Bioinformatics, 31(21), 2015, 3537–3543
doi: 10.1093/bioinformatics/btv410
Advance Access Publication Date: 10 July 2015
Original Paper

Databases and ontologies

C-It-Loci: a knowledge database for tissue-enriched loci

Tyler Weirick[1,2], David John[1,2], Stefanie Dimmeler[1,2] and Shizuka Uchida[1,2,*]

[1]Institute of Cardiovascular Regeneration, Centre for Molecular Medicine, Goethe University Frankfurt and [2]German Center for Cardiovascular Research, Partner side Rhein-Main, Frankfurt am Main, Germany

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on March 23, 2015; revised on June 24, 2015; accepted on July 7, 2015

Abstract

Motivation: Increasing evidences suggest that most of the genome is transcribed into RNAs, but many of them are not translated into proteins. All those RNAs that do not become proteins are called 'non-coding RNAs (ncRNAs)', which outnumbers protein-coding genes. Interestingly, these ncRNAs are shown to be more tissue specifically expressed than protein-coding genes. Given that tissue-specific expressions of transcripts suggest their importance in the expressed tissue, researchers are conducting biological experiments to elucidate the function of such ncRNAs. Owing greatly to the advancement of next-generation techniques, especially RNA-seq, the amount of high-throughput data are increasing rapidly. However, due to the complexity of the data as well as its high volume, it is not easy to re-analyze such data to extract tissue-specific expressions of ncRNAs from published datasets.

Results: Here, we introduce a new knowledge database called 'C-It-Loci', which allows a user to screen for tissue-specific transcripts across three organisms: human, mouse and zebrafish. C-It-Loci is intuitive and easy to use to identify not only protein-coding genes but also ncRNAs from various tissues. C-It-Loci defines homology through sequence and positional conservation to allow for the extraction of species-conserved loci. C-It-Loci can be used as a starting point for further biological experiments.

Availability and implementation: C-It-Loci is freely available online without registration at http://c-it-loci.uni-frankfurt.de.

Contact: uchida@med.uni-frankfurt.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A series of articles by 'Functional Annotation Of Mammalian genome (FANTOM)' projects (Carninci *et al.*, 2005), 'ENCyclopedia Of DNA Elements (ENCODE)' consortium (Consortium, 2012), and others clearly indicate that a majority of genome is transcribed in the form of RNAs, yet only few percentages of them fall under the category of protein-coding genes. The current estimate is that less than 3% of the mammalian genome encodes for protein-coding genes (Lander *et al.*, 2001; Uchida *et al.*, 2012; Uchida and Dimmeler, 2015). Previously, many of the remaining RNAs were discarded as transcriptional noises and experimental errors. However, through the discovery of microRNAs (miRNAs) and other non-coding RNAs (ncRNAs) (e.g. long non-coding RNAs (lncRNAs)), it became evident that RNAs have functions beyond templates for protein expression. The concept of ncRNAs is not new, as ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) do not encode for proteins but are necessary for protein translation. Given such a complex reality of transcriptomes, lncRNAs are an immensely important topic of study

for understanding biological systems. Furthermore they are of major interest in medicine with a huge range of potential applications from the detection of neurological diseases and treatments, anticancer therapies, and even cardiac regeneration (Uchida *et al.*, 2012; Qureshi and Mehler, 2013; Tang *et al.*, 2013). However, compared to protein-coding genes, the functions of lncRNAs are poorly understood. One of the major reasons is that lncRNAs exhibit low sequence conservation, making evolutionary comparison difficult. Indeed, a recent study using a computational pipeline for homology-based comparison found that only 3.4% of lncRNAs share sequence conservation between human and other non-marsupial mammals (Derrien *et al.*, 2012) [compared to 85% of protein-coding genes between human and mice (Batzoglou *et al.*, 2000)]. Although lncRNAs are poorly conserved among species, their expression patterns are distinct compared to those of protein-coding genes as a majority of lncRNAs (both with high and low sequence conservation) is expressed in a tissue-specific manner (Cabili *et al.*, 2011). Furthermore, when a set of highly conserved lncRNAs was examined, their tissue expressions are conserved between species (Derrien *et al.*, 2012). Since it is highly unlikely that the level of (convergent) evolution has occurred to allow for the specificity of lncRNAs when comparing to their poor sequence conservation, it is possible that evolutionary pressure is effective but not on a sequence-specific manner, which might suggests a sequence-independent way to understand the conservation of lncRNAs is necessary. Indeed, a number of methods have already been explored. One is based on the observation that the secondary structures of a number of lncRNAs are conserved (Johnsson *et al.*, 2014). However, a large-scale attempt using such methodology is still impeded by poor understanding of secondary structures of RNA in general (Johnsson *et al.*, 2014). Other studies have observed that promoter regions of lncRNAs are relatively conserved even when compared to those of protein-coding genes (Guttman *et al.*, 2009; Derrien *et al.*, 2012). Although defining species-conservation via promoters is valuable, recent evidences from chromatin conformational capture assays (e.g. 3C, 4C, Hi-C) suggest that a promoter and enhancer could be up to 120kb from the transcription start sites (TSS) of a gene to control its expression (Sanyal *et al.*, 2012). If this applies to the mammalian genome, it would be difficult to define which region is controlling the expression of one gene but not the other.

To overcome these challenges, in this study, we introduced positional conservation based on the idea that certain regions of the genome are conserved from one species to another. This region is defined as a pair of adjacent genes, which share adjacency and homology when compared across another species. Applying this concept to three organisms (human, mouse and zebrafish), we built a knowledge database called 'C-It-Loci' (http://c-it-loci.uni-frankfurt.de) using published RNA-seq datasets from various studies covering different tissues. Using C-It-Loci, we revisited the idea of tissue specificity of protein-coding genes and lncRNAs. Furthermore, we examined the definition of housekeeping genes and applied its concept to screen for housekeeping lncRNAs. To show the applicability of C-It-Loci, we screened for tissue-specific lncRNAs and confirmed their expressions in the target tissues by surveying publicly-available biological experimental data.

2 Methods

C-It-Loci (Fig. 1a) contains the information about three species (human, mouse and zebrafish) based on their annotations (e.g. gene names/symbols, biotypes) from Ensembl (version 77) as follows: GRCh38 (hg38), GRCm38 (mm10) and Zv9 (danRer7) for *Homo sapiens*, *Mus musculus* and *Danio rerio*, respectively (Flicek *et al.*, 2014). Of note, in the case of human genome, hg38 significantly differs from its previous human genome assembly hg19 (Supplementary Table S1).

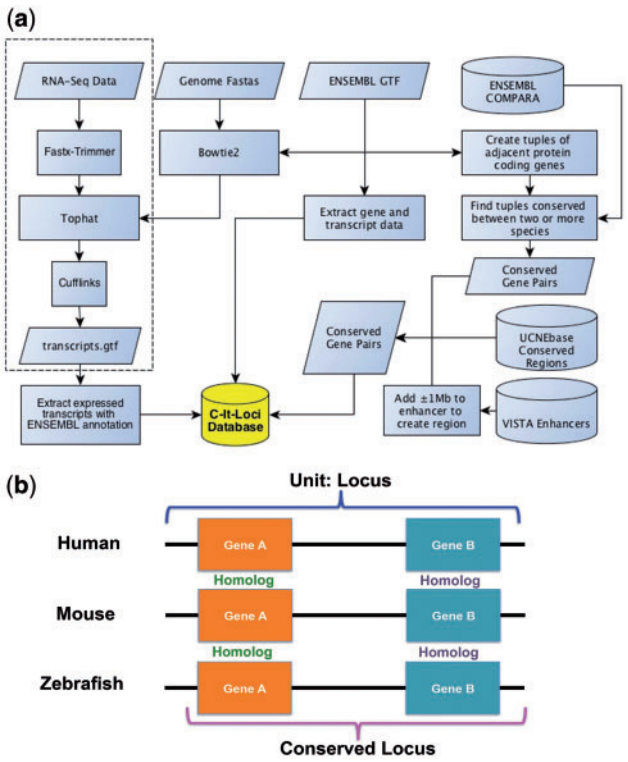

Fig. 1. Scheme of C-It-Loci. (**a**) Flowchart of building of C-It-Loci. All the analyzed results were imported as MySQL data tables into C-It-Loci. (**b**) Definition of CGP. The genomic coordinates from one protein-coding gene ('Gene A') to the immediately downstream protein-coding gene ('Gene B') are defined as one locus unit. When homologous protein-coding genes are found in another species for both protein-coding genes in the locus, this locus is defined as 'conserved locus', which we called 'C-It-Loci Genomic Positions (CGP)'

2.1 Generation of conserved regions

Three types of conserved regions were considered. The first type is 'positional conservation' (called 'C-It-Loci Genomic Positions (CGP)') based on the presences of conserved protein pairs (Fig. 1b). They are defined as the set of all pairs of adjacent protein-coding genes within an organism's genome, which are also adjacent when compared to orthologs of another species. Corresponding orthologous pairs were found using the Ensembl Compara database via Ensembl's REST API (Yates *et al.*, 2015). A total of 9757 conserved gene pairs are shared between two or all species in C-It-Loci.

The second region type is based on the ultraconserved elements, which are species-conserved regions that are shown to be transcriptional regulators of key developmental genes (Bejerano *et al.*, 2004). The information about these regions was downloaded from UCNEbase (Dimitrieva and Bucher, 2013). UCNEbase (http://ccg.vital-it.ch/UCNEbase) is a database of ultra-conserved non-coding elements (UCNEs). It contains two types of data: the individual UCNEs themselves and genomic regulatory blocks

(UGRBs). UCNEs were defined as non-coding DNA regions with $\geq 95\%$ sequence identity between human and chicken and identified via whole-genome alignment. UGRBs or 'UCNE clusters' are arrays of UCNEs, which are syntenically conserved (i.e. orthologs

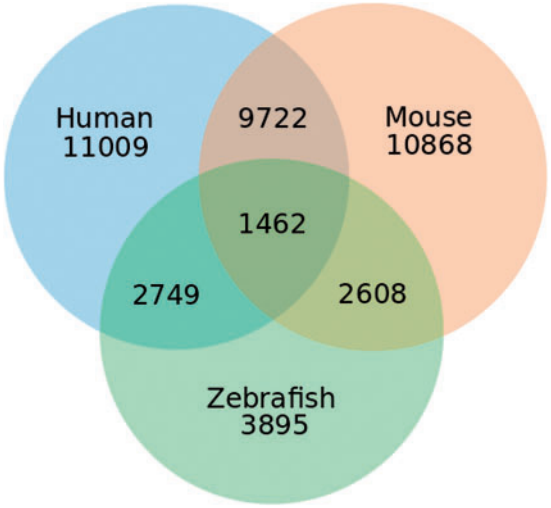

Fig. 2. Homologous regions. The Venn diagram shows three types of conserved regions among three organisms used in this study

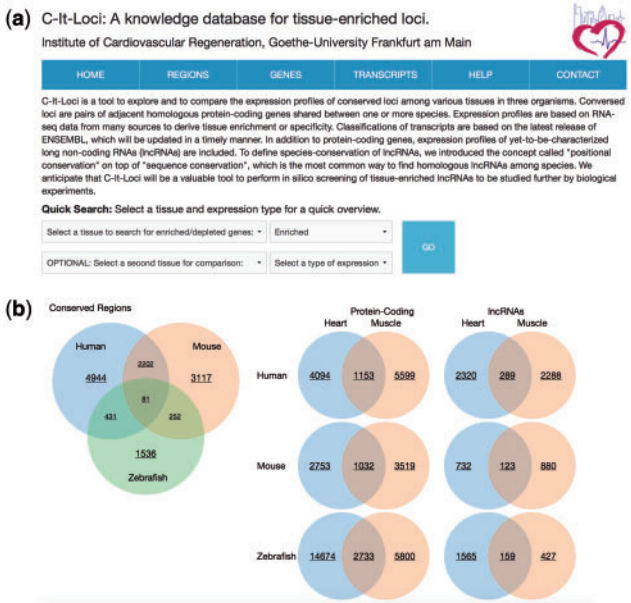

Fig. 3. Screen shots of C-It-Loci. (**a**) Top page. A quick search function is provided to screen for tissue 'expressed', 'enriched' and 'specific' conserved regions as well as transcripts that are separated between protein-coding genes and lncRNAs. Two tissues can be compared for the expressions. (**b**) Results of quick search. Transcripts enriched in both heart and muscle are screened

that are in the same order) between the human and chicken genomes and generally within > 0.5 mega bases (mb) of each other. Orthologous and paralogous regions were identified with and between other species using the program SSEARCH and additional statistical methods regions with E-values less than 1e-4. UGRBs range in size from 4.9 mb to 2 kilobases (kb) and an average of 16 UCNEs. A total of 240 UCNE clusters are included in C-It-Loci.

The third type is based on the species-conserved cis-regulatory elements (enhancers) that are experimentally validated in transgenic mice (Pennacchio *et al.*, 2006). The information about these regions was downloaded from the VISTA Enhancer Browser database (Visel *et al.*, 2007) (http://enhancer.lbl.gov). The enhancer elements used in VISTA were selected for their conservation between human and several non-mammalian vertebrates or extreme conservation (≥ 200bp regions with 100% identity) between human, mouse and rat. To allow for the detection of transcripts around these enhancers, 300 kb upstream and downstream were added to the VISTA regions as enhancers have been shown to act independent of strand and at distances of up to several hundred kb (Maston *et al.*, 2006). A total of 2158 VISTA-conserved regions are included in C-It-Loci.

For the conversion between different genomic assemblies, CrossMap (Zhao *et al.*, 2014) was utilized. In total, 1462 regions share among all species, whereas 9722 between human and mouse, 2749 between human and zebrafish, and 2608 mouse and zebrafish (Fig. 2). The average sizes (in base pairs) of the regions for CGP, UCNEbase, and VISTA regions are 178 097, 337 429 and 601 879, respectively, and their standard deviations are 314 064, 556 697 and 1317, respectively. There were 11 725 regions containing lncRNAs. Of these, 8409 regions share a lncRNA among more than one tissue, and 511 within all three species. Furthermore, 7881 of these regions also contained the same biotype of lncRNA, and 3768 the same enriched tissue, region, and biotype. The overlapped regions among three conservation types are shown in Supplementary Figure S2.

2.2 RNA-seq assembly

Raw sequence datasets were downloaded from the NCBI Sequence Read Archive as SRA files or from EBI as fastq files (Kodama *et al.*, 2012; McWilliam *et al.*, 2013). In total, 119 RNA-seq datasets were utilized in this study (Supplementary Table S2): 28 human (Taxonomy ID: 9606), 78 mouse (Taxonomy ID: 10090) and 13 zebrafish (Taxonomy ID: 7955). In the case of SRA files, fastq-dump (version 2.1.7) was use to convert the SRA files to fastq files (http://www.ncbi.nlm.nih.gov/sra). The first 10 bases of all reads (i.e. adaptor sequences) were removed with fastx-trimmer (version 0.0.13) (Pearson *et al.*, 1997). Although there are different bioinformatics tools to analyze RNA-seq available, we chose to map the trimmed reads with Tophat (version 2.0.11) and to annotate them with Cufflinks (version 2.2.1) as this pipeline has been shown to outperform others when the reference genome is given (Engstrom *et al.*, 2013). Both programs were used with the default setting with the Ensembl version 77 GTF annotations. Only annotated transcripts with FPKM (Fragments Per Kilobase of exon per Million fragments mapped) values greater than 1e-5 were added to the C-It-Loci database.

2.3 The C-It-Loci database

All the information and analyzed RNA-seq data were stored in MySQL (Supplementary Fig. S1). For the web interface, the CakePHP web framework was utilized. C-It-Loci allows users to explore the data from the following three main views: 'REGIONS (conserved regions)', 'GENES' and 'TRANSCRIPTS' (Fig. 3). Each view allows users to browse the data in order of accession as well as to execute a search by a gene name or accession. Furthermore, query tags are allowed to specify searches based on various other types of data (e.g. taxonomy, tissue specificity). The search terms can be combined in any combination and number with the Boolean operators 'and', 'or', 'not' and '()'. A wildcard is also allowed.

From the top page of C-It-Loci, a quick search function is provided for a tissue of interest or in comparison to another tissue. Here, it is possible to screen for expressed (FPKM values above zero), enriched (FPKM values greater than or equal to the average of all tissues for the target transcript), or specific (expressed only in the target tissue) transcripts.

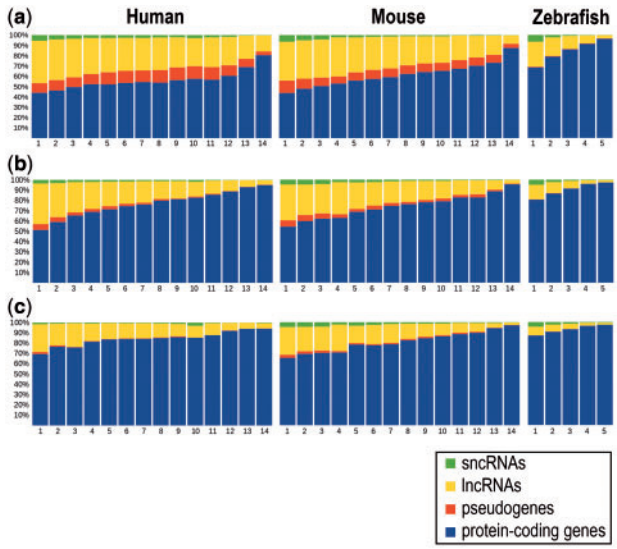

Fig. 4. Percent distributions of transcripts. (**a**) Expressed (FPKM > 0); (**b**) FPKM > 1; and (**c**) FPKM > 5. The x-axis indicates the number of tissues, and the y-axis represents the percent distribution for each condition. 'sncRNAs' stands for short non-coding RNAs. In general, the similar trend of tissue-specific expressions of lncRNAs compared to those of protein-coding genes are observed for human, mouse and zebrafish

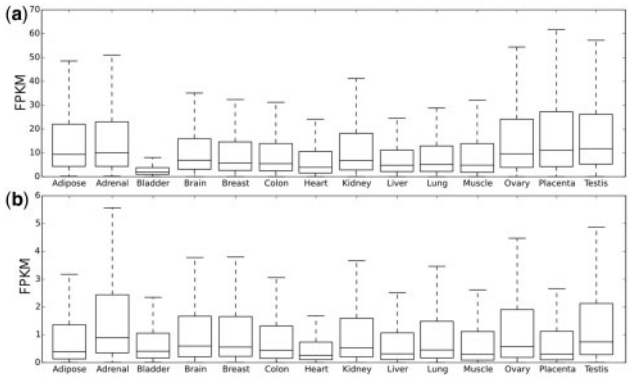

Fig. 5. Distributions of human housekeeping (HK) transcripts. (**a**) HK protein-coding genes. (**b**) HK lncRNAs. In a box-and-whisker plot, the top and the bottom of the box indicate the third (Q3) and first quartiles (Q1), respectively. The band inside the box shows the median, whereas the whiskers represent Q3+1.5*IQR (interquartile range, Q3-Q1) for the top and Q1-1.5*IQR for the bottom. Across tissues, the box-and-whisker plots show that the median FPKM values of HK protein-coding genes vary significantly, especially for bladder. HK lncRNAs show similar distribution across tissues, but their FPKM values are 10-fold lower than those of protein-coding genes

Gene Ontology (GO) annotations were obtained from the GO annotations available for each transcript on the Ensembl database. The GO terms available from the gene view are simply the unique set of GO terms for each transcript produced by the gene. For each GO term, a link to AmiGo 2 is provided (http://amigo.geneontology.org/amigo) (Ashburner *et al.*, 2000).

2.4 Exon array analysis

The sample dataset consisting of 11 tissues of human and mouse were downloaded from Affymetrix, Inc. The CEL files were uploaded to our noncoder web interface (Gellert *et al.*, 2013). The data were pre-processed using RMA algorithm (Bolstad *et al.*, 2003). To calculate the correlation between exon arrays and RNA-seq, the information from ENSEMBL was paired with the microarray data. Any exon array entry matching multiple transcripts was mapped to all potential RNA-seq transcripts. To allow for the comparison between exon array and RNA-seq datasets, any Transcript Cluster IDs on exon array with no RNA-Seq expression were removed from the further analysis.

2.5 Screening for housekeeping lncRNAs

The previously described set of housekeeping protein-coding genes (Eisenberg and Levanon, 2013) was downloaded from http://www.tau.ac.il/~elieis/HKG/. Based on this list, a standard deviation (SD) was calculated for each housekeeping protein-coding gene using FPKM values across tissues. Then, an average SD was derived for housekeeping protein-coding genes. Using this average SD as a threshold, when a standard deviation of the FPKM values of the lncRNA is lower than that of the average SD, we defined this lncRNA to be a 'housekeeping lncRNA'.

2.6 Expression patterns of lncRNAs compared to those of protein-coding genes

Previous studies indicate that lncRNAs are more tissue specifically expressed than protein-coding genes (Derrien *et al.*, 2012). To test whether such a trend holds true, the transcripts in C-It-Loci were screened across various tissues (Fig. 4). Compared to protein-coding genes, high percent of transcripts are specifically expressed only in a few tissues. Furthermore, this trend holds even more when the threshold for minimum FPKM values was increased. Based on these results, we could confirm that lncRNAs are generally lowly expressed compared to protein-coding genes, and their expression patterns are more tissue-specific than those of protein-coding genes. However, we also noted the presence of lncRNAs whose expressions are detected in all tissues included in C-It-Loci even FPKM values above 5, which corresponds to more than one RNA molecule per cell when homogenous cell type was considered (Mortazavi *et al.*, 2008).

Given that there exist so-called 'housekeeping (HK) genes' that are expressed in various tissues at the relatively constant level, we questioned whether there are such HK lncRNAs exist or not. By using the list of 3804 HK protein-coding genes (Eisenberg and Levanon, 2013), we first examined their expression levels in the tissues deposited into C-It-Loci. As reported in the original article (Eisenberg and Levanon, 2013), the expression levels of these HK protein-coding genes vary quite significantly when the whole transcript is considered compared to that of exon, which the authors used, indicating that the expressions of HK protein-coding genes are not as constant as one would expect when they are examined at the tissue level (Fig. 5a). Next, we calculated an average standard deviation of FPKM values of 3804 HK protein-coding genes across all tissues and use this standard deviation (1.258 in log2 of FPKM values) to screen for HK lncRNAs. Using this criterion, we were able to identify 959 HK lncRNAs. Although their standard deviation of FPKM values are similar to those of HK protein-coding genes, their median FPKM values for human are 10-fold lower than those of HK protein-coding genes (Fig. 5b).

Next, we questioned whether HK protein-coding genes and lncRNAs are located closely to each other on the genome or not. In the case of HK protein-coding genes, 534 loci are in between HK protein-coding genes and 2777 loci with one HK protein-coding gene in either up or downstream genomic location. When HK lncRNAs were examined, 428 out of 959 HK lncRNAs are located in 609 loci. Of 609 loci, 36 loci are in between HK protein-coding

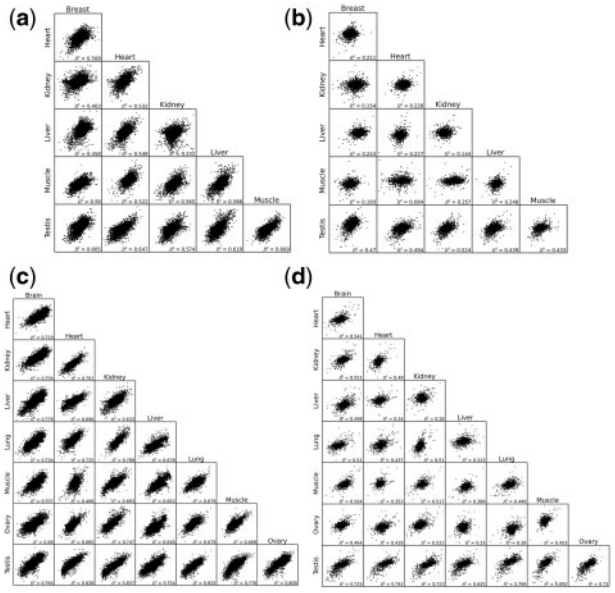

Fig. 6. Comparison to microarray data. (**a**) Human protein-coding genes; (**b**) Human lncRNAs; (**c**) Mouse protein-coding genes; and (**d**) Mouse lncRNAs. The ratios were derived for each tissue comparison as indicated in the figure. Pearson correlations are shown in the figure. In general, the correlations for protein-coding genes are better than those of lncRNAs

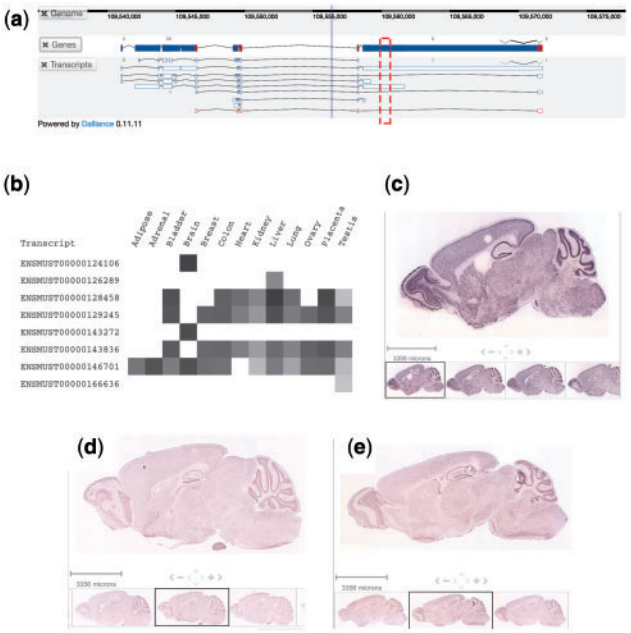

Fig. 7. Case study for brain-specific mouse lincRNAs. (**a**) Genomic locations of *Meg3* transcripts. In C-It-Loci, when each transcript is clicked, a popup window opens to display the details of each transcript as well as a link to the corresponding Ensembl page. A red square indicates the position of *in situ* probe used by the Allen Brain Atlas, which matches to Meg3-001 (ENSMUST00000146701) and Meg3-004. (**b**) Heat map of Meg3 transcripts. Of 11 isoforms, two (Meg3-002 (ENSMUST00000124106) and Meg3-004 (ENSMUST00000143272)) are expressed exclusively in the brain. (c-e) Representative image of *in situ* hybridization of mouse adult brain provided by the Allen Brain Atlas for (**c**) *Meg3* (http://mouse.brain-map.org/experiment/show?id=71281027); (**d**) *A130030D18Rik* (http://mouse.brain-map.org/experiment/show?id=74277353); and (**e**) *A930104D05Rik* (http://mouse.brain-map.org/experiment/show?id=72471695). Image credit: Allen Institute for Brain Science

genes, whereas 236 loci with one HK protein-coding gene in either up or downstream genomic location, which suggest that they are under similar transcriptional regulations.

2.7 Comparison to microarray data

Since 1990s, microarrays have been a technology of choice for transcriptomics studies. In our previous study (Gellert *et al.*, 2013), we compared the results of microarrays (exon arrays) to those of RNA-seq data for the selected set of samples. Here, we revisited our previous study with more tissues and more comprehensive analysis of RNA-seq data (Fig. 6). In the case of protein-coding genes, the correlations between fold changes of RNA-seq and exon array data are very high. The best scores are Pearson correlation of $R2 = 0.669$ between skeletal muscle and testis for human and $R2 = 0.825$ between lung and testis for mice and. In contrast, correlations for lncRNAs are noticeably weaker in general. However, in some cases, correlations are comparable to those of protein-coding genes: $R2 = 0.669$ between skeletal muscle and testis for human and $R2 = 0.723$ between heart and testis for mice. The lower correlations for lncRNAs are likely due to a number of factors. First, the probe coverage of lncRNAs on exon arrays is lower compared to those of protein-coding genes (at least 4 probes per exon of a protein-coding gene) as exon arrays are designed to detect exons of protein-coding genes rather than lncRNAs. Second, as we pointed in our previous study (Gellert *et al.*, 2013) and confirmed further in Figure 4, lncRNAs are lower expressed than protein-coding genes opening the possibility of greater experimental errors. Given such differences, compare to microarrays, RNA-seq data provide more reliable expressions of lncRNAs in various tissues, which C-It-Loci is based on.

2.8 Case study: brain-specific mouse lincRNAs

Although C-It-Loci is based on the experimental data (i.e. RNA-seq), it is important to test the validity of the information and functionalities of C-It-Loci in the context of biological experiments. For this purpose, we carried out a case study for long intergenic ncRNAs (lincRNAs). The Allen Brain Atlas (http://www.brain-map.org) provides detailed information about transcripts expressed in the human and mouse brains, including *in situ* hybridization results of mouse brain. To test the validity of our C-It-Loci, from the 'TRANSCRIPTS' tab, mouse lincRNAs that are brain-specific within CGP conserved region are screened using the following query: 'TAXID:10090 and BIOTYPE:lincRNA and SPECIFIC:Brain and REGIONTYPE:CGP'. This query yielded 22 transcripts from 20 genes (Supplementary Table S3). Of these 20 genes, three are included in the Allen Brain Atlas. In the case of *Meg3*, both Meg3-002 (ENSMUST00000124106) and Meg3-004 (ENSMUST00000143272) are highly expressed (FPKM values of 435.831 and 193.165, respectively) (Fig. 7a and b). Indeed, when the *in situ* hybridization images provided by the Allen Brain Atlas were examined, *Meg3* is highly expressed in the mouse adult brain (Fig. 7c). Two others are lowly expressed: *A130030D18Rik* (FPKM = 0.046) and *A930104D05Rik* (FPKM = 0.033). When *in situ* images were examined, their expressions are region-specific in the brain, which explains their low FPKM values when whole brain was considered for RNA-seq data (Fig. 7d and e). By examining the *in situ* hybridization results of mouse brain, it is possible to confirm the expression patterns and to observe the correlation of FPKM values to the experimental data using different technique for the detection of transcripts.

3 Discussion

In this study, we analyzed RNA-seq data of various tissues from human, mouse and zebrafish. Using the latest genomic assemblies from Ensembl, we annotated each detected transcripts with detailed biotypes known to the scientific community to avoid ambiguity currently existing in the field of lncRNAs (Uchida and Dimmeler, 2015). The analyzed datasets can be explored through our C-It-Loci knowledge database, which allows users to screen for tissue-expressed, enriched and specific transcripts. C-It-Loci was built to assist researchers with a limited knowledge about analysis of RNA-seq data as well as bioinformatics and computational programs. An easy-to-use quick search function is provided at the top page of C-It-Loci so that users can perform an *in silico* screening of transcripts that are expressed, enriched and/or specific by comparing up to two tissues. As an output of such *in silico* screening, easy-to-view Venn diagrams are provided for conserved regions and transcripts for protein-coding genes and lncRNAs separately. These Venn diagrams are clickable, which jumps to the table that lists all the identified conserved regions or transcripts. From this table, users can explore the detailed information about each entry. In the detailed information page for conserved region, gene, or transcript, Ensembl ID is provided along with an official gene name and symbol, genomic position and biotype. Furthermore, link out to Ensembl database and the UCSC Genome Browser are provided. Within the same page, genome browser is implemented so that users can visually inspect the location of the identified gene and its corresponding transcripts (i.e. isoforms). In addition, a heat map is provided to examine the expression patterns of the identified gene and its isoforms. A tab is implemented to view only for those expressed tissues as well as numerical FPKM values are provided in the table format. In the case of the identified gene that follows under conserved regions, a list of conserved regions and their region types are provided along with the links to the corresponding information. Since gene ontology terms are often used in the research community to categorize protein-coding genes, GO terms are provided along with the link to the corresponding information page provided by AmiGO 2.

There are many databases available that contain expressions of protein-coding genes across various tissues, including our own 'C-It' knowledge database (Gellert *et al.*, 2010). In the case of lncRNAs, there are three databases available that include the expressions of lncRNAs across tissues; namely, lncRNAMap (Chan *et al.*, 2014), lncRNAtor (Park *et al.*, 2014) and NONCODE (Xie *et al.*, 2014). lncRNAMap is limited to human lncRNAs and provides a link out to the UCSC Genome Browser for expression data, whereas lncRNAtor and NONCODE include more species. Of these three databases, lncRNAtor includes the most amount of next-generation sequencing datasets, including RNA-seq. Compared to these databases, our C-It-Loci differs in the following points. First, the latest genomic assembly for human (hg38) is utilized, whereas all the other databases are based on hg19, which was released in February 2009, and numerous transcripts are not included or ambiguously annotated (Supplementary Table S1). Second, C-It-Loci allows an *in silico* screening of tissue-expressed, enriched and specific transcripts by comparing two tissues, whose function is missing in all other databases. Supplementary Table S4 is provided to give an overview of tissue-specific transcripts in each tissue. Such information can only be obtained through C-It-Loci, which would be valuable to conduct loss-of-function studies to elucidate the functions of such tissue-specific transcripts, including yet-function-unknown lncRNAs. Third, C-It-Loci defines homologous lncRNAs among three organisms, which are not so well defined in all the other databases. Of note, while our study was underway, LNCipededia (Volders *et al.*, 2015), which is a database for human lncRNAs, introduced a similar method of defining homology based on genomic position. However, LNCipedia does not provide a list of conserved mouse and zebrafish lncRNAs. In our C-It-Loci, we utilized three types of conserved regions and provided all the transcripts that follow under each conserved region. Of all the databases of protein-coding genes and lncRNAs currently available to the research community, our C-It-Loci is the only database that includes both protein-coding genes and lncRNAs as well as other types of transcripts [e.g. small nucleolar RNAs (snoRNAs)]. We anticipate that our C-It-Loci serves as a start point for conducting functional studies in tissue-enriched/specific transcripts, including lncRNAs.

Acknowledgements

The authors thank Dr. Pascal Gellert and the members of Institute of Cardiovascular Regeneration for helpful discussions.

Funding

This study was supported by the German Center for Cardiovascular Research (BMBF) to S.D. and S.U., the LOEWE Center for Cell and Gene Therapy (State of Hessen) to S.D. and S.U., the Deutsche Forschungsgemeinschaft (SFB834 to S.D. and S.U; and UC 67/2-1 to S.U.), and the MicroRNA-based Therapeutic Strategies in Vascular Disease (MIRVAD) by the Fondation Leducq to S.D.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Batzoglou,S. *et al.* (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.

Bejerano,G. *et al.* (2004) Ultraconserved elements in the human genome. *Science (New York, N.Y.)*, **304**, 1321–1325.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, **19**, 185–193.

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.

Carninci,P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)*, **309**, 1559–1563.

Chan,W.L. *et al.* (2014) lncRNAMap: a map of putative regulatory functions in the long non-coding transcriptome. *Comput. Biol. Chem.*, **50**, 41–49.

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.

Dimitrieva,S. and Bucher,P. (2013) UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.*, **41**, D101–D109.

Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet. TIG*, **29**, 569–574.

ENCODE Project Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Engstrom,P.G. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.

Flicek,P. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.

Gellert,P. *et al.* (2010) C-It: a knowledge database for tissue-enriched genes. *Bioinformatics (Oxford, England)*, **26**, 2328–2333.

Gellert,P. *et al.* (2013) Noncoder: a web interface for exon array-based detection of long non-coding RNAs. *Nucleic Acids Res.*, **41**, e20.

Guttman,M. *et al*. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.

Johnsson,P. *et al*. (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica et Biophysica Acta*, **1840**, 1063–1071.

Kodama,Y. *et al*. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Landt,S.G. *et al*. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.

Maston,G.A. *et al*. (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.

McWilliam,H. *et al*. (2013) Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597–W600.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Park,C. *et al*. (2014) lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. *Bioinformatics (Oxford, England)*, **30**, 2480–2485.

Pearson,W.R. *et al*. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.

Pennacchio,L.A. *et al*. (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.

Qureshi,I.A. and Mehler,M.F. (2013) Long non-coding RNAs: novel targets for nervous system disease diagnosis and therapy. *Neurotherapeutics J. Am. Soc. Exp. NeuroTherapeutics*, **10**, 632–646.

Sanyal,A. *et al*. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.

Tang,J.Y. *et al*. (2013) Long noncoding RNAs-related diseases, cancers, and drugs. *TheScientificWorldJournal*, **2013**, 943539.

Uchida,S. and Dimmeler,S. (2015) Long noncoding RNAs in cardiovascular diseases. *Circ. Res.*, **116**, 737–750.

Uchida,S. *et al*. (2012) Deeply dissecting stemness: making sense to non-coding RNAs in stem cells. *Stem Cell Rev.*, **8**, 78–86.

Visel,A. *et al*. (2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.

Volders,P.J. *et al*. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.*, **43**, D174–D180.

Xie,C. *et al*. (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.*, **42**, D98–D103.

Yates,A. *et al*. (2015) The Ensembl REST API: Ensembl data for any language. *Bioinformatics (Oxford, England)*, **31**, 143–145.

Zhao,H. *et al*. (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics (Oxford, England)*, **30**, 1006–1007.