

Module 9: Non-coding RNA resources

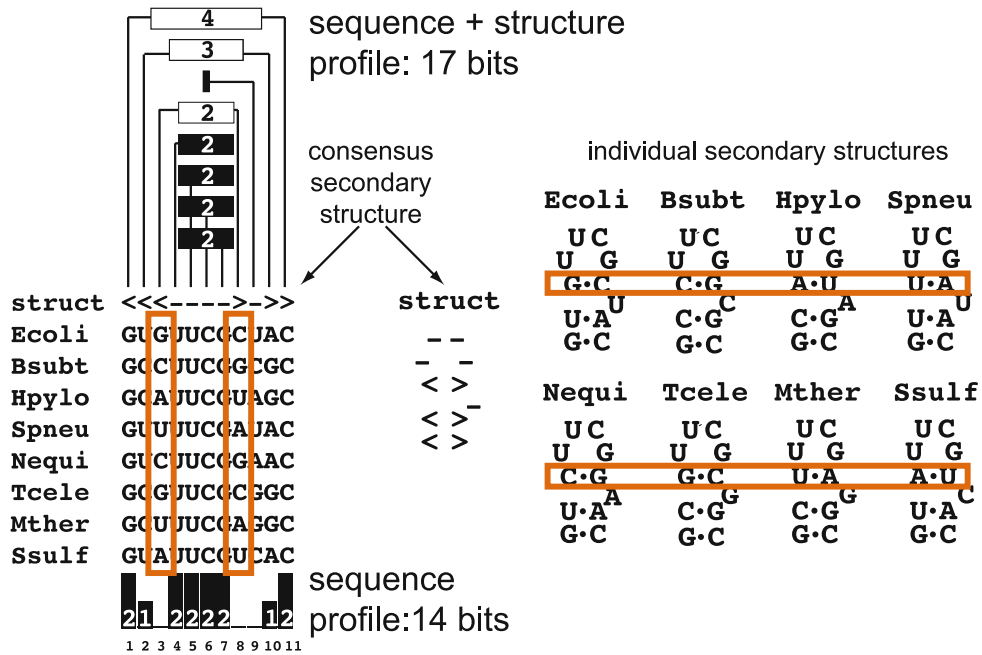
Aims

- Introduce several non-coding RNA databases
- Focus on regulatory microRNAs and resources that catalogue their genomic targets
- Human annotated nc-RNAs
- Outline the use of RNACentral, a database of non-RNA sequences

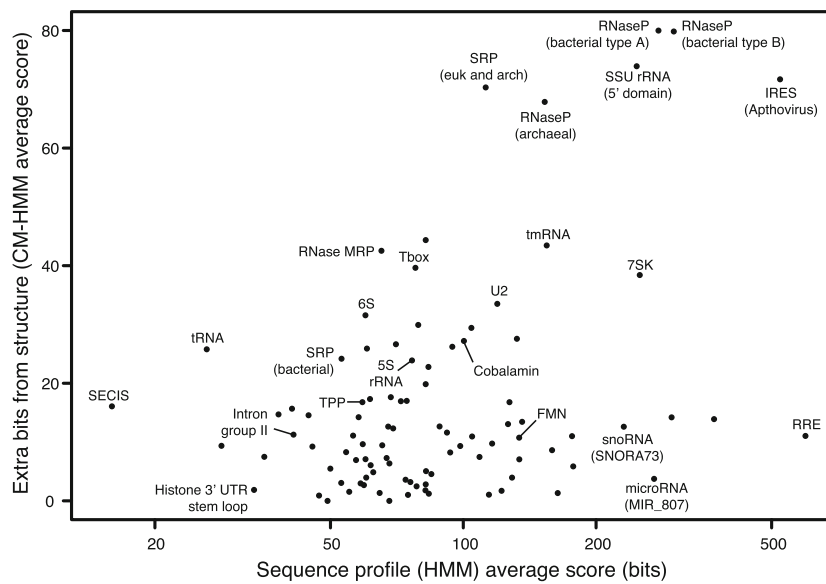
Introduction

The central paradigm of molecular biology is that DNA is transcribed into RNA and RNA is translated to protein. However, biology is never that simple! There has been tremendous growth in the number of reported sequences of non-coding RNAs (ncRNAs). Large-scale genome sequencing has enabled the identification of new representatives of well-known functional classes and many new types of ncRNA to be defined, including piRNAs and circRNAs. Consequently, the role of RNA in regulatory and functional processes has received an increasing amount of attention. Notably microRNAs have been found to have important roles in development and cancer. Due to the increase in experimentation, several database have emerged that catalogue non-coding RNAs, but typically with a single resource dedicated to a particular type/class of ncRNA. In this module we will investigate some of these resources that describe different non-coding RNAs.

To achieve their function, most functional non-coding RNAs adopt a defined structure. Unlike protein sequences, the secondary structure (formed by base pairings) can be more conserved than the primary sequence. This is illustrated below for a series of stem loops (modified from Chapter 9, Annotating functional RNAs in genomes using Infernal, by Eric Nawrocki, in RNA Sequence, Structure and Function: Computational and Bioinformatic Methods).



In the highlighted alignment section (orange box) the sequences show little sequence conservation. However, in all cases the secondary is conserved (shown on the right panel), with the standard Watson-Crick base pairing present. Covariance models, similar to profile HMMs, model the sequence conservation and additionally any secondary structure. Below shows a graph (taken from the same book chapter) that indicates how different RNA families benefit from modelling the secondary structure.





7.1 Non-Coding RNA Families

Rfam – The largest collection of non-coding RNAs families is Rfam, which is produced by the same laboratory as the Pfam database. Rfam contains a wide-ranging catalogue of non-coding RNA families, with each entry containing an alignment, consensus secondary structures that are used to construct a co-variance model. The co-variance model is used to identify new instances of the Rfam entry on new sequences. Rfam has similar concepts to Pfam and is available on the web at the following URL: <http://rfam.xfam.org>

Rfam can allow you to:

- 1) Find out the function of RNA genes and elements
- 2) Identify secondary structure and sequence variation of RNA genes and elements

Exploring an Rfam entry

EMBL-EBI  HOME | SEARCH | BROWSE | FTP | BLOG | HELP  keyword search Go

Rfam 12.0 (July 2014, 2450 families)

The Rfam database is a collection of RNA families, each represented by **multiple sequence alignments**, **consensus secondary structures** and **covariance models (CMs)**. [More...](#)

QUICK LINKS

SEQUENCE SEARCH Analyze your RNA sequence for Rfam matches

VIEW AN RFAM FAMILY View Rfam family annotation and alignments

VIEW AN RFAM CLAN View Rfam clan details

KEYWORD SEARCH Query Rfam by keywords

TAXONOMY SEARCH Fetch families or sequences by NCBI taxonomy

JUMP TO Go [Example](#)

Enter any type of accession or ID to jump to the page for a Rfam family, sequence or genome

Or view the [help](#) pages for more information

STEP 1 -
Go to the Rfam homepage, click 'view an rfam family'. Enter "yybP-ykoY" into the textfield

STEP 2 -
Click on 'Alignments'

HOME | SEARCH | BROWSE | FTP | BLOG | HELP

Rfam
keyword search Go

Family: yybP-ykoY (RF00080)
Description: *yybP-ykoY leader*

1006 sequences 885 species 0 structures

Summary

Wikipedia annotation [Edit Wikipedia article](#)

The Rfam group coordinates the annotation of Rfam families in [Wikipedia](#). This family is described by a Wikipedia entry entitled **yybP-ykoY leader**. You can see the Wikipedia page for this family [here](#). [More...](#)

The **yybP-ykoY leader** RNA element was originally discovered in *E. coli* during a large scale screen and was named SraF.^[1] This family was later found to exist *upstream* of related families of protein genes in many bacteria, including the *yybP* and *ykoY* genes in *B. subtilis*. The specific functions of these proteins are unknown, but this structured RNA element may be involved in their genetic regulation as a riboswitch.^[2]

References

- Argaman, L; Hershberg R; Vogel J; Bejerano G; Wagner EG; Margalit H; Altuvia S (2001). "Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*". *Curr Biol* **11** (12): 941-950. doi:10.1016/S0960-9822(01)00270-6. PMID 11448770.
- Barrick, JE; Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser JK, Breaker RR (2004). "New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control". *Proc Natl Acad Sci USA* **101** (17): 6421-6426. doi:10.1073/pnas.0308014101. PMC 404060. PMID 15096624.

External links

- Page for *yybP-ykoY leader* at Rfam
- This molecular or cell biology article is a stub. You can help Wikipedia by expanding it.

This page is based on a [Wikipedia article](#). The text is available under the [Creative Commons Attribution/Share-Alike License](#).

yybP-ykoY leader

Predicted secondary structure and sequence conservation of *yybP-ykoY*

Identifiers

Symbol	yybP-ykoY
Alt. Symbols	SraF
Rfam	RF00080 G

Other data

RNA type	Cis-reg
Domain(s)	Bacteria
SO	0000233 G

The sequences tab contains up-to the first 300 hits in the family, with the complete list of full hits downloadable. If you want to align all of these sequences you will have to download the CM and Internal and run *cmalign*.

Now we will see how to get the **seed** multiple sequence alignment for the RNA family, to help understand the sequence variation present in the family.

STEP 3 -
Select 'Seed' and 'HTML',
then click 'view'

The screenshot shows the Rfam website interface. At the top right, there is a search bar and navigation links: SEARCH | BROWSE | FTP | BLOG | HELP. Below this, there are statistics: 1006 sequences, 0 species, and 0 structures. The main content area is titled 'Alignment' and contains several sections: 'View options', 'Formatting options', and 'Download'. A yellow callout box labeled 'STEP 3 - Select 'Seed' and 'HTML', then click 'view'' points to the 'View' button in the 'View options' section. The 'View options' section includes a 'Viewer' dropdown menu set to 'jalview' and a 'View' button. The 'Formatting options' section includes an 'Alignment format' dropdown set to 'Stockholm' and 'Download/view' radio buttons with 'Download' selected. The 'Download' section provides a link to download a gzip-compressed Stockholm-format file.

The screenshot shows the EMBL-EBI website interface for a seed sequence alignment. The title is 'Seed sequence alignment for RF00080'. The alignment is displayed as a grid of nucleotide sequences from various species, with some base pairs highlighted in blue and red. A callout box labeled 'Link to EMBL entry' points to a small icon next to the species names. Another callout box labeled 'Base pairs are coloured' points to the highlighted nucleotides. At the bottom, there is a 'Toggle labels between species names and sequence' button and a 'Close window' checkbox.

STEP 4 -
Select 'Secondary structure' tab

Note the letter coding on these structures. This is because they use the IUPAC ambiguity codes. The following table provided you with the designation of the different letters.

Symbol	Meaning	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A, C or T	not-G, H follows G in the alphabet
B	G, T or C	not-A, B follows A
V	G, C or A	not-T (not-U), V follows U
D	G, A or T	not-C, D follows C
N	G, A, T or C	aNy



Compare the predicted secondary structure with that on the Wikipedia page. Do they differ? Are they the same?

These larger secondary structures can often be broken down into smaller structural components, or motifs. Rfam now tries to show these motifs on the

secondary structure. However, due to their small size, they can be very difficult to predict in the absence of a known 3D structure.

STEP 5 -
Select 'Motif Matches' tab

This will list three different motif matches. None of these matches are particularly strong, so should be treated with caution, however, they do provide clues to how the secondary structure element function.

EMBL-EBI  HOME | SEARCH | BROWSE | FTP | BLOG | HELP  keyword search




Family: *yybP-ykoY* (RF00080)
Description: *yybP-ykoY* leader

Summary
Sequences
Alignment
Secondary structure
Species
Trees
Structures
Motif matches
Database references
Curation
Jump to...



Motif matches

There are 3 motifs which match this family.

This section shows the Rfam motifs that match sequences within the seed alignment of this family. Users should be aware that the motifs are structural constructs and do not necessarily conform to taxonomic boundaries in the way that Rfam families do. [More...](#)

Motif Accession	Motif Description	Number of Hits	Fraction of Hits	Sum of Bits	Image
RM00008	GNRA tetraloop	8	0.276	71.3	
RM00022	Rho independent terminator 2	5	0.172	66.4	
RM00029	UNCG tetraloop	5	0.172	58.2	

STEP 6 -
Select 'GNRA tetraloop' motif

EMBL-EBI  HOME | SEARCH | BROWSE | FTP | BLOG | HELP  keyword search

Motif: *GNRA* (RM00008)
Description: *GNRA* tetraloop

Summary
Alignments
Structures
Family matches
References
Curation
Jump to...

Summary

Wikipedia annotation [Edit Wikipedia article](#)

The Rfam group coordinates the annotation of Rfam data in [Wikipedia](#). This motif is described by a Wikipedia entry entitled [Tetraloop](#). [More...](#)

Tetraloops are a type of four-base [hairpin loop motifs](#) in RNA secondary structure that cap many [double helices](#).^[2] There are many variants of the tetraloop, the published ones include ANYA,^{[3][4]} CUYG,^[5] GNRA,^[6] UMAC^[7] and UNCG.^[8]

Three types of tetraloops are common in ribosomal RNA: GNRA, UNCG and CUUG. The GNRA tetraloop has a guanine-adenine base-pair where the guanine is 5' to the helix and the adenine is 3' to the helix. Tetraloops with the sequence UMAC have essentially the same backbone fold as the GNRA tetraloop,^[9] but may be less likely to form tetraloop-receptor interactions. They may therefore be a better choice for closing stems when designing artificial RNAs.

See also

- [RNA Tertiary Structure](#) (section [Tetraloop-receptor interactions](#))

References

- [^] Cate, J.H., Gooding, A.R., Pedell, E., Zhou, K., Golden, B.L., Kundrot, C.E., Cech, T.R., Doudna, J.A. (1996). "Crystal structure of a group I ribozyme domain: principles of RNA packing.". *Science* **273** (5282): 1676-1685. doi:10.1126/science.273.5282.1678 [PMID 8781224](#) [G](#).
- [^] Woese, C.R., Winkers, S., Gutell, R.R. (1990). "Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops" ". *Proc. Natl. Acad. Sci. USA* **87** (21): 8467-

Structure of a GNRA tetraloop from a group I self-splicing intron.^[1]
















The page above describes the motif and contains the alignment used to build the CM for the motif prediction, know 3D structures, references and curation details, similar to a family page. It also lists all 'Family matches', which contains all Rfam entries that contain the GNRA tetraloop.

Motif: GNRA (RM00008)
 Description: *GNRA tetraloop*

Summary
 Alignments
 Structures

Family matches

There are **343** Rfam families which match this motif.
 This section shows the families which have been annotated with this motif. Users should be aware that the motifs are structural constructs and do not necessarily conform to taxonomic boundaries in the way that Rfam families do. [More...](#)

Family Accession	Family Description	Number of Hits	Fraction of Hits	Sum of Bits	Image
RF00001	5S ribosomal RNA	279	0.392	2678.1	
RF00004	U2 spliceosomal RNA	18	0.087	162.3	
RF00007	U12 minor spliceosomal RNA	11	0.177	125.0	
RF00008	Hammerhead ribozyme (type III)	8	0.098	81.7	
RF00009	Nuclear RNase P	49	0.422	580.3	
RF00010	Bacterial RNase P class A	453	0.989	17762.2	
RF00011	Bacterial RNase P class B	102	0.895	2028.5	
RF00012	Small nucleolar RNA U3	16	0.184	174.1	
RF00013	6S / SsrS RNA	33	0.221	312.3	
RF00014	DsrA RNA	3	0.600	24.0	
RF00017	Metazoan signal recognition particle RNA	71	0.780	1144.2	
RF00018	CsrB/RsmB RNA family	23	0.605	261.7	
RF00021	Spot 42 RNA	11	0.579	103.0	
RF00022	GcvB RNA	8	0.296	98.0	
RF00023	transfer-messenger RNA	153	0.321	1591.1	

References
 Curation

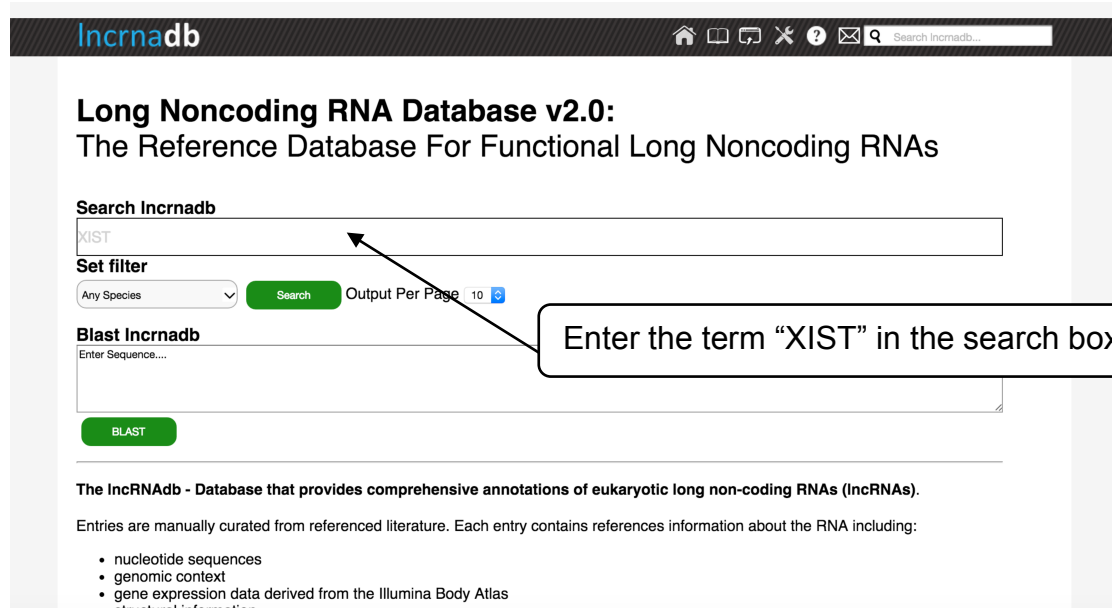
Jump to...
 enter ID/acc

In your own time, go to the secondary structure page for Rfam entry RF00005, tRNA. Compare the sequence conservation (seqcons) and base pair conservation (bpcons). Where is the region of strongest sequence conservation (colours closer to red are more strongly conserved)? Does this correlate with the base pair conservation? Why do think this is?

7.2 Long non-coding RNAs

LncRNA are non-coding transcripts (>200 bp) thought to be involved in various aspects of gene regulation. The precise number of lncRNAs is still of much debate, as the definition is largely arbitrary and the evidence limited to sequence data. To date, it appears that most lncRNAs have **little sequence or secondary conservation**, with typically only a few patches of sequence conservation. Consequently, database such as Rfam do not model lncRNAs very well. The lncRNADB curates lncRNAs from the literature, to provide details about their function.

STEP 1 – Go to the lncRNADB home page: <http://lncrnadb.org>



IncRNadb

Long Noncoding RNA Database v2.0:
The Reference Database For Functional Long Noncoding RNAs

Search IncRNadb

XIST

Set filter

Any Species Output Per Page 10

Blast IncRNadb

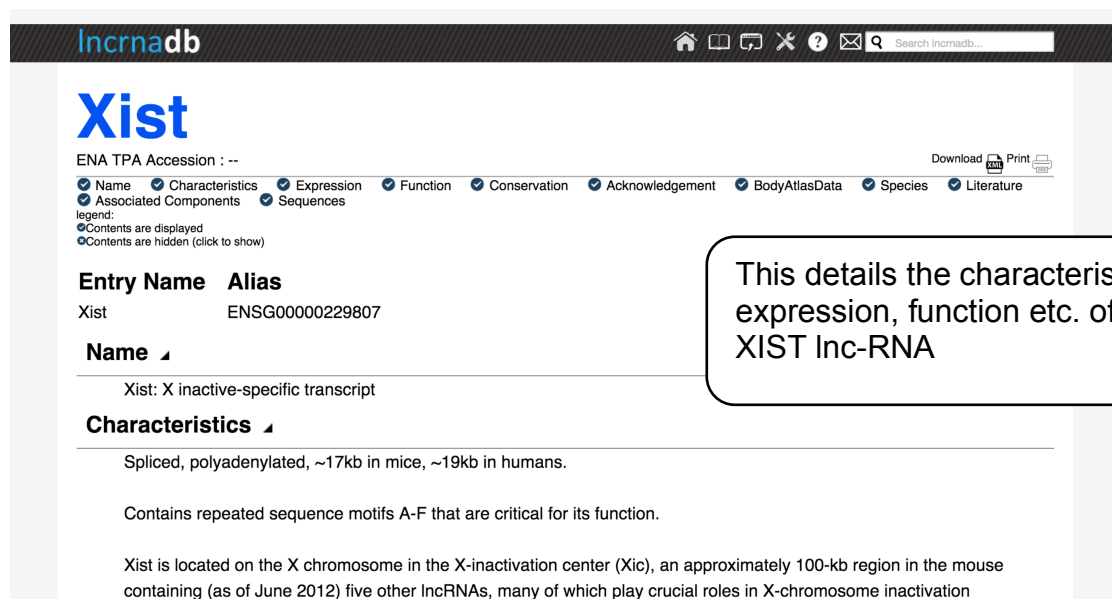
Enter Sequence...

The IncRNadb - Database that provides comprehensive annotations of eukaryotic long non-coding RNAs (lncRNAs).

Entries are manually curated from referenced literature. Each entry contains references information about the RNA including:

- nucleotide sequences
- genomic context
- gene expression data derived from the Illumina Body Atlas
- structural information

From the search results, find the XIST entry and click on it to produce the following entry page.



IncRNadb

Xist

ENA TPA Accession : --

Download

Name Characteristics Expression Function Conservation Acknowledgement BodyAtlasData Species Literature

Associated Components Sequences

legend:
 Contents are displayed
 Contents are hidden (click to show)

Entry Name **Alias**

Xist ENSG00000229807

Name ▾

Xist: X inactive-specific transcript

Characteristics ▾

Spliced, polyadenylated, ~17kb in mice, ~19kb in humans.

Contains repeated sequence motifs A-F that are critical for its function.

Xist is located on the X chromosome in the X-inactivation center (Xic), an approximately 100-kb region in the mouse containing (as of June 2012) five other lncRNAs, many of which play crucial roles in X-chromosome inactivation

Go through the lncRNA database entry for XISTs and write a brief summary about the function of this lncRNA. How many functionally important regions does it have? What is the taxonomic distribution?

Another major source of non-coding RNAs come from human annotation of genomes (e.g. HAVANA team), which identify non-coding RNAs. These normally arise when there is excellent transcript evidence, yet no apparent

protein coding evidence. Some ncRNAs are supported by literature; others are not and are annotated based on the sequenc/transcript evidence.

VEGA

STEP 1 – Go to the VEGA home page: <http://vega.sanger.ac.uk>

The screenshot shows the VEGA website interface. At the top, there is a search bar and navigation links. The main content area includes a 'Browse a genome' section with links for Mouse, Zebrafish, Pig, Human, and Rat. There is also a 'Browse a region' section with links for Tasmanian devil and Chimpanzee. A central section highlights 'Major histocompatibility complex (MHC) annotation' with a list of non-reference regions for Human, Mouse, and Pig. To the right, there is an 'Our Data' section with a list of features and datasets.

Use the VEGA website to investigate the non-coding RNA of TSIX.

STEP 2 – Enter **TSIX** onto the Vega search box and select the human gene: OTTHUMG00000184725

The screenshot shows the Vega website interface for the gene TSIX. The top navigation bar includes the Vega logo, BLAST/BLAT, Help & Documentation, and a search bar. The main content area is divided into several sections:

- Gene-based displays:** A sidebar menu with options like Summary, Splice variants, Transcript comparison, Supporting evidence, Sequence, External references, Comparative Genomics, Genomic alignments, Orthologues, Alt. alleles, External data, Personal annotation, Other genome browsers, and Ensembl.
- Gene: TSIX OTTHUMG00000184725:**
 - Description:** TSIX transcript, XIST antisense RNA
 - Synonyms:** LINC00013, NCRNA00013, XIST-AS1
 - Location:** Chromosome X: 73,792,205-73,829,231 forward strand.
 - INSDC coordinates:** chromosome:VEGA57:CM000685.2:73792205:73829231:1
 - Transcripts:** This gene has 1 transcript (splice variant) [Hide transcript table]
- Summary:**
 - Curated Locus:** TSIX (HGNC Symbol)
 - Synonyms:** LINC00013, NCRNA00013, XIST-AS1 [To view all genes linked to the name click here.]
 - Gene type:** Known lincRNA [Definition]
 - Author:** This gene was annotated by Havana <vega@sanger.ac.uk>
 - Version & date:** Version 1, last modified on 28/02/2014 (Created on 11/01/2013)
 - Other assemblies:** This gene maps to 73,792,205-73,829,231 in GRCh38 (Ensembl) coordinates. Jump to this stable ID in Ensembl
 - Curation Method:** Manual annotation from Havana
 - Alternative genes:** Ensembl gene: ENSG00000270641
- Genomic browser:** A visualization of the genomic region from 73.79Mb to 73.83Mb on chromosome X. It shows the TSIX gene (green bar) and various XIST transcripts (blue bars) on the opposite strand. Contigs are also visible at the bottom.

TSIX is the antisense counter part to XIST. Use the Vega website to investigate the supporting and external references. Which publication is associated with this Vega entry?

Use the Vega website to produce an alignment of the human and mouse XIST genes.

STEP 3 – Click on 'Genomic alignments' (left side menu)

Gene: TSIX OTTHUMG00000184725

Description: TSIX transcript, XIST antisense RNA

Synonyms: LINC00013, NCRNA00013, XIST-AS1

Location: Chromosome X: 73,792,205-73,829,231 forward strand.

INSDC coordinates: chromosome:VEGA57:CM000685.2:73792205:73829231:1

Transcripts: This gene has 1 transcript (splice variant) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
TSIX-001	OTTHUMT00000469120	37027	No protein	LincRNA	-	

Genomic alignments

Alignment: -- Select an alignment -- [Go](#)

[Download alignment](#)

No alignment specified
Please select the alignment you wish to display from the box above.

[Go to a graphical view of this alignment](#)

From the drop down list of alignments, select Mouse (Mus_musculus).

Gene: TSIX OTTHUMG00000184725

Description: TSIX transcript, XIST antisense RNA

Synonyms: LINC00013, NCRNA00013, XIST-AS1

Location: Chromosome X: 73,792,205-73,829,231 forward strand.

INSDC coordinates: chromosome:VEGA57:CM000685.2:73792205:73829231:1

Transcripts: This gene has 1 transcript (splice variant) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
TSIX-001	OTTHUMT00000469120	37027	No protein	LincRNA	-	

Genomic alignments

Alignment: Mus_musculus chromosome X [Go](#)

[Download alignment](#)

Go to a graphical view of this alignment

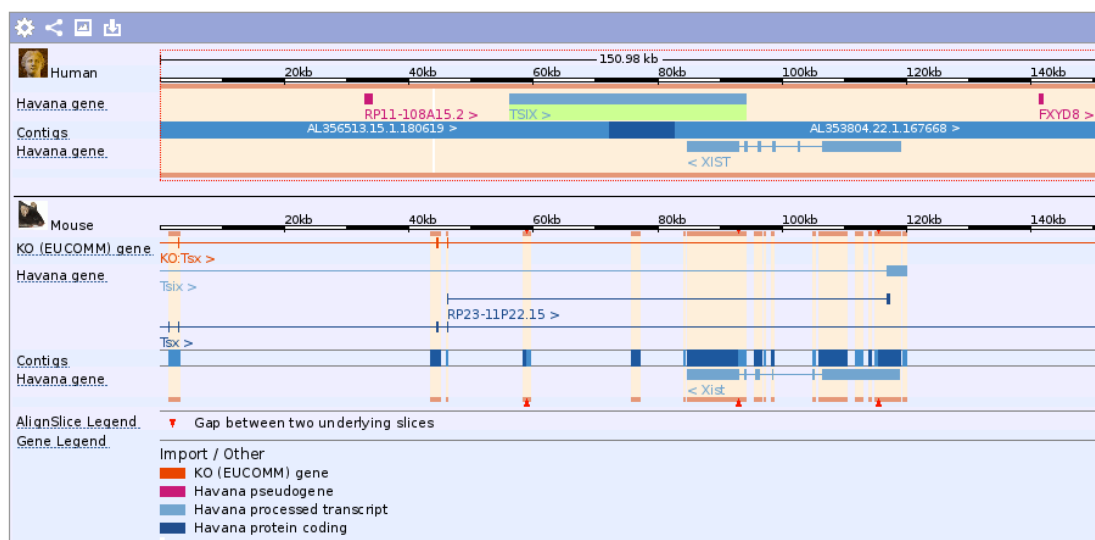
A total of 7 alignment blocks

Block	Start	End	Human	Mouse
Block 1	7469	1919	X:73820631-73828099	X:103460345-103467421
Block 2	1919	1073	X:73827353-73829271	X:103467383-103469009
Block 3	1073	541	X:73811874-73812946	X:103451662-103452672
Block 4	541	495	X:73795000-73795540	X:103450594-103451060
Block 5	495	232	X:73794506-73795000	X:103450004-103450415
Block 6	232	115	X:73819990-73820221	X:103459464-103459675
Block 7	115		X:73820664-73820778	X:9262829-9262941

Vega Genome Browser release 58 - Nov 2014 © WTSI / EBI
View in Vega release 57

Privacy policy | Contact Us | Help

From the graphical view, zoom out to see both the XIST and TSIX genes in human. Are both genes syntenic between human and mouse?



7.3 MicroRNAs and their targets

miRBase, TarBase and mircoCOSM

MicroRNAs are short regulatory RNAs that affect gene expression and translation. Currently miRBase lists nearly 700 human microRNAs, each of which might regulate tens or hundreds of protein coding transcripts (stored in microCOSM). Clearly microRNAs are important and so it is useful to understand how to get the latest information about known and predicted microRNAs and the genes that they target. MiRBase, TarBase and microCOSM are great starting places for analysis.

STEP 1 – Go to the mirBase home page
<http://www.mirbase.org>

miRBase

Home | Search | Browse | Help | Download | Blog | Submit

miRNA count: 21264 entries
Release 19: August 2012

Search by miRNA name or keyword

Download published miRNA data
[Download page](#) | [FTP site](#)

This site is featured in:
[NetWatch - Science 303:1741 \(2004\)](#)
[Highlights, Web watch - Nature Reviews Genetics 5:244 \(2004\)](#)

miRBase: the microRNA database

miRBase provides the following services:

- The [miRBase database](#) is a searchable database. Sequence database represents a predicted hairpin structure, information on the location and sequence of mature miRNAs. Sequences are available for [searching](#) and [browsing](#), and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also [available for download](#).
- The [miRBase Registry](#) provides miRNA gene hunters with unique names for novel miRNA genes prior to publication of results. Visit the [help pages](#) for more information about the naming service.

To receive email notification of data updates and feature changes please subscribe to the [miRBase announcements mailing list](#). Any queries about the website or naming service should be directed at mirbase@manchester.ac.uk.

miRBase is hosted and maintained in the [Faculty of Life Sciences](#) at the [University of Manchester](#) with funding from the [BBSRC](#), and was previously hosted and supported by the [Wellcome Trust Sanger Institute](#).

Search Results

We found **615** unique results for your query ("*mir-34*"), in **5** sections of the database.

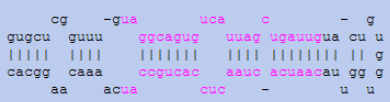
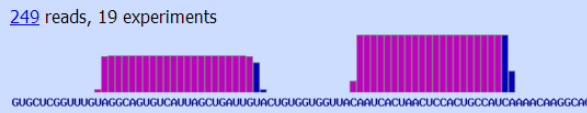
Section	Description	Number of hits
miRNA name	match the accession or ID of a hairpin precursor entry	272
Previous ID	match the previous ID of a hairpin precursor entry	14
Mature name	match the accession or ID of a mature miRNA sequence	342
Previous Mature ID	match the previous mature ID of a mature entry	77
Dead entry	match the accession or ID of a dead entry	1
Dead entry previous ID	match the accession or ID of a dead entry	0
Gene symbol	find miRNA entries based on gene symbols	0
Description	search miRNA entry description	0
Comments	search miRNA entry comments	0
PubMed ID	find miRNA entries based on literature reference PubMed ID	0
Literature reference	search title and authors of associated literature references	0

The above key shows a brief description of each of the database sections, along with the number of hits found in each one. Only unique miRNA entries are shown in the results table below. Click the column headings to sort the results table, or [restore to the original order](#).

Accession ↕	ID ↕	miRNA name ↕	Previous ID ↕	Mature name ↕	Previous Mature ID ↕	Dead entry ↕	Links
MI0000268	hsa-mir-34a	✓					
MI0000403	mmu-mir-34c	✓					
MI0000584	mmu-mir-34a	✓					
MI0000629	mo-mir-344a-1	✓					

STEP 3 - Click on link for the miRNA of interest, hsa-mir-34b (half way down)

Stem-loop sequence hsa-mir-34b

Accession	MI0000742
Symbol	HGNC:MIR34B
Description	Homo sapiens miR-34b stem-loop
Gene family	MIPF0000039; mir-34
Community annotation	<p>This text is a summary paragraph taken from the Wikipedia entry entitled mir-34 microRNA precursor family. miRBase and Rfam are facilitating community annotation of microRNA families and entries in Wikipedia. Read more...</p> <p>The mir-34 microRNA precursor family are non-coding RNA molecules that, in mammals, give rise to three major mature miRNAs. The miR-34 family members were discovered computationally and later verified experimentally. The precursor miRNA stem-loop is processed in the cytoplasm of the cell, with the predominant miR-34 mature sequence excised from the 5' arm of the hairpin. The mature miR-34a is a part of the p53 tumor suppressor network of proteins; therefore, it is hypothesized that miR-34 dysregulation is involved in the development of some cancers.</p> <p>Show Wikipedia entry View @ Wikipedia Edit Wikipedia entry</p>
Stem-loop	 <p>cg -gua uca c - g gugcu guuu ggcagug uuag ugauugua cu u g cacgg caaa ccguac acuc acuaadau gg g aa acua cuc - u u</p> <p>Get sequence</p>
Deep sequencing	<p>249 reads, 19 experiments</p>  <p>GUGCUCGGUUGUGAGGCGAGUGCAUUAAGCUGAUUGUCUGUGGUGGUUACAAUCACUAACUCACUGCCAUCAAAACAAAGGCAC</p>
Comments	Houbaviy et al. cloned 3 closely related sequences from mouse embryonic stem cells [1], and named them miR-34a, miR-34b and miR-172. These names have been remapped to miR-34c (MI0000403), miR-34b (MI0000404) and miR-34a (MI0000584) to clarify homology with human sequences. The predominant mature miRNA in human is expressed from the 3' arm (in contrast to previous annotation) [2]. Both arms express mature products in mouse.
Genome context	<p><i>Coordinates (GRCh37.p5)</i> chr11: 111383663-111383746 [+]</p> <p><i>Overlapping transcripts</i> sense ENST00000540312; AP002008.1-201; intron 1</p>
Clustered miRNAs	<p>< 10kb from hsa-mir-34b</p> <p>hsa-mir-34b chr11: 111383663-111383746 [+]</p> <p>hsa-mir-34c chr11: 111384164-111384240 [+]</p>
Database links	<p>ENTREZGENE: 407041; MIR34B</p> <p>HGNC: 31636; MIR34B</p>

Annotations:

- Predicted structure of miRNA precursor
- See genomic context with nearby genes
- See related miRNAs and make alignments

Mature sequence hsa-miR-34b-5p

Accession	MIMAT0000685
Previous IDs	hsa-miR-34b;hsa-miR-34b*
Sequence	13 - uaggcagugucuuagcugauug - 35
Deep sequencing	127 reads, 10 experiments
Evidence	by similarity; MI0000404
Validated targets	TARBASE: hsa-miR-34b-5p
Predicted targets	<p>DIANA-MICROT: hsa-miR-34b-5p</p> <p>MICRORNA.ORG: hsa-miR-34b-5p</p> <p>MIRDB: hsa-miR-34b-5p</p> <p>RNA22-HSA: hsa-miR-34b-5p</p> <p>TARGETMINER: hsa-miR-34b-5p</p> <p>PICTAR-VERT: hsa-miR-34b</p>

Annotation:

- Information about predicted targets by different algorithms

The top two sections contain a wide variety of information on the microRNA including comments on previous nomenclature of the sequences. You can

also see what the predicted targets are for this miRNA with a variety of software systems.

Mature sequence hsa-miR-34b-3p	
Accession	MIMAT0004676
Previous IDs	hsa-miR-34b
Sequence	50 - caaucauaacucuccacugooau - 71 Get sequence
Deep sequencing	122 reads, 13 experiments
Evidence	experimental; cloned [2]
Validated targets	TARBASE: hsa-miR-34b-3p
Predicted targets	DIANA-MICROT: hsa-miR-34b-3p MICRORNA.ORG: hsa-miR-34b-3p MIRDB: hsa-miR-34b-3p RNA22-HSA: hsa-miR-34b-3p TARGETMINER: hsa-miR-34b-3p TARGETSCAN-VERT: hsa-miR-34b PICTAR-VERT: hsa-miR-34b

References	
1	PMID: 12919684 "Embryonic stem cell-specific MicroRNAs" Houbaviy HB, Murray MF, Sharp PA Dev Cell. 5:351-358(2003).
2	PMID: 17604727 "A mammalian microRNA expression atlas based on small RNA library sequencing" Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foa R, Schliwka J, Fuchs U, Novosel A, Muller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M Cell. 129:1401-1414(2007).

The two mature sequence sections show that there is evidence that a mature miRNA is expressed on the miR* strand, which means both arms of the precursor might become mature miRNAs.

Finally a wide variety of useful search options are provided linked from the search tab at the top of the page.

The screenshot shows the miRBase website navigation bar with tabs: Home, Search, Browse, Help, Download, Blog, Submit, and hsa-mir-34b. Below the navigation bar is the 'Search miRBase' section with several search options: 'By miRNA identifier or keyword', 'By genomic location', 'For clusters', 'By tissue expression', and 'By sequence'. A yellow callout box with a black border points to the 'Search' tab in the navigation bar, containing the text 'STEP 4 - Click on search tab'.

This interface allows you to carry out several useful analyses quickly, such as find all the microRNAs on human chromosome 1. You can also find clusters of

microRNAs that are close on a genome and might be expressed from a single transcript.

From the mirBase example shown above, there can be both validated and predicted targets. In the following worked examples, we will look at two resources, one that contains validated targets and one that provides computational predictions. TarBase contains validated targets, where as MicroCOSM has computational predictions.

TarBase worked example – in this example, find the genes that are regulated by mmu-miR-20a-5p.

STEP 1 - Go to TarBase

<http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index>

The screenshot shows the DIANA TOOLS website. The header includes logos for the European Union, NSRF (National Strategic Reference Framework) 2007-2013, and the Alexander Fleming Biomedical Sciences Research Center (IMIS). The main navigation bar has links for HOME, SOFTWARE, PUBLICATIONS, and CONTACT. The left sidebar contains a login form with fields for Username and Password, a 'Remember me next time' checkbox, and a 'Login' button. Below the login form are links for 'Forgot your password?' and 'Sign up for free!' or 'take a tour'. The main content area features an 'IMPORTANT NOTE' about a power upgrade on Saturday, December 13. Below this is a search bar with a magnifying glass icon and a help icon. The main text welcomes users to DIANA-TarBase v7.0 and provides a detailed description of the database, including its history and the number of entries. On the right side, there is a 'Filters' section with dropdown menus for Species, Method Type, Method, and Source.

STEP 2 – enter mmu-miR-20a-5p in the search textfield.

The resulting page indicates the different genes that are regulated by the miRNA and the methods that have been used to establish the target of the microRNA. Uses the filters to determine which gene is ‘up’ regulated by this microRNA (Regulation type)?

STEP 3 – Select the Stat3 gene (right down arrow)

Stat3 (mmu)		mmu-miR-20a-5p		RS	qP	WB	O	0.906
Publication	Methods	Tissue	Cell line	Tested cell line	Exp. condition			
Gianni Carraro et al. 2009	RS	Breast Cancerous Tissues	NA	N/A	N/A			
Location	Method	Result	Regulation	Valid. type	Source			
3UTR	Luciferase Reporter Assay	POSITIVE	↓	DIRECT	Tarbase 7.0			
3UTR	Luciferase Reporter Assay	POSITIVE	↓	DIRECT	Tarbase 7.0			
Gianni Carraro et al. 2009	qP WB O	NA	NA	N/A	N/A			
Location	Method	Result	Regulation	Valid. type	Source			
UNKNOWN	qPCR	POSITIVE	↓	INDIRECT	Tarbase 7.0			
UNKNOWN	Western Blot	POSITIVE	↓	INDIRECT	Tarbase 7.0			
UNKNOWN	Other	POSITIVE	↓	INDIRECT	Tarbase 7.0			

This lists the details of the publication and method uses to establish the microRNA target.

MicroCOSM worked example - In this example you will be identifying microRNAs that might regulate your protein of interest. In this case we'll look at the important cancer gene P53.

STEP 1 - Go to microCOSM <http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/>

All miRNA hits for *Homo sapiens* where search terms are 2 hits found.

Gene Name	Transcript	Description	GO Terms	Score	Energy	P-value	Length	Total Sites	No. Cons. Species	No. miRNAs
TP53	ENST00000359597	Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13). [Source:Uniprot/SWISSPROT,Acc:P04637]		264	-351	9.142e-05	1000	16	1	14
TP53	ENST00000269305	Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13). [Source:Uniprot/SWISSPROT,Acc:P04637]		288	-355	0.000272854	1188	18	3	23

STEP 4 -
Click on view

Two ensembl transcripts are found.

Hit information for ENST00000359597

Gene Name	TP53
Transcript	ENST00000359597
Gene	ENSG00000141510
Description	Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13). [Source:Uniprot/SWISSPROT,Acc:P04637]

Mouse over the targets for more information

Organisation of 3' UTR, coloured blocks show predicted miRNA binding sites

Alignment of miRNA sequence to predicted miRNA binding sites in 3' UTR

Rfam ID	Score	Energy	Base P	Poisson P	Org P	Start	End	Alignment
mmu-miR-709	18.0408	-30.84	4.132880e-02	3.850650e-03	3.850650e-03	591	619	UGUAAGGUGGAGGUC UUGUAAGGUGGAGGUC
hsa-miR-30b*	17.5717	-29.14	3.916900e-02	3.841180e-02	3.841180e-02	782	803	UUGUAAGGUGGAGGUC UUGUAAGGUGGAGGUC
mmu-miR-709	17.0157	-29	9.043240e-02	3.850650e-03	3.850650e-03	973	991	UGUAAGGUGGAGGUC UGUAAGGUGGAGGUC
hsa-miR-92a	16.9658	-30.57	1.883850e-02	1.883850e-02	1.883850e-02	748	767	UGUAAGGUGGAGGUC UGUAAGGUGGAGGUC

The results of the microCOSOM predictions are shown above. The prediction of microRNA target sites remains a difficult problem and all prediction methods are prone to false positives. So these require extensive manual inspection to decide if they are likely to be important. There are a small number of known microRNA binding sites

Recently David Corney *et al.* published a paper in Cancer Research suggesting that miR34b and miR34c are regulated by P53. These two microRNAs were not predicted to regulate P53.

7.4 A database of non-coding RNA sequences

Up until 2014, there was no centralised repository for RNA sequences. Prior to this, researchers would have to go to each individual specialist database to get the information. However, these databases do not reflect all non-coding information and further information may have been directly deposited in the ENA database (part of the INSDC). To provide help to this situation, RNACentral was set up to bring non-coding RNA sequence data from different databases. To avoid duplication of sequences in RNACentral, the database groups identical sequences into a single entry, and assign an **Unique RNA Sequence identifier**, regardless of source database or organism.

STEP 1 - Go to the RNACentral home page <http://rnacentral.org>

STEP 2 - Entry URS00000478B7 into the search box

This shows the RNACentral page for URS00000478B7.

Unique RNA Sequence URS00000478B7 Interactive tour

A unique RNA sequence entry in RNACentral groups together all identical RNA sequences no matter what species they are from.

Overview Taxonomy 2D 3D Download ▾

Overview

Description: Homo sapiens SRP_RNA
 299 nucleotides 5 databases (ENA, lncRNAdb, RefSeq, Rfam, SRPDB) 1 organism first seen 29 May 2014 last updated 25 Jul 2014

Annotations 1-5 of 5 Filter table

Database	Description	Species
SRPDB	Homo sapiens (human) signal recognition particle RNA > SRPDB: Homo.sapi._X01037 > Source ENA entry: HG323706.1:1..299:ncRNA	Homo sapiens
lncRNAdb	Homo sapiens (human) Small nucleolar RNA 7SL > lncRNAdb: 7SL > Source ENA entry: HG975378.1:1..299:ncRNA	Homo sapiens
Rfam	Homo sapiens Metazoan signal recognition particle RNA > RFAM family: RF00017 (Metazoa_SRP), seed alignment > Source ENA entry: X01037.1 (nucleotides 5:303)	Homo sapiens
Rfam	Homo sapiens Metazoan signal recognition particle RNA > RFAM family: RF00017 (Metazoa_SRP), seed alignment > Source ENA entry: X04248.1 (nucleotides 1:299)	Homo sapiens
RefSeq	Homo sapiens RNA, 7SL, cytoplasmic 1 (RN7SL) > RefSeq: NR_002715.1 > NCBI GeneID: 6029 > HGNC gene RN7SL1	

Sequence

299 nucleotides (56 A; 83 C; 105 G; 55 U, 0 N)

```

GCCGGGCGCGGUGGGCGCGUGCCUGUAGUCCAGCUACUCGGAGGCUGAGGCUGGAGGAGUCUCUGGAGCCAGGAGUCUGGGCGUGAGGCGCCAGCCCGGCGGG
UCCGACUAAGUUCGGCAUCAUAUGGUGACCUCGCCGGAGCGGGGACCACCAGGUUGCCUAAGGAGGGGUGAACCGGCCAGGUCGGAACGGAGCAGGUCAAAAUCU
CCCGUCUGAUCAGUAGUGGAUCGCGCCUGGAAUAGCCACUGCACUCCAGCCUGGGCAACAUAGCGAGACCCCGUCUCU
    
```

This shows the SRP RNA from human from 5 different source databases. In this instance they all agree. Note, a single URS identifier can contain more than one species.

But if you did not know the URS identifier for the human how might you find all SRP RNAs where these five databases agree? RNACentral has a very powerful faceted search interface, where you can select different features to quickly drill down to the sequences of interest.

STEP 3 – Enter “RNA” in the search text box

This will provide a list of *all* sequences contained in RNACentral.

The screenshot shows the RNAcentral search interface. At the top, the RNAcentral logo is on the left, and a search bar contains the text 'RNA'. Below the search bar, there are navigation links: 'v1.0 Expert databases - API - Sequence search', 'Downloads', 'Help', and 'Contact'. The main content area displays 'Results 15 out of 8,102,559 sequences'. On the left, there are two panels: 'Expert databases' and 'RNA types'. The 'Expert databases' panel lists various databases with checkboxes: ENA (6,984,057), Rfam (2,493,782), RefSeq (30,900), VEGA (27,317), gtRNAdb (10,625), miRBase (8,795), RDP (4,779), tmRNA Website (2,857), SRPDB (503), and lncRNAdb (62). The 'RNA types' panel lists various RNA types with checkboxes: rRNA (5,612,511), misc RNA (1,111,150), tRNA (818,026), piRNA (208,933), other (129,140), miRNA (92,000), snRNA (90,503), snoRNA (80,526), siRNA (45,059), hammerhead ribozyme (40,210), lncRNA (40,139), SRP RNA (14,375), and precursor RNA (13,014). The main results area lists 15 sequences, each with its name, accession number, and nucleotide count. The first result is 'Prochlorococcus marinus subsp. misc RNA/RNase P RNA URS0000532385' with 308 nucleotides. Other results include 'Prochlorococcus marinus subsp. misc RNA/RNase P RNA URS0000560E5A' (333 nucleotides), 'Synechococcus sp. misc RNA/RNase P RNA URS000004F88D' (305 nucleotides), 'misc RNA/RNase P RNA/other from 270 species URS00004BB8BB' (377 nucleotides), 'Prochlorococcus marinus subsp. misc RNA/RNase P RNA URS00005CC4EE' (311 nucleotides), 'Dickeya dadantii RNase P RNA URS00002070F6' (381 nucleotides), 'Prochlorococcus marinus str. PAC1A misc RNA/RNase P RNA URS000042D60B' (333 nucleotides), 'Prochlorococcus marinus str. PAC1B misc RNA/RNase P RNA URS00004572C6' (334 nucleotides), and 'Prochlorococcus marinus str. TAK9803-2 misc RNA/RNase P RNA URS000018913C' (308 nucleotides).

Use the left panel to refine the search

STEP 4 – Select ‘SRP RNA’ and the five expert databases (not PDBe)

The screenshot shows the RNAcentral search interface after refining the search. The search bar now contains 'SRP RNA'. The main content area displays 'Results 1 sequence'. On the left, the 'Expert databases' panel shows five databases selected with checkboxes: ENA (1), Rfam (1), RefSeq (1), SRPDB (1), and lncRNAdb (1). The 'RNA types' panel shows 'SRP RNA (1)' selected. The 'Organisms' panel is empty. The main results area lists one sequence: 'Homo sapiens SRP RNA URS00000478B7' with 299 nucleotides.

There is only one sequence in RNAcentral where these 5 databases all agree on an SRP RNA annotation.

For some organisms, it is possible to view the non-coding RNAs in genomic context. Using the search interface, search for the following:

STEP 5 – Enter 'HOTAIR' into the search box and select the expert database VEGA

The screenshot shows the RNAcentral search interface. The search bar contains the text "HOTAIR AND expert_db:'VEGA'". Below the search bar, there are navigation links for "v1.0 Expert databases", "API", and "Sequence search". The search results section displays "Results 12 out of 12 sequences". On the left, under "Expert databases", the "VEGA (12)" option is selected. The top result is "Homo sapiens (human) long non-coding RNA OTTHUMT00000328665.1 (HOTAIR gene)" with accession number "URS000011D1F0" and "562 nucleotides".

STEP 6 – Select to top match (URS000011D1F0)

The screenshot shows the RNAcentral search bar with the placeholder text "organism, expert database, gene, ncRNA type, accession". Below the search bar, there are navigation links for "v1.0 Expert databases", "API", and "Sequence search".

Unique RNA Sequence URS000011D1F0

[Interactive tour](#)

A unique RNA sequence entry in RNAcentral groups together all identical RNA sequences no matter what species they are from.

Overview [Taxonomy](#) [2D](#) [3D](#) [Download](#)

Overview

Description: Homo sapiens (human) long non-coding RNA OTTHUMT00000328665.1 (HOTAIR gene)
562 nucleotides **2 databases** (ENA, VEGA) **1 organism** first seen 29 May 2014 last updated 05 Sep 2014

Annotations 1-1 of 1

Database	Description	Species
Vega (GENCODE)	<p>Homo sapiens (human) long non-coding RNA OTTHUMT00000328665.1 (HOTAIR gene)</p> <p>> Vega transcript OTTHUMT00000328665 from gene OTTHUMG00000152934</p> <p>> 4 alternative transcripts: URS00000513030 (560 nts), URS000019A694 (572 nts), URS00001A335C (918 nts), URS0000301B08 (2,421 nts).</p> <p>> Source ENA entry: HG504802.1:1..562:ncRNA</p> <p>> View genomic location 12:53,963,901-53,967,355 Ensembl UCSC</p>	Homo sapiens

Sequence

STEP 7 – Click on 'View genomic location'

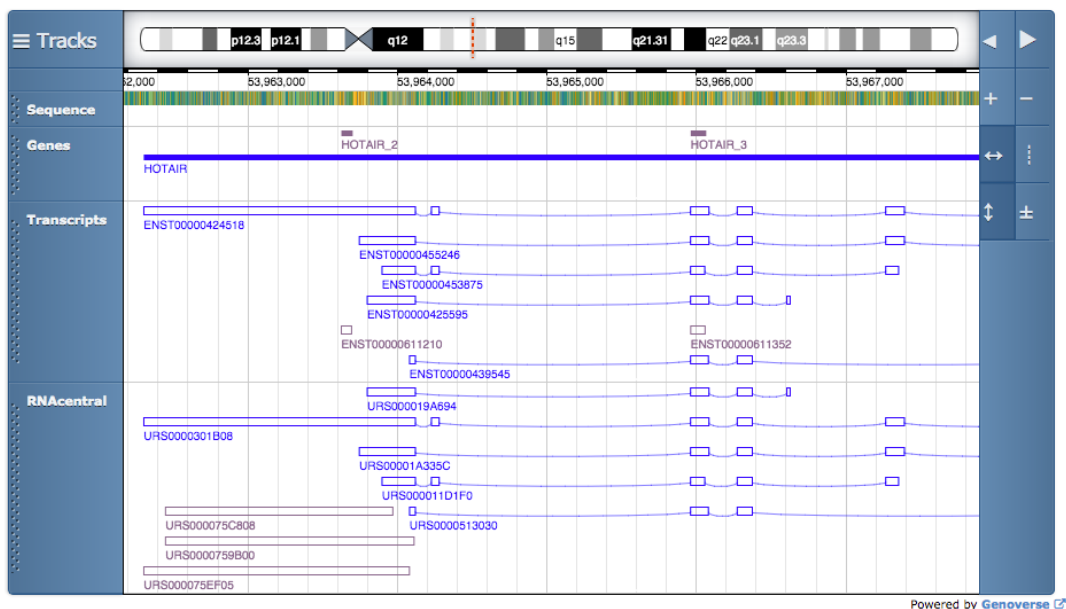
562 nucleotides (165 A; 136 C; 10 G; 11 U; 2 T)

```
AGACCAACACCCUUCUCCUGGGGCUCCACCCGGGACUAGACCCUCAGGUCCUAAUUAUCCCGGAGGUGUCUCAUUCAGAAAGGUCCUGUCUCCGCUUCGAGUGG
AAUGGAAACGGAUUUAGAAGCCUUCAGUAGGGGAGUUGGGGAGUGGAGAGAGAGGCCAGCCAGAGUUACAGACGGCGGCGAGAGGCCCAAAGAGUCUGAUGUUACAAGACC
AGAAAUGCCACGGCCGUCUUCGUCAGAGAAAAGGCUGAAAUGGAGGACCGGCGCCUUCUUUAUAAGUAUUCACAUUGGCGAGAGAAGUGUCGCAACCAAACAGCAA
UUACACCCAAAGCUCGUUUGGGCCUUAAGCCAGUACCGACCUGGUAAGAAAAGCAACCAGCAAGCUAGAGAGAGAGCCAGAGGAGGGAAGAGAGCGCCAGAGCAAGGUGAA
AGCGAACCCACGCAGAGAAAUGCAGGCAAGGGAGCAAGGCCGAGUUCGGAAACAACGUGGCAGAGGGCAAGACGGGCACUCACAGACAGAGGUUUUAUGUAUUUUUAU
UUUUUUAAAUCUGAUUU
```


This launches a basic genome browser and allows the RNACentral annotations to be displayed along side the gene annotations and transcript data from Ensembl.

Genome browser *Homo sapiens* 12:53,963,901-53,967,355

Homo sapiens (human) long non-coding RNA OTTHUMT00000328665.1 (HOTAIR gene)

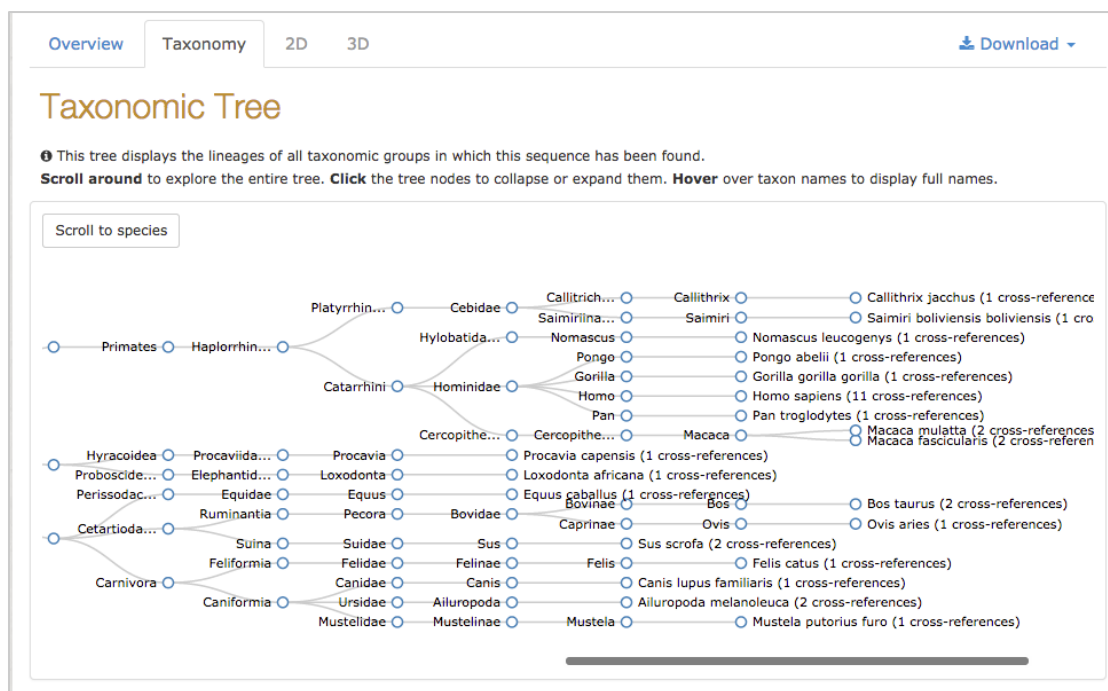


Use this browser to identify the genes either side of the long HOTAIR gene.

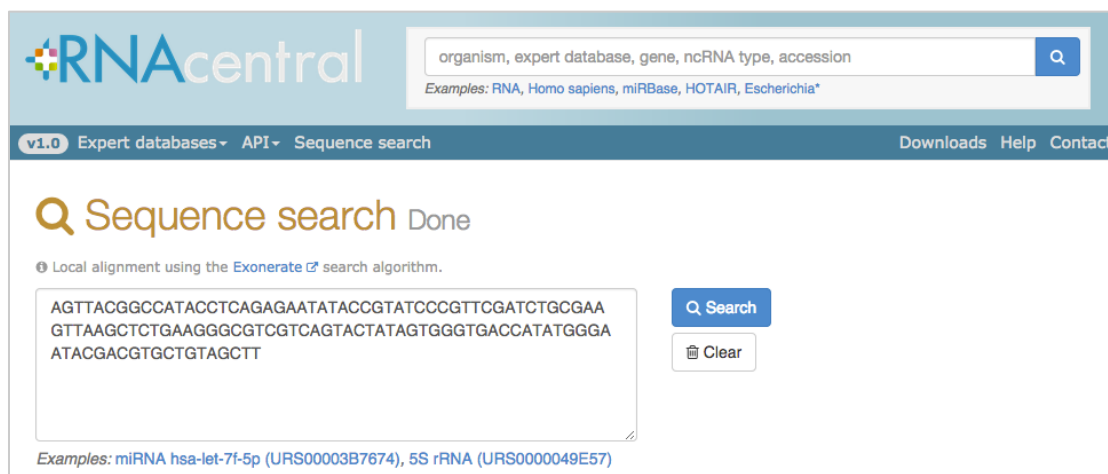
To illustrate how RNACentral groups identical sequences in a single entry go to the URS000047C79B entry.

STEP 7 – Enter URS000047C79B in the search box

How many species have been grouped together in this entry? Click on the taxonomy tab to show the distribution of these species.



Finally, it is also possible to search RNACentral using a query sequence. Click on 'Sequence search' in the menu:



Paste the following sequence into the sequence search box:

>Query

```
AGTTACGGCCATACCTCAGAGAATATACCGTATCCCGTTCGATCTGCGAA
GTTAAGCTCTGAAGGGCGTCGTCAGTACTATAGTGGGTGACCATATGGGA
ATACGACGTGCTGTAGCTT
```

How many alignments are returned? What ncRNA does this sequence represent?