understand the complex interaction networks of these molecules. In general, expression of lncRNAs is more specific for cell type and developmental stage than that of protein-coding genes [20]. Functional prediction of lncRNAs is more difficult than, *e.g.*, that for smaller miRNAs, because function of a lncRNA is not solely determined by its nucleotide sequence, but by the resulting secondary structure enabling it to interact with other biomolecules [21]. This is supported by the fact that lncRNA sequences are less conserved than miRNAs or protein-coding genes except for their promoter regions [22]. Genomic variants in lncRNA sequences may induce abnormal expression and function of their harboring lncRNAs, *e.g.*, by gaining or losing binding sites for interaction partners or by altering the secondary structure even at distant positions of the RNA molecule, possibly explaining part of genetic susceptibility to certain diseases [23]. Many of the aforementioned disease-associated lncRNAs like *ANRIL*, *MIAT*, and *HCG22* have gene variants, whose structural impact is not yet understood. Generally, there is a large gap between the number of identified lncRNAs and their known functional impact.

Therefore there is a need for comprehensive lncRNA databases to utilize the huge experimental datasets from current high-throughput technologies joined by massively parallel sequencing such as RNA-Seq, chromatin immunoprecipitation (ChIP)-Seq, RNA immunoprecipitation (RIP)-Seq, cross-linking immunoprecipitation (CLIP)-Seq or chromatin isolation by RNA purification (ChIRP)-Seq [24]. In addition to the main genomics data portals from NCBI, EMBL and UCSC, which also provide data on non-coding genes, several specialized databases have been developed that collect and integrate data in the context of lncRNAs [25,26]. All databases discussed here are accessible via a web-based interface and have been published in peer-reviewed journals (**Table 1** and **Table 2**). Apart from these, there are further data repositories with downloadable data files like the Human lincRNA Catalog from Broad Institute [20]. While some databases have been performing well for several years, many specialized databases have been developed in very recent time, highlighting the strong momentum of this research field.

In the following, we will give an overview of selected databases for different kinds of lncRNA-related information (**Figure 1**). The suggested analysis outline is exemplified by the CAD-related human lncRNA *ANRIL*. All specifications of database contents and query results refer to status of March, 2016 (Table 1 and Table 2). To start a query for a lncRNA of interest, basic information about lncRNA type, chromosomal location, nucleotide sequence, expression profiles, and functional annotation may be retrieved at NONCODE [27–31] and lncRNAdb [32]. Roughly, the classification of lncRNA types is based on their genomic context concerning sense, antisense, bidirectional, intergenic (lincRNAs), or intronic lncRNAs [33]. Known biological functions such as gene ontology (GO) annotation and disease associations are documented in LncRNADisease [34]. In addition to providing lncRNA expression profiles, lncRNAtor [35] offers co-expression data for protein-coding genes to identify potential functional connections between coding and non-noncoding transcripts. To gain insights in regulation of lncRNA expression, ChIPBase [36] contains information on TFs that regulate the expression of non-coding genes. In the next step, the interactions of lncRNAs with other biomolecules may be examined by using starBase [37] and NPInter [38]. These databases provide experimentally-validated data on interactions with proteins, DNA, and other RNA types, especially miRNAs. Finally, genomic variations within the lncRNA gene sequence can be analyzed to explore their potential functional influence on the lncRNA transcript using lncRNASNP [39].

NONCODE 2016

NONCODE was first published in 2005 as an integrated knowledge database of ncRNAs [27] and has repeatedly been updated since then [28–31]. Its latest version NONCODE 2016 offers data for 16 species including 167,150 human lncRNAs [31]. In addition to lncRNA class, chromosomal location, sequence, Coding-Non-Coding Index (CNCI) for protein coding prediction and expression profiles, the database included conservation annotation and disease association as new features in its latest version. The collected data are curated from published literature and include input from other public databases such as Ensembl [40], RefSeq [41], lncRNAdb [32], and GENCODE [42]. The database established a lncRNA nomenclature consisting of "NON", a three character code that specifies the species, "T" or "G" for transcript or gene, respectively, followed by six sequential numbers and a version number where applicable. For *ANRIL* (NONHSAG051899), we find 22 transcript isoforms of type 'antisense' on chromosome 9, which are mostly expressed in lung, lymph nodes, prostate, skeletal muscle, and white blood cells. However, although this molecule has been linked in the literature to