miRcode: A map of putative microRNA target sites in the long non-coding transcriptome

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1093/bioinformatics/bts344</td>
</tr>
<tr>
<td>Accessed</td>
<td>November 28, 2017 10:14:33 PM EST</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:10445586">http://nrs.harvard.edu/urn-3:HUL.InstRepos:10445586</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>

(Article begins on next page)
miRcode: a map of putative microRNA target sites in the long non-coding transcriptome

Ashwini Jeggari1, Debora S Marks2 and Erik Larsson1[

1Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, SE-405 30, Sweden and 2Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

Associate Editor: Martin Bishop

ABSTRACT
Summary: Although small non-coding RNAs, such as microRNAs, have well-established functions in the cell, long non-coding RNAs (lncRNAs) have only recently started to emerge as abundant regulators of cell physiology, and their functions may be diverse. A small number of studies describe interactions between small and lncRNAs, with lncRNAs acting either as inhibitory decoys or as regulatory targets of microRNAs, but such interactions are still poorly explored. To facilitate the study of microRNA-lncRNA interactions, we implemented miRcode: a comprehensive searchable map of putative microRNA target sites across the complete GENCODE annotated transcriptome, including 10 419 lncRNA genes in the current version.

Availability: http://www.mircode.org
Contact: erik.larsson@gu.se

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on April 10, 2012; revised on May 18, 2012; accepted on June 10, 2012

1 INTRODUCTION
Large-scale studies in recent years have revealed that mammalian genomes encode thousands of long (>200 nt) transcripts that lack coding capacity, but are otherwise messenger RNA-like. These are collectively referred to as long non-coding RNAs (lncRNAs) (Mercer et al., 2009). Although their overall biological importance has been debated, early functional examples were discovered more than 20 years ago, notably H19 (Branman et al., 1990) and XIST (Brown et al., 1991). Novel lncRNAs are now being uncovered at an increasing rate, with molecular functions that include recruitment of histone-modifying complexes to chromatin (e.g. HOTAIR and HOTTIP (Rinn et al., 2007; Wang et al., 2011)) and modulation of transcription and splicing by molecular interaction with relevant factors (e.g. GAS5 and MALAT1 (Bernard et al., 2010; Kino et al., 2010)).

A small number of studies describe interactions between small and lncRNAs, with lncRNAs acting either as inhibitory decoys of microRNAs (Ebert and Sharp, 2010) or as regulatory targets. In humans, miR-671 targets an antisense transcript of the human CDR1 gene (Hansen et al., 2011), and miR-29 can regulate the lncRNA MEG3 in hepatocellular cancer, although only indirectly (Braconi et al., 2011). Long non-coding transcripts that derive from ultra-conserved regions (T-UCRs) have also been suggested to be microRNA targets (Calin et al., 2007). In plants, the IPS1 lncRNA inhibits miR-399 through a sponge/decoy effect (Franco-Zorrilla et al., 2007). Herpesvirus-encoded RNAs can bind and inhibit human host miR-27 (Cazalla et al., 2010), and the HULC lncRNA can bind and sequester miR-372 in liver cancer (Wang et al., 2010). A pseudogene of the PTEN tumor suppressor can compete for microRNA binding with its coding counterpart (Poliseno et al., 2010), and microRNA inhibition by lncRNAs is important during muscle differentiation (Cesana et al., 2011). The decoy hypothesis is further supported by the observation that microRNAs with many targets in the cell tend to have a diluted effect on each individual target (Arvey et al., 2010).

A recent study used lentiviral small hairpin RNAs to silence 147 lncRNAs at an average efficacy of 75% (Guttman et al., 2011), demonstrating that lncRNAs in general are susceptible to regulation by Argonaute-small RNA complexes despite frequent nuclear localization. However, existing web-accessible microRNA target prediction databases, such as PicTar (Krek et al., 2005), miRanda (Betel et al., 2008) or TargetScan (Friedman et al., 2009), are focused on 3′-untranslated region (UTR) of coding genes and fail to provide predictions for the long non-coding transcriptome.

To simplify the study of microRNA-lncRNA interactions, we here describe miRcode: a comprehensive map of putative microRNA target sites across the GENCODE long non-coding transcriptome (10 419/15 977 lncRNAs genes/transcripts in the current version based on GENCODE V11). miRcode is designed to be an easy to use, web-based tool, with search functionalities to aid hypothesis generation starting from an lncRNA or microRNA of interest. Custom genome browser views and downloadable tab-delimited files are also accessible through the miRcode web interface. miRcode additionally covers other GENCODE gene classes, including 12 549 pseudogenes and 19 999 coding genes both in typical (3′-UTR) and atypical (5′-UTR and CDS) positions.

2 IMPLEMENTATION
miRcode identifies putative target sites based on established principles: seed complementarity and evolutionary conservation (see Supplementary Material for detailed methods). The seed region, encompassing bases 2–8 from the 5′-end of the microRNA, is the major sequence determinant of microRNA targeting (Lewis et al., 2003). The miRcode pipeline (Fig. 1), implemented using Perl, Matlab, PHP and MySQL, searches for complementary matches to the GENCODE (Harrow et al., 2006) transcripts. We consider 7-mers...
and adenosine-flanked 6-mer and 7-mer matches, but not 6-mers as these are only marginally effective (Grimson et al., 2007; Selbach et al., 2008).

GENCODE represents a comprehensive, high-quality description of the polyA transcriptome. It is updated on a regular basis and based largely on full-length or near full-length complementary DNA evidence and additionally contains many known RNA genes and microRNAs. Although all of GENCODE is analyzed and accessible in miRcode, we define a subset of IncRNA genes that produce only predicted non-coding transcripts with a mature (spliced) length of >200 nt. IncRNAs are further subdivided into intergenic (not overlapping with any coding gene) and non-intergenic.

To assess evolutionary conservation, a 46-way Multiz vertebrate genomic multiple alignment (Blanchette et al., 2004; Fujita et al., 2011) is remapped onto transcripts, and site conservation levels are determined based on site presence in primates, non-primate mammals and non-mammal vertebrates. Transcript regions (5′-UTR, CDS and 5′-UTR) in case of coding genes) and possible overlaps with repeat sequences are recorded for each site. Sites are mapped first to transcripts to allow identification of splice-junction-overlaps with repeat sequences are recorded for each site. Sites are returned in the form of lists, aggregated on genes, where conservation levels (fraction of species where site is present) are presented separately for mammals and non-mammal vertebrates. Transcript regions (5′-UTR, CDS and 5′-UTR) in case of coding genes) and possible overlaps with repeat sequences are recorded for each site. Sites are mapped first to transcripts to allow identification of splice-junction-overlaps with repeat sequences are recorded for each site. Sites are returned in the form of lists, aggregated on genes, where conservation levels (fraction of species where site is present) are presented separately for mammals and non-mammal vertebrates. Transcript regions (5′-UTR, CDS and 5′-UTR) in case of coding genes) and possible overlaps with repeat sequences are recorded for each site. Sites are mapped first to transcripts to allow identification of splice-junction-overlaps with repeat sequences are recorded for each site. Sites are returned in the form of lists, aggregated on genes, where conservation levels (fraction of species where site is present) are presented separately for mammals and non-mammal vertebrates. Transcript regions (5′-UTR, CDS and 5′-UTR) in case of coding genes) and possible overlaps with repeat sequences are recorded for each site. Sites are mapped first to transcripts to allow identification of splice-junction-overlaps with repeat sequences are recorded for each site. Sites are returned in the form of lists, aggregated on genes, where conservation levels (fraction of species where site is present) are presented separately for mammals and non-mammal vertebrates.

Workflow for mapping of conserved putative microRNA target sites in lncRNAs

To assess evolutionary conservation, a 46-way Multiz vertebrate genomic multiple alignment (Blanchette et al., 2004; Fujita et al., 2011) is remapped onto transcripts, and site conservation levels are determined based on site presence in primates, non-primate mammals and non-mammal vertebrates. Transcript regions (5′-UTR, CDS and 5′-UTR) in case of coding genes) and possible overlaps with repeat sequences are recorded for each site. Sites are mapped first to transcripts to allow identification of splice-junction-overlaps with repeat sequences are recorded for each site. Sites are returned in the form of lists, aggregated on genes, where conservation levels (fraction of species where site is present) are presented separately for mammals and non-mammal vertebrates. Transcript regions (5′-UTR, CDS and 5′-UTR) in case of coding genes) and possible overlaps with repeat sequences are recorded for each site. Sites are mapped first to transcripts to allow identification of splice-junction-overlaps with repeat sequences are recorded for each site. Sites are returned in the form of lists, aggregated on genes, where conservation levels (fraction of species where site is present) are presented separately for mammals and non-mammal vertebrates.

3 FUNCTIONALITY

The miRcode interface provides basic search functionality for finding putative microRNA-target sites in IncRNAs of interest or predicted targets of specific microRNAs. Sites are returned in the form of lists, aggregated on genes, where conservation levels (fraction of species where site is present) are presented separately for primates, non-primate placental mammals and non-mammal vertebrates. In addition, custom UCSC browser views enable browsing of target sites in a genome context. Tab-delimited text files and BED files provide convenient access to whole-transcriptome target predictions for use in computational projects.

In summary, we provide, in several formats, a pan GENCODE microRNA site map to facilitate further investigation into microRNA regulation of IncRNAs as well as other atypical target regions such as pseudogenes and 5′-UTRs.

Funding: Grants from the Swedish Medical Research Council; the Assar Gabrielsson Foundation; the Magnus Bergvall Foundation; and the Lars Hierta Memorial Foundation.

Conflict of Interest: None declared.

REFERENCES


