RNAcentral

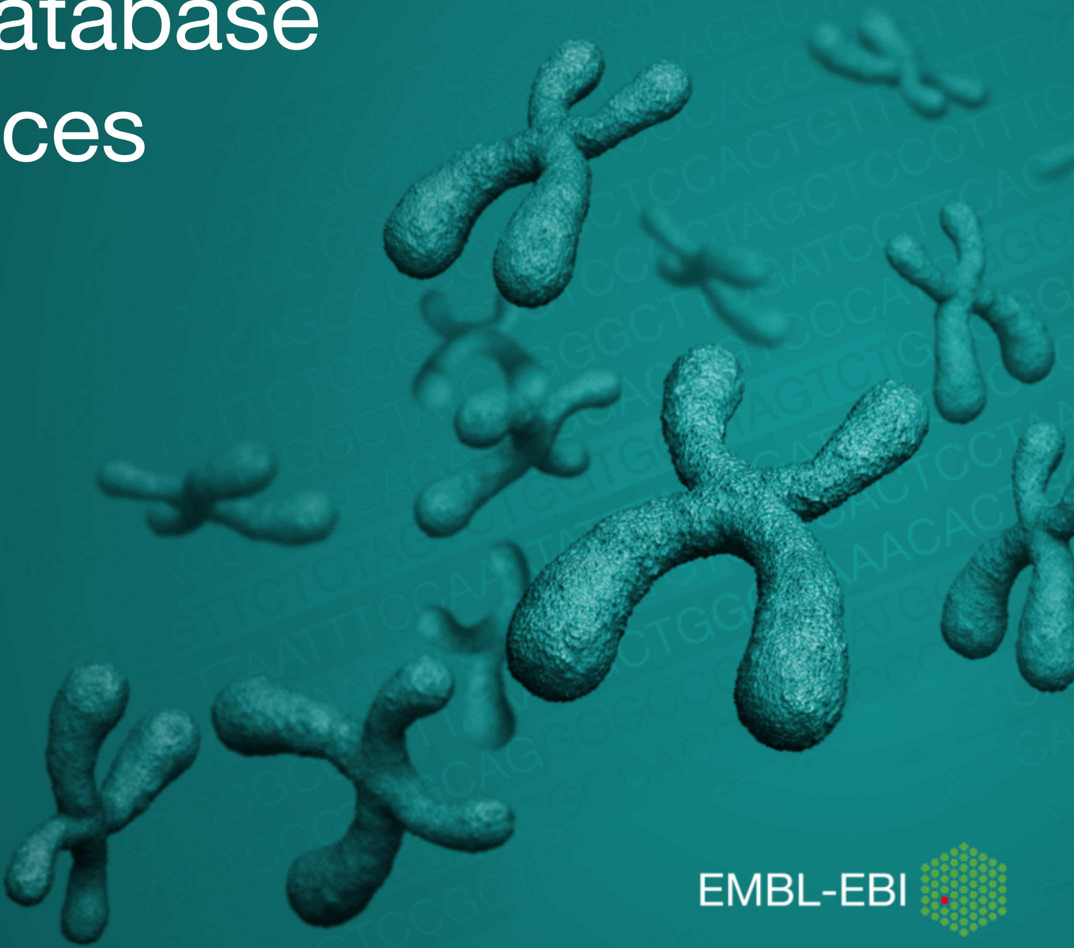
an International Database of ncRNA Sequences

Anton Petrov

apetrov@ebi.ac.uk

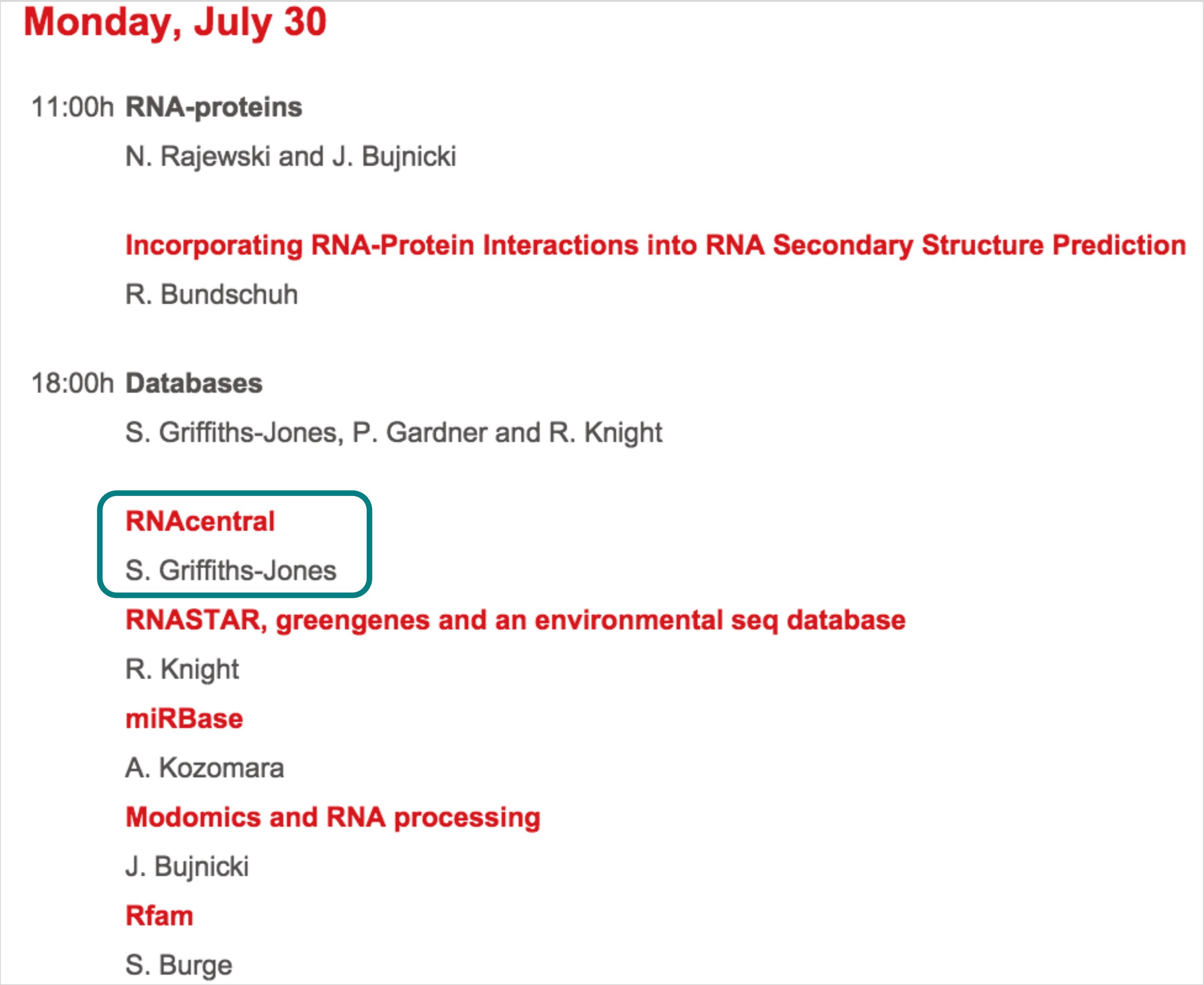
Benasque

July 27th, 2015



EMBL-EBI

RNAcentral at Benasque 2012



Monday, July 30

11:00h **RNA-proteins**

N. Rajewski and J. Bujnicki

Incorporating RNA-Protein Interactions into RNA Secondary Structure Prediction

R. Bundschuh

18:00h **Databases**

S. Griffiths-Jones, P. Gardner and R. Knight

RNAcentral

S. Griffiths-Jones

RNASTAR, greengenes and an environmental seq database

R. Knight

miRBase

A. Kozomara

Modomics and RNA processing

J. Bujnicki

Rfam

S. Burge

http://benasque.org/2012rna/cgi-bin/talks/allprint.pl

Why do we need RNAcentral?

Before RNAcentral:

- lots of specialized databases
- no single entry point

  for ncRNA sequence analysis
- lack of standard identifiers

http://www.officesignspro.com/Funny-Road-Signs-2/

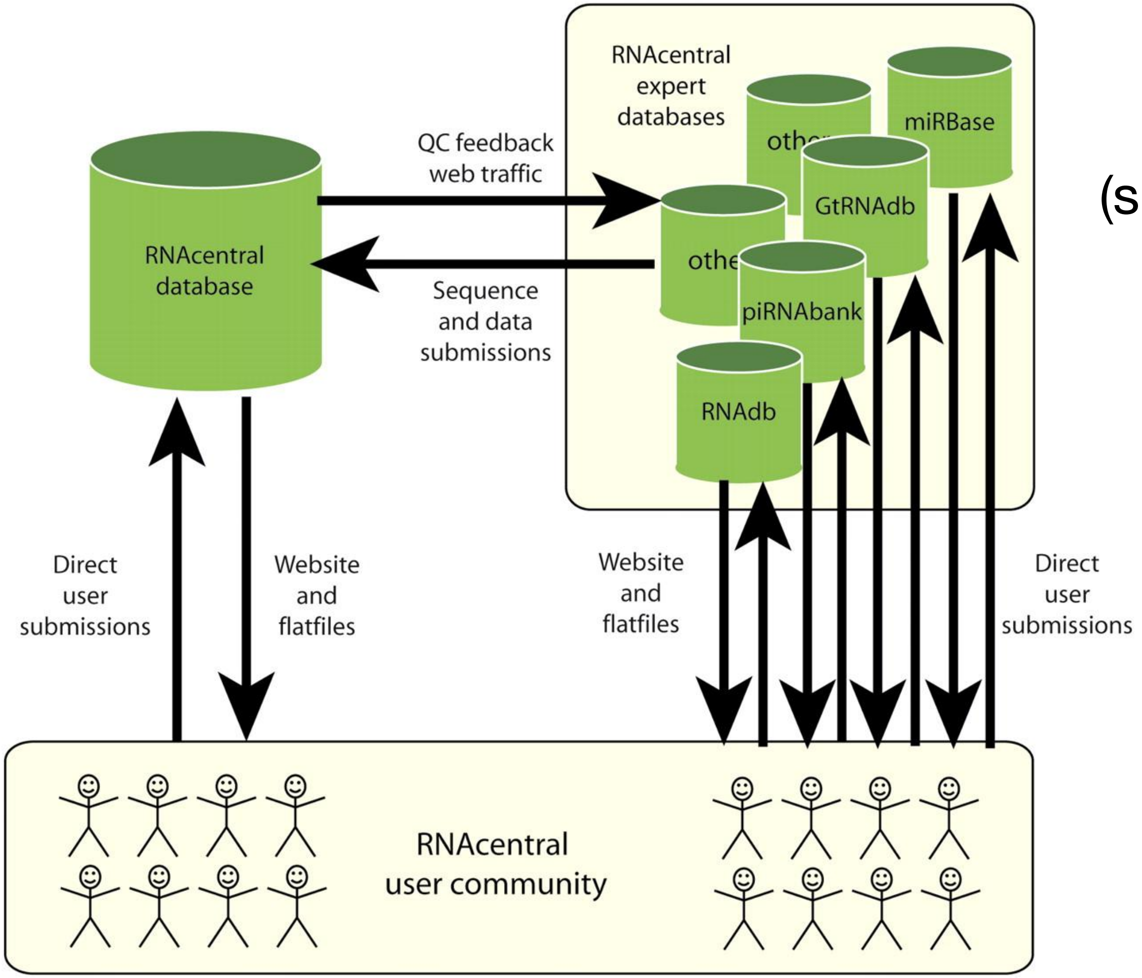
What is RNAcentral?

- RNAcentral is a **comprehensive** and **up-to-date** database of **accessioned** ncRNA sequences that collates and integrates information from an international consortium of established RNA sequence databases.
- RNAcentral provides **broad coverage** of ncRNA types and the taxonomic space.
- Four releases since **June 2014**.

What does RNAcentral provide?

- **unified access** to data from multiple sources
- **stable identifiers** for distinct RNA sequences
- **cross-references** to other databases
- sequence and metadata **search**
- **API** for programmatic data access
- **FTP** archive

Where does the data come from?



Expert Databases

(such as miRBase or Vega)

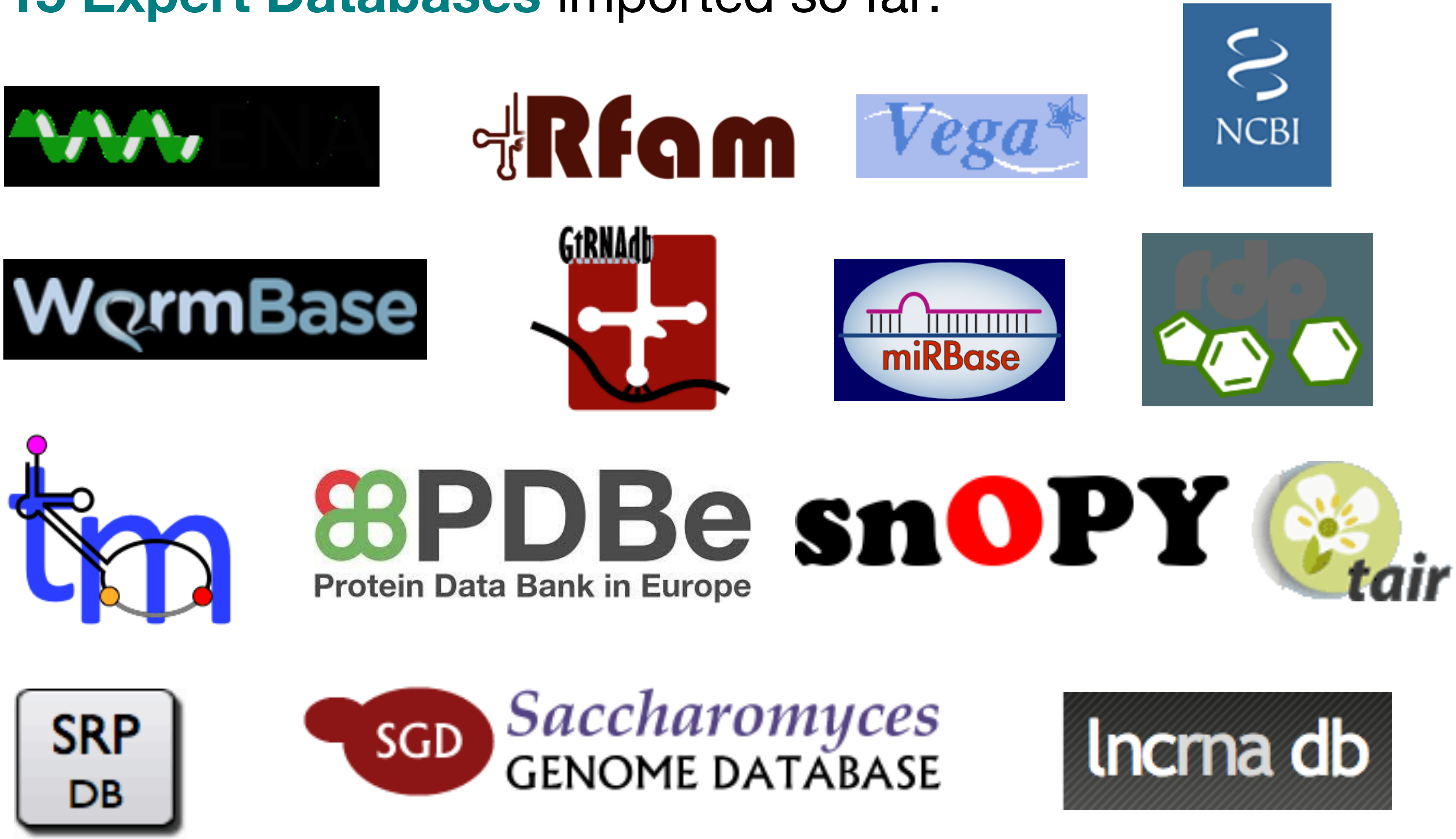
supply data to

RNAcentral.

Bateman et al., 2011

What data is in RNAcentral?

15 Expert Databases imported so far:



> 20 more Expert Databases to import

Comparative RNA Web Site

FlyBase

GREENGENES
The 16S rRNA Gene Database and Tools

HGNC
HUGO Gene Nomenclature Committee

tmR DB

LNCipedia.org

Modomics
a database of RNA modification pathways

NONCODE

NPInter

piRNABank

sRNAMap

PLncDB:Plant Long noncoding RNA Database

PomBase

snoRNA Database

RNApathwaysDB

tRNAdb
Transfer RNA database

silva
high quality ribosomal RNA databases

RGD

MGI
CELEBRATING 25 YEARS

TarBase

LncBase

Full list of RNAcentral Expert Databases:
http://rnacentral.org/expert-databases

Data from **INSDC** is imported automatically

International Nucleotide Sequence Database Collaboration =

ENA European Nucleotide Archive, EMBL-EBI +

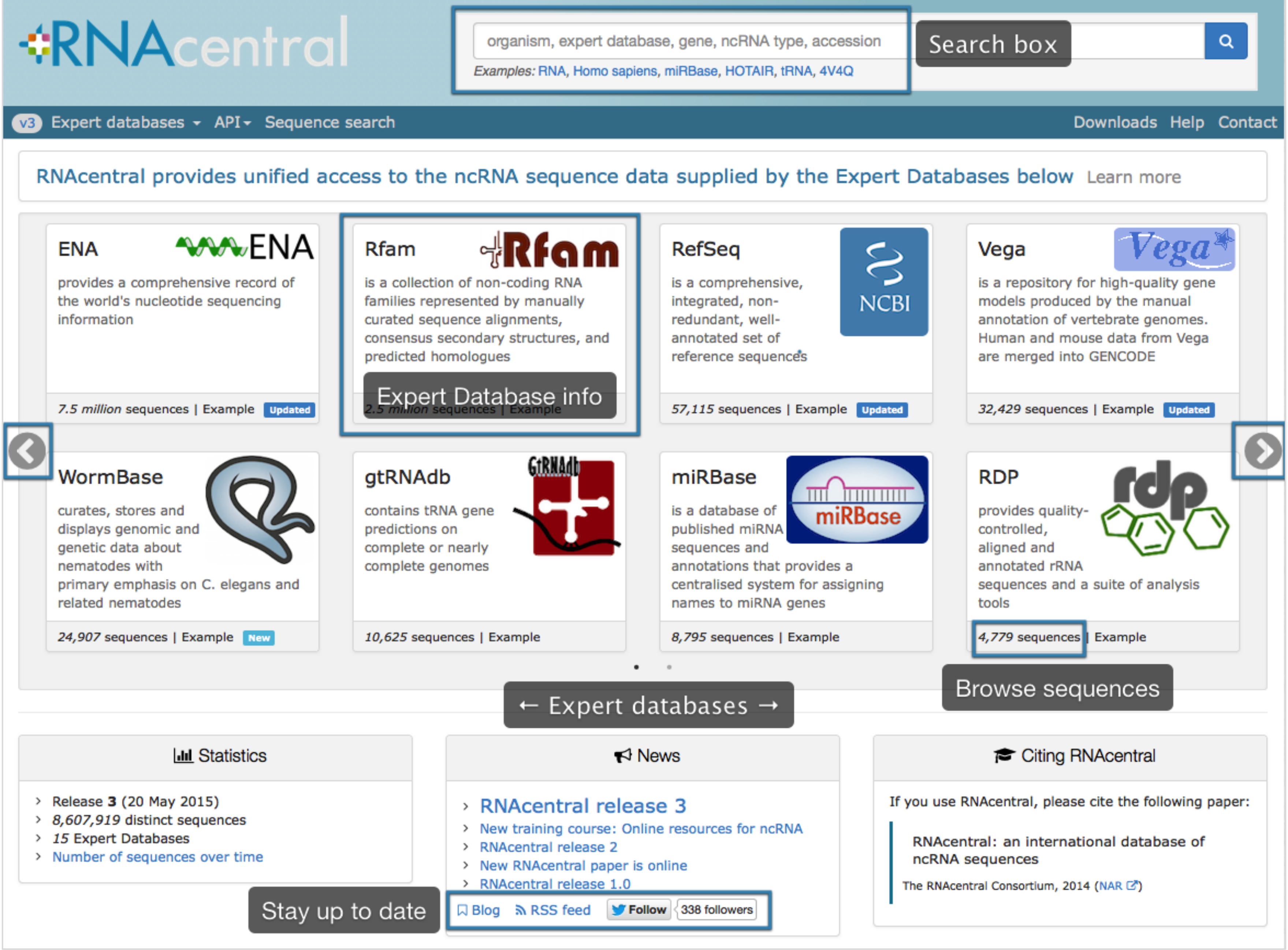
GenBank NCBI +

DDBJ DNA Data Bank of Japan

http://www.insdc.org/

Demo

http://rnacentral.org



RNAcentral

organism, expert database, gene, ncRNA type, accession

Examples: RNA, Homo sapiens, miRBase, HOTAIR, tRNA, 4V4Q

RNAcentral provides unified access to the ncRNA sequence data supplied by the Expert Databases below Learn more

ENA
provides a comprehensive record of the world's nucleotide sequencing information
7.5 million sequences | Example Updated

Rfam
is a collection of non-coding RNA families represented by manually curated sequence alignments, consensus secondary structures, and predicted homologues
2.5 million sequences | Example

RefSeq
is a comprehensive, integrated, non-redundant, well-annotated set of reference sequences
57,115 sequences | Example Updated

Vega
is a repository for high-quality gene models produced by the manual annotation of vertebrate genomes. Human and mouse data from Vega are merged into GENCODE
32,429 sequences | Example Updated

WormBase
curates, stores and displays genomic and genetic data about nematodes with primary emphasis on C. elegans and related nematodes
24,907 sequences | Example New

gtRNAdb
contains tRNA gene predictions on complete or nearly complete genomes
10,625 sequences | Example

miRBase
is a database of published miRNA sequences and annotations that provides a centralised system for assigning names to miRNA genes
8,795 sequences | Example

RDP
provides quality-controlled, aligned and annotated rRNA sequences and a suite of analysis tools
4,779 sequences | Example

Statistics
- Release **3** (20 May 2015)
- *8,607,919* distinct sequences
- *15* Expert Databases
- Number of sequences over time

News
- RNAcentral release 3
- New training course: Online resources for ncRNA
- RNAcentral release 2
- New RNAcentral paper is online
- RNAcentral release 1.0

Blog RSS feed Follow 338 followers

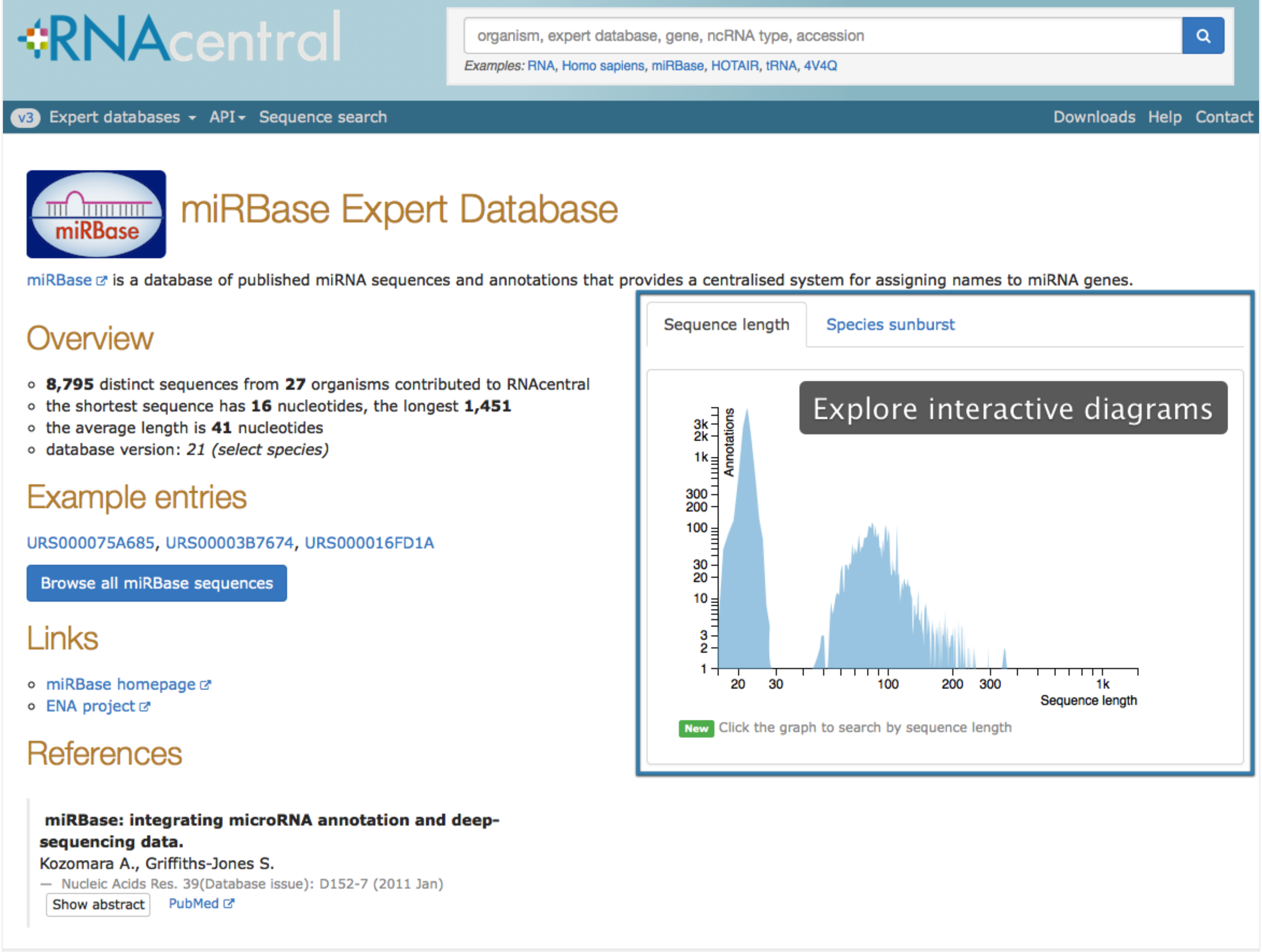
Citing RNAcentral

If you use RNAcentral, please cite the following paper:

> RNAcentral: an international database of ncRNA sequences
>
> The RNAcentral Consortium, 2014 (NAR)

RNAcentral homepage: http://rnacentral.org/

EMBL-EBI



RNAcentral
organism, expert database, gene, ncRNA type, accession
Examples: RNA, Homo sapiens, miRBase, HOTAIR, tRNA, 4V4Q
v3 Expert databases API Sequence search
Downloads Help Contact
miRBase

miRBase Expert Database

miRBase is a database of published miRNA sequences and annotations that provides a centralised system for assigning names to miRNA genes.

Overview

- **8,795** distinct sequences from **27** organisms contributed to RNAcentral
- the shortest sequence has **16** nucleotides, the longest **1,451**
- the average length is **41** nucleotides
- database version: *21 (select species)*

Example entries

URS000075A685, URS00003B7674, URS000016FD1A

Browse all miRBase sequences

Links

- miRBase homepage
- ENA project

References

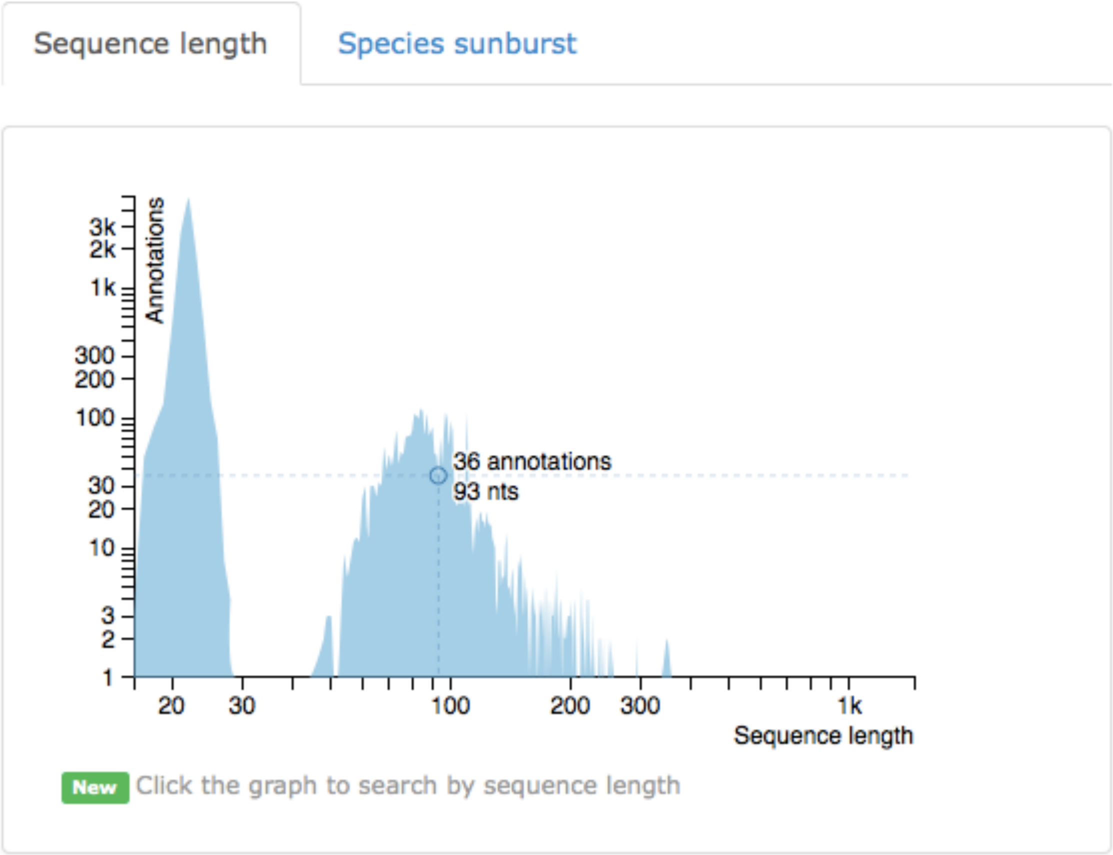
miRBase: integrating microRNA annotation and deep-sequencing data.
Kozomara A., Griffiths-Jones S.
— Nucleic Acids Res. 39(Database issue): D152-7 (2011 Jan)
Show abstract PubMed

Click the graph to search by sequence length

Example Expert Database page: http://rnacentral.org/expert-database/mirbase

EMBL-EBI

Exploring sequences by length

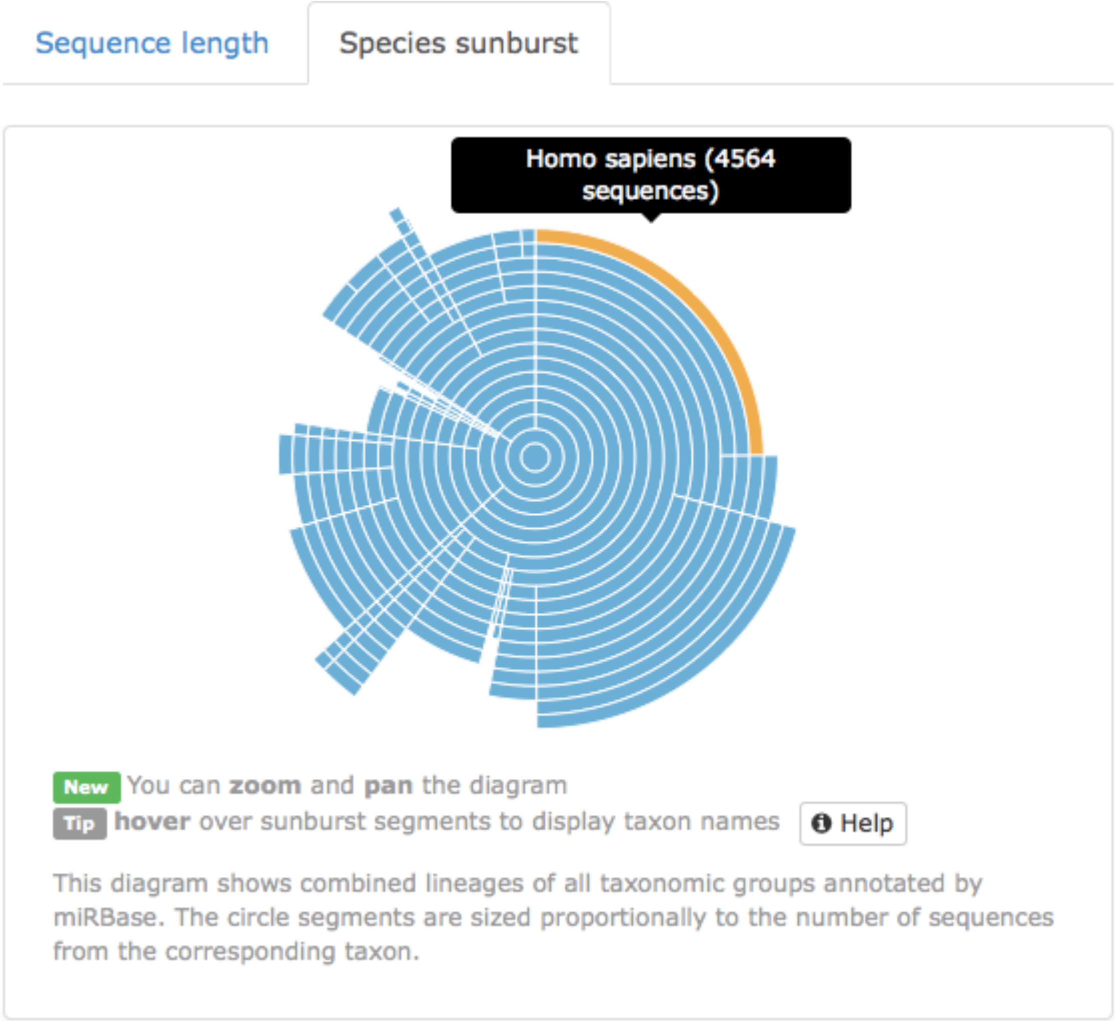


Click the graph to search by sequence length

You can explore sequence length distribution and **launch searches** using the interactive graph.

Example Expert Database page: http://rnacentral.org/expert-database/mirbase

Exploring species distribution



You can **zoom** and **pan** the diagram

hover over sunburst segments to display taxon names

This diagram shows combined lineages of all taxonomic groups annotated by miRBase. The circle segments are sized proportionally to the number of sequences from the corresponding taxon.

You can view what species the data comes from using the interactive **sunburst diagram**.

Expert Database example page: http://rnacentral.org/expert-database/mirbase

Unique RNA Sequence identifiers

Each distinct sequence gets a unique RNAcentral identifier regardless of what species it is coming from.

- Format: **URS + 10-digit hexadecimal number**
- Example: URS00000B15DA
- Sequences must be **at least 10 nucleotides** long
  - ~ 8.6 million ids assigned so far
  - > 1 trillion possible ids

Species-specific identifiers

RNAcentral also provides **species-specific identifiers**.

- Format: **URS identifier / NCBI taxid**
- Examples:
  - URS00003B7674**/9606**

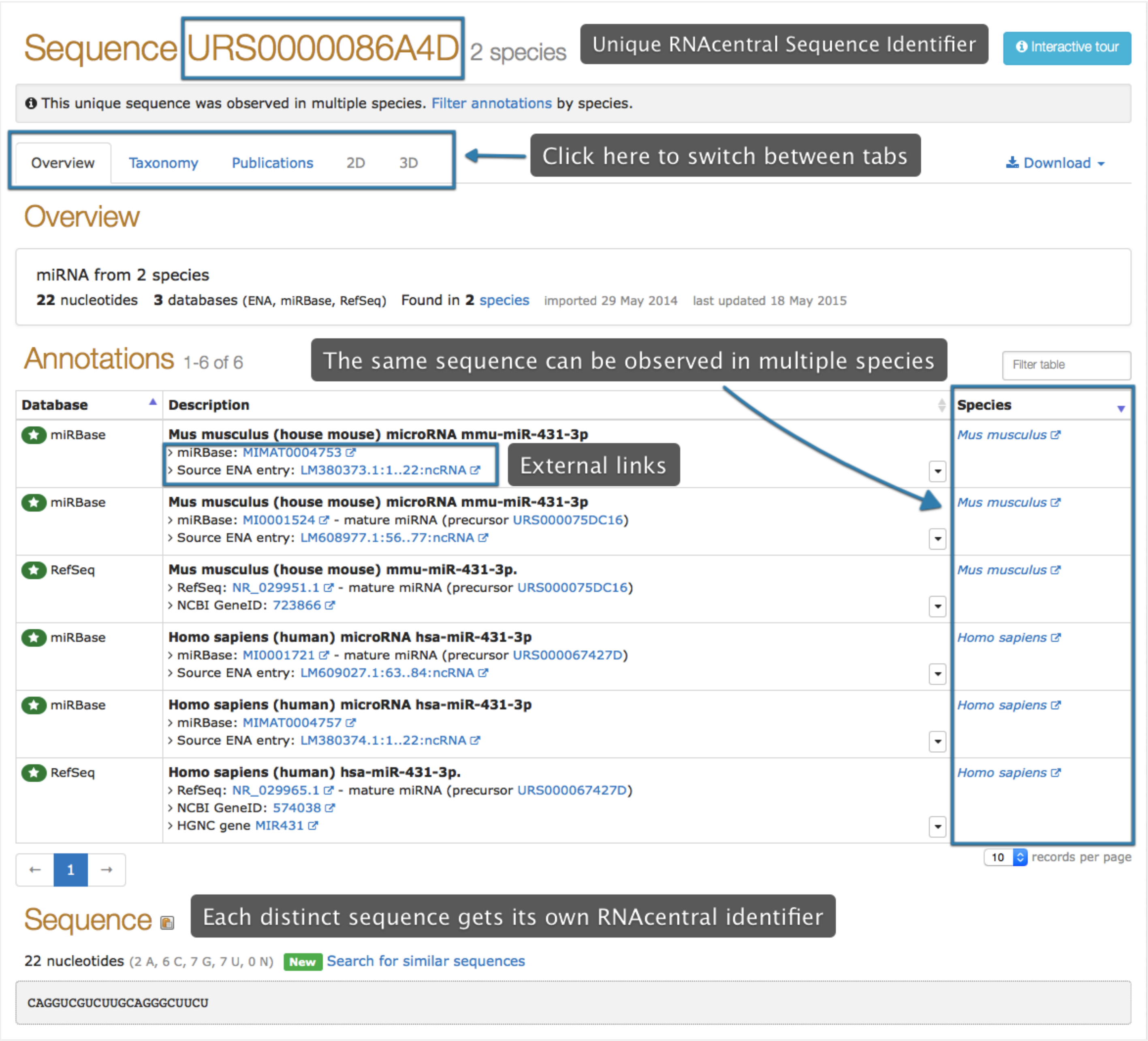
    can also use **underscore** instead of **slash**

Gene Ontology uses RNAcentral IDs for annotating miRNAs:

http://amigo.geneontology.org/amigo/gene_product/RNAcentral:URS00004C9052_9606

Expert species-specific page:
http://rnacentral.org/rna/URS00003B7674/9606

Demo



Sequence URS0000086A4D 2 species

Unique RNAcentral Sequence Identifier

This unique sequence was observed in multiple species. Filter annotations by species.

Click here to switch between tabs

Overview

miRNA from 2 species

22 nucleotides **3** databases (ENA, miRBase, RefSeq) Found in **2** species imported 29 May 2014 last updated 18 May 2015

Annotations 1-6 of 6

The same sequence can be observed in multiple species

Database	Description	Species
miRBase	**Mus musculus (house mouse) microRNA mmu-miR-431-3p** › miRBase: MIMAT0004753 › Source ENA entry: LM380373.1:1..22:ncRNA	*Mus musculus*
miRBase	**Mus musculus (house mouse) microRNA mmu-miR-431-3p** › miRBase: MI0001524 - mature miRNA (precursor URS000075DC16) › Source ENA entry: LM608977.1:56..77:ncRNA	*Mus musculus*
RefSeq	**Mus musculus (house mouse) mmu-miR-431-3p.** › RefSeq: NR_029951.1 - mature miRNA (precursor URS000075DC16) › NCBI GeneID: 723866	*Mus musculus*
miRBase	**Homo sapiens (human) microRNA hsa-miR-431-3p** › miRBase: MI0001721 - mature miRNA (precursor URS000067427D) › Source ENA entry: LM609027.1:63..84:ncRNA	*Homo sapiens*
miRBase	**Homo sapiens (human) microRNA hsa-miR-431-3p** › miRBase: MIMAT0004757 › Source ENA entry: LM380374.1:1..22:ncRNA	*Homo sapiens*
RefSeq	**Homo sapiens (human) hsa-miR-431-3p.** › RefSeq: NR_029965.1 - mature miRNA (precursor URS000067427D) › NCBI GeneID: 574038 › HGNC gene MIR431	*Homo sapiens*

External links

Sequence

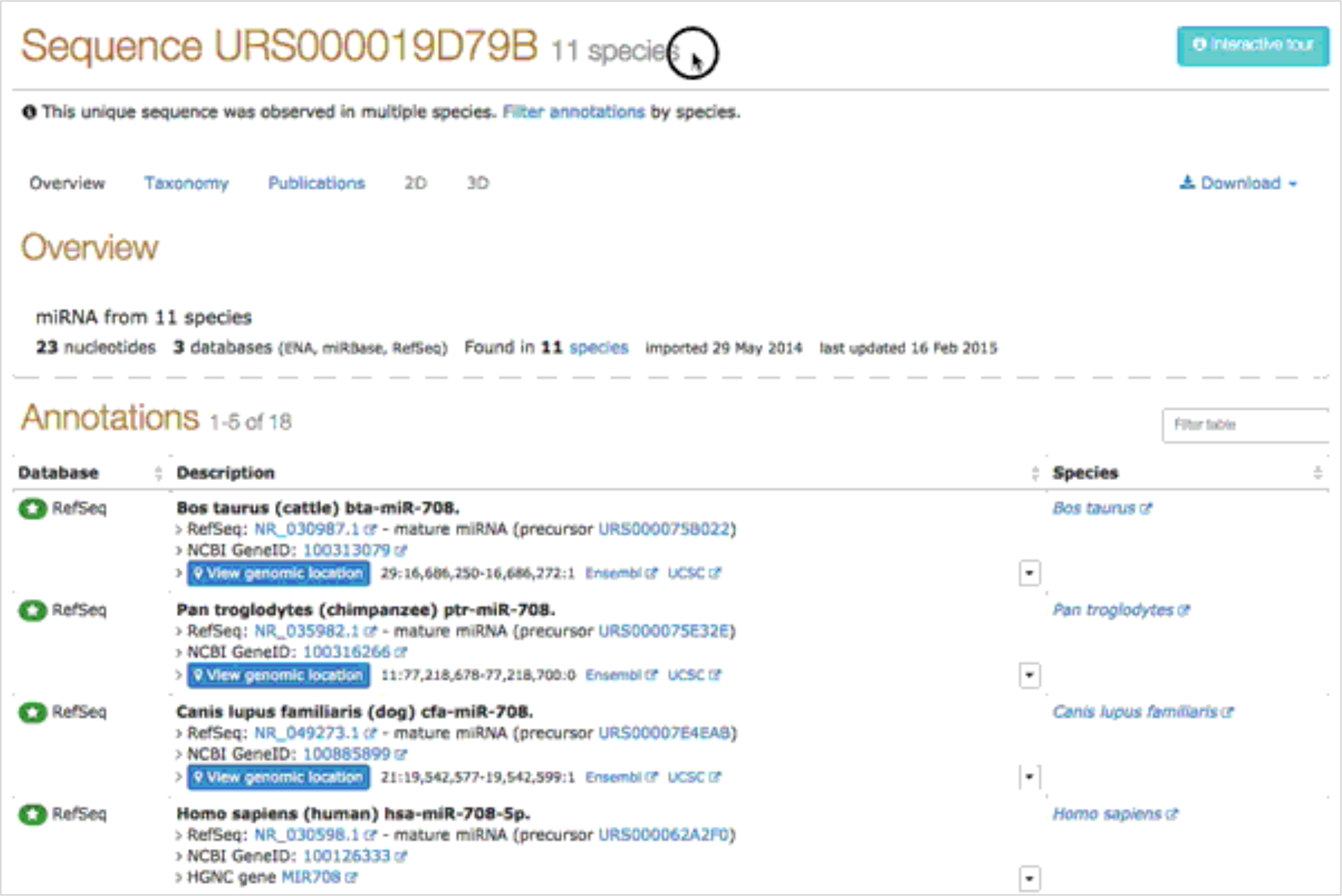
Each distinct sequence gets its own RNAcentral identifier

22 nucleotides (2 A, 6 C, 7 G, 7 U, 0 N) New Search for similar sequences

```
CAGGUCGUCUUGCAGGGCUUCU
```

Example RNAcentral sequence page:
http://rnacentral.org/rna/URS0000086A4D

Switching between species using taxonomic tree



Sequence URS000019D79B 11 species

This unique sequence was observed in multiple species. Filter annotations by species.

Overview

miRNA from 11 species

23 nucleotides **3** databases (ENA, miRBase, RefSeq) Found in **11** species imported 29 May 2014 last updated 16 Feb 2015

Annotations 1-5 of 18

Database	Description	Species
RefSeq	**Bos taurus (cattle) bta-miR-708.** › RefSeq: NR_030987.1 - mature miRNA (precursor URS000075B022) › NCBI GeneID: 100313079 › View genomic location 29:16,686,250-16,686,272:1 Ensembl UCSC	*Bos taurus*
RefSeq	**Pan troglodytes (chimpanzee) ptr-miR-708.** › RefSeq: NR_035982.1 - mature miRNA (precursor URS000075E32E) › NCBI GeneID: 100316266 › View genomic location 11:77,218,678-77,218,700:0 Ensembl UCSC	*Pan troglodytes*
RefSeq	**Canis lupus familiaris (dog) cfa-miR-708.** › RefSeq: NR_049273.1 - mature miRNA (precursor URS00007E4EAB) › NCBI GeneID: 100885899 › View genomic location 21:19,542,577-19,542,599:1 Ensembl UCSC	*Canis lupus familiaris*
RefSeq	**Homo sapiens (human) hsa-miR-708-5p.** › RefSeq: NR_030598.1 - mature miRNA (precursor URS000062A2F0) › NCBI GeneID: 100126333 › HGNC gene MIR708	*Homo sapiens*

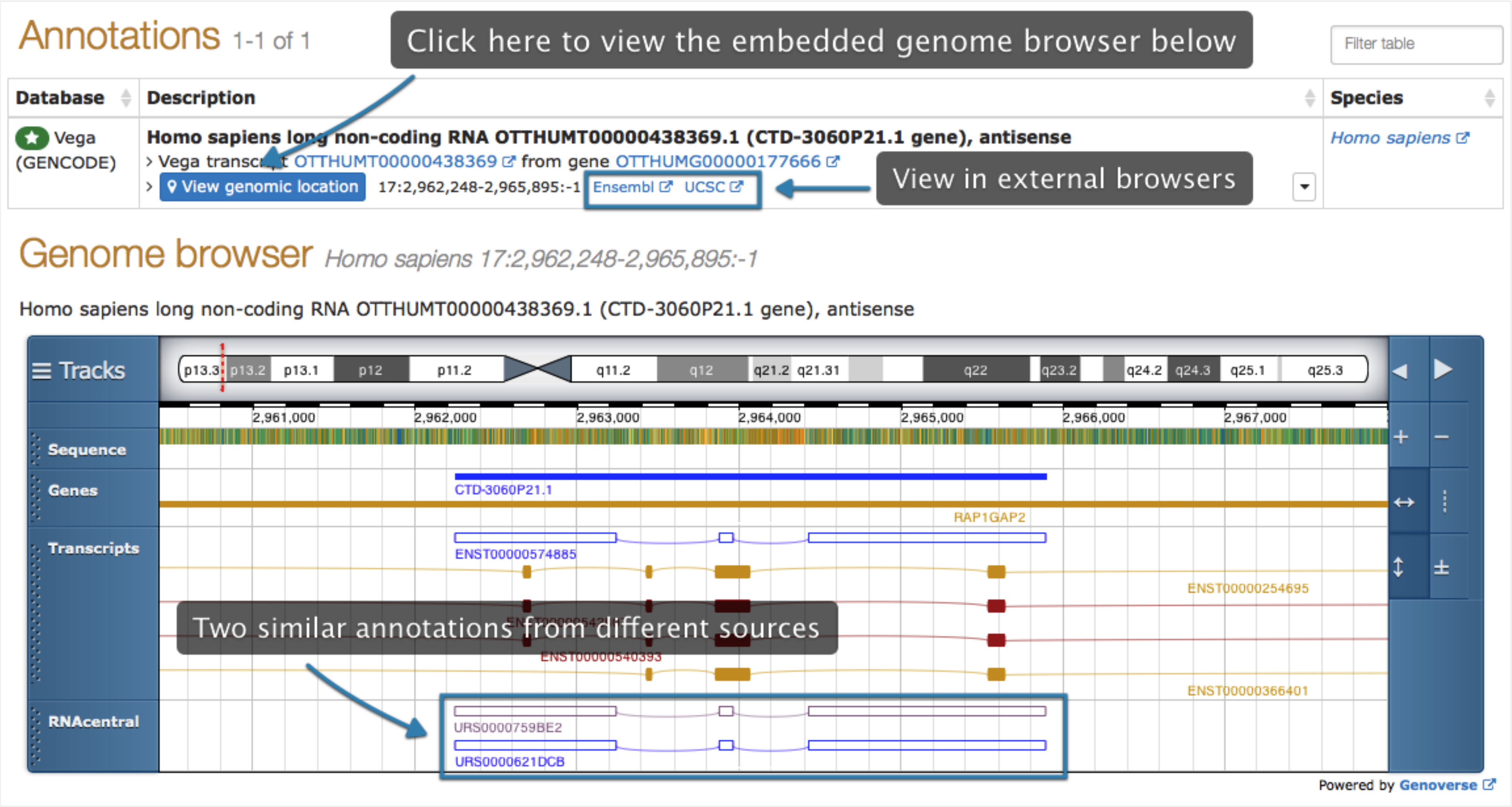
View animated figure at:
http://blog.rnacentral.org/2015/06/rnacentral-release-2.html

Genomic mapping

Many entries in RNAcentral come from **reference genomes**.

These entries can be viewed in their genomic context using an **embedded genome browser** and their coordinates can be downloaded in GFF/GFF3/BED formats.

Viewing entries in their genomic context



Annotations 1-1 of 1

Click here to view the embedded genome browser below

Filter table

Database	Description	Species
Vega (GENCODE)	**Homo sapiens long non-coding RNA OTTHUMT00000438369.1 (CTD-3060P21.1 gene), antisense** › Vega transcript OTTHUMT00000438369 from gene OTTHUMG00000177666 › View genomic location 17:2,962,248-2,965,895:-1 Ensembl UCSC	*Homo sapiens*

View in external browsers

Genome browser *Homo sapiens 17:2,962,248-2,965,895:-1*

Homo sapiens long non-coding RNA OTTHUMT00000438369.1 (CTD-3060P21.1 gene), antisense

Two similar annotations from different sources

Powered by Genoverse

RNA sequence example page: http://rnacentral.org/rna/URS0000621DCB

Metadata search

- **faceting** helps to explore and filter the data
- **advanced search**

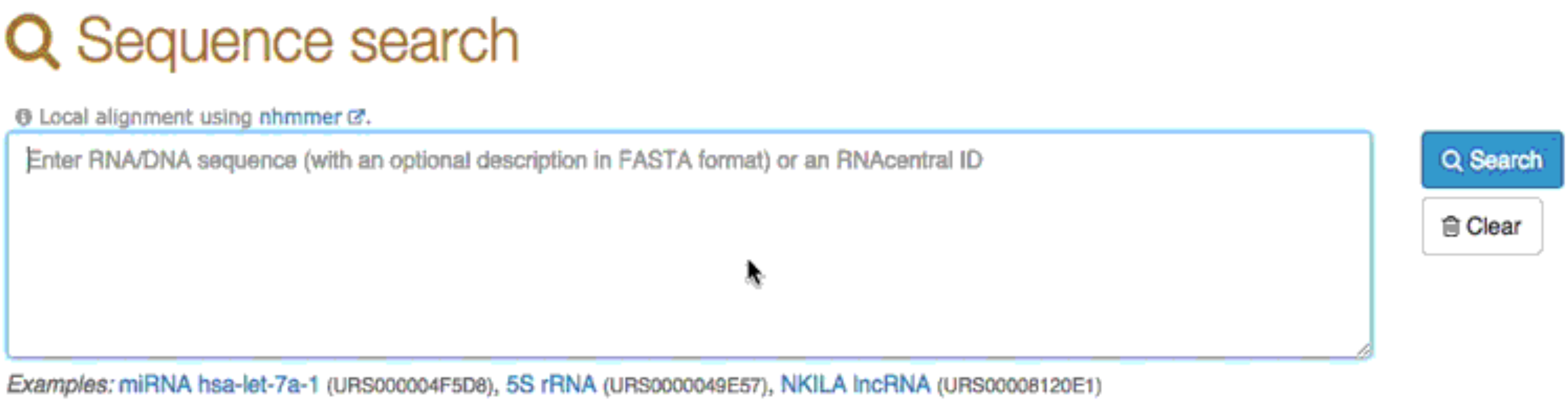
  logical operators, field-specific search and more
- results can be **exported** in several formats

Sequence search

- powered by ***nhmmer***
- results are stored for **7 days** and can be accessed using **unique URLs**
- results can be **sorted** by sequence identity, coverage etc

http://rnacentral.org/sequence-search

Instant retrieval of exact sequence matches



Sequence search

Local alignment using nhmmer.

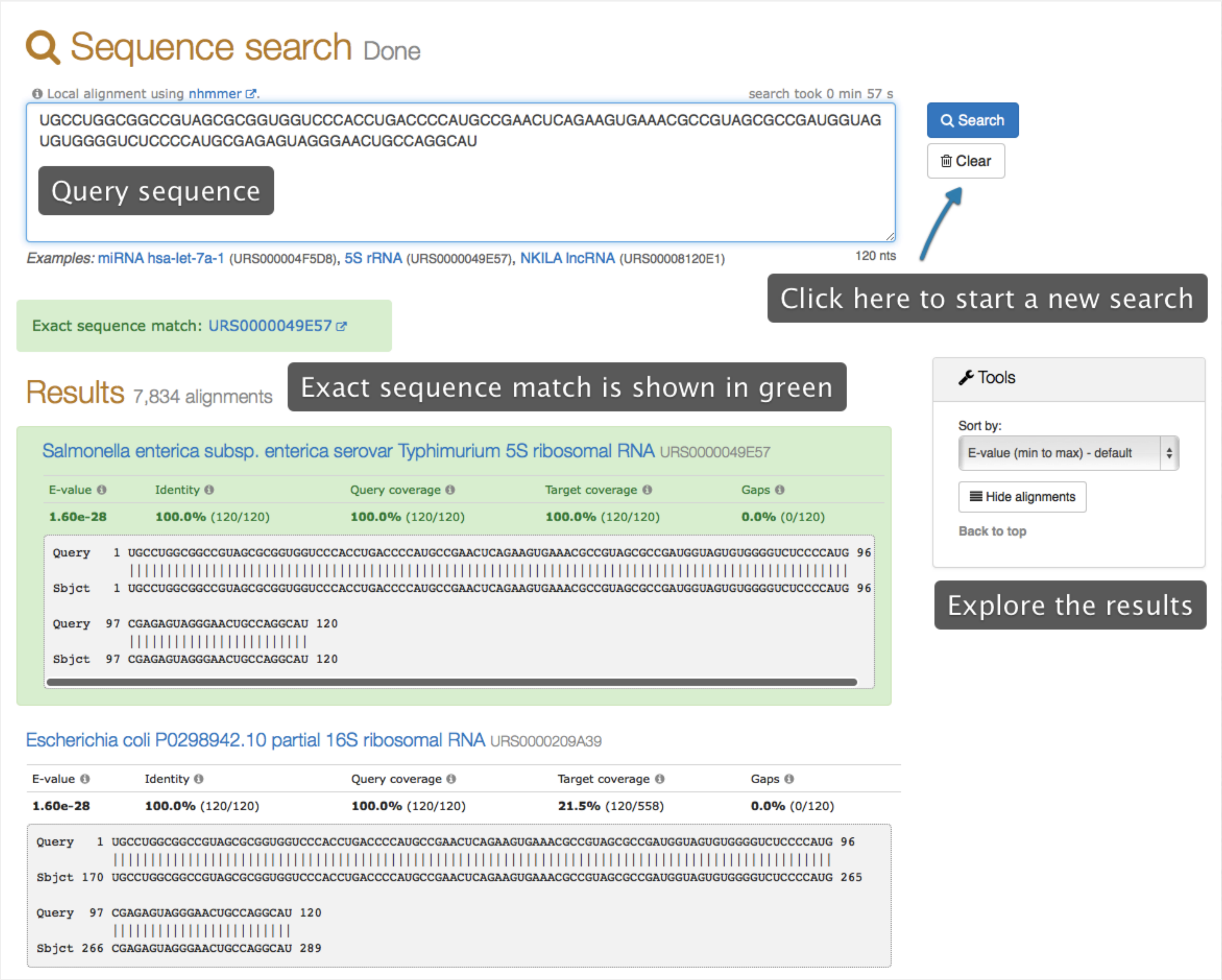
Enter RNA/DNA sequence (with an optional description in FASTA format) or an RNAcentral ID

Examples: miRNA hsa-let-7a-1 (URS000004F5D8), 5S rRNA (URS0000049E57), NKILA lncRNA (URS00008120E1)

- The **exact sequence match** is retrieved instantly (*if it exists*)
- You can **cancel** the search if you only need the exact sequence match

View animated figure at:
http://blog.rnacentral.org/2015/06/rnacentral-release-3.html

EMBL-EBI



Sequence search Done

Local alignment using nhmmer. search took 0 min 57 s

UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAG
UGUGGGGUCUCCCCAUGCGAGAGUAGGAACUGCCAGGCAU

Query sequence

Examples: miRNA hsa-let-7a-1 (URS000004F5D8), 5S rRNA (URS0000049E57), NKILA lncRNA (URS00008120E1) 120 nts

Click here to start a new search

Exact sequence match: URS0000049E57

Results 7,834 alignments

Exact sequence match is shown in green

Explore the results

Salmonella enterica subsp. enterica serovar Typhimurium 5S ribosomal RNA URS0000049E57

E-value	Identity	Query coverage	Target coverage	Gaps
1.60e-28	**100.0%** (120/120)	**100.0%** (120/120)	**100.0%** (120/120)	**0.0%** (0/120)

```
Query   1  UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUG  96
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct   1  UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUG  96

Query  97  CGAGAGUAGGAACUGCCAGGCAU  120
           |||||||||||||||||||||||
Sbjct  97  CGAGAGUAGGAACUGCCAGGCAU  120
```

Escherichia coli P0298942.10 partial 16S ribosomal RNA URS0000209A39

E-value	Identity	Query coverage	Target coverage	Gaps
1.60e-28	**100.0%** (120/120)	**100.0%** (120/120)	**21.5%** (120/558)	**0.0%** (0/120)

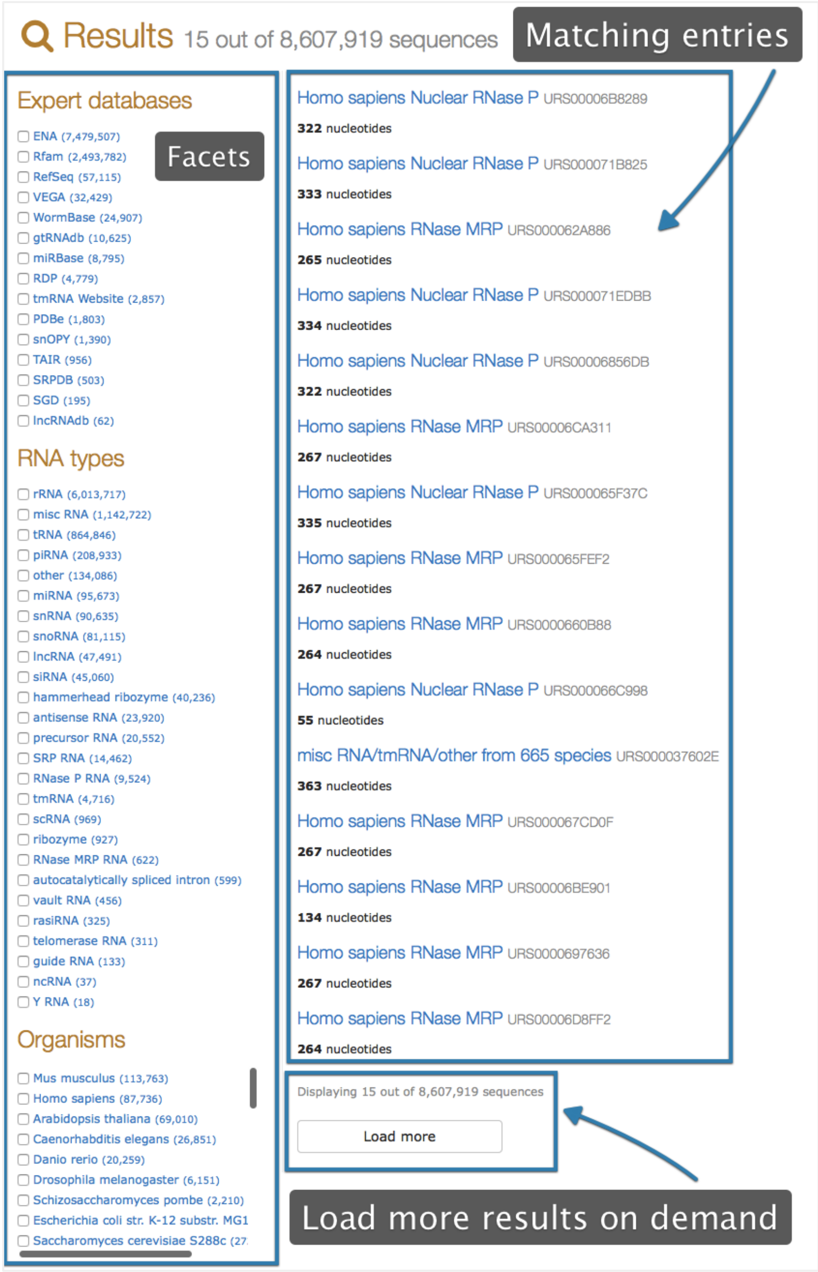
```
Query    1  UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUG  96
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  170  UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUG  265

Query   97  CGAGAGUAGGAACUGCCAGGCAU  120
            |||||||||||||||||||||||
Sbjct  266  CGAGAGUAGGAACUGCCAGGCAU  289
```

Example sequence search result:
http://rnacentral.org/sequence-search/?id=07a325aa-c909-4c8f-a7a5-aee1e553ec1b

Demo

Example search result

- RNAcentral search is fast and intuitive
- by default all metadata associated with all entries is searched
- one can construct specific searches using the query syntax



Results 15 out of 8,607,919 sequences

Matching entries

Facets

Expert databases

- □ ENA (7,479,507)
- □ Rfam (2,493,782)
- □ RefSeq (57,115)
- □ VEGA (32,429)
- □ WormBase (24,907)
- □ gtRNAdb (10,625)
- □ miRBase (8,795)
- □ RDP (4,779)
- □ tmRNA Website (2,857)
- □ PDBe (1,803)
- □ snOPY (1,390)
- □ TAIR (956)
- □ SRPDB (503)
- □ SGD (195)
- □ lncRNAdb (62)

RNA types

- □ rRNA (6,013,717)
- □ misc RNA (1,142,722)
- □ tRNA (864,846)
- □ piRNA (208,933)
- □ other (134,086)
- □ miRNA (95,673)
- □ snRNA (90,635)
- □ snoRNA (81,115)
- □ lncRNA (47,491)
- □ siRNA (45,060)
- □ hammerhead ribozyme (40,236)
- □ antisense RNA (23,920)
- □ precursor RNA (20,552)
- □ SRP RNA (14,462)
- □ RNase P RNA (9,524)
- □ tmRNA (4,716)
- □ scRNA (969)
- □ ribozyme (927)
- □ RNase MRP RNA (622)
- □ autocatalytically spliced intron (599)
- □ vault RNA (456)
- □ rasiRNA (325)
- □ telomerase RNA (311)
- □ guide RNA (133)
- □ ncRNA (37)
- □ Y RNA (18)

Organisms

- □ Mus musculus (113,763)
- □ Homo sapiens (87,736)
- □ Arabidopsis thaliana (69,010)
- □ Caenorhabditis elegans (26,851)
- □ Danio rerio (20,259)
- □ Drosophila melanogaster (6,151)
- □ Schizosaccharomyces pombe (2,210)
- □ Escherichia coli str. K-12 substr. MG1
- □ Saccharomyces cerevisiae S288c (27

Homo sapiens Nuclear RNase P URS00006B8289

322 nucleotides

Homo sapiens Nuclear RNase P URS000071B825

333 nucleotides

Homo sapiens RNase MRP URS000062A886

265 nucleotides

Homo sapiens Nuclear RNase P URS000071EDBB

334 nucleotides

Homo sapiens Nuclear RNase P URS00006856DB

322 nucleotides

Homo sapiens RNase MRP URS00006CA311

267 nucleotides

Homo sapiens Nuclear RNase P URS000065F37C

335 nucleotides

Homo sapiens RNase MRP URS000065FEF2

267 nucleotides

Homo sapiens RNase MRP URS0000660B88

264 nucleotides

Homo sapiens Nuclear RNase P URS000066C998

55 nucleotides

misc RNA/tmRNA/other from 665 species URS000037602E

363 nucleotides

Homo sapiens RNase MRP URS000067CD0F

267 nucleotides

Homo sapiens RNase MRP URS00006BE901

134 nucleotides

Homo sapiens RNase MRP URS0000697636

267 nucleotides

Homo sapiens RNase MRP URS00006D8FF2

264 nucleotides

Displaying 15 out of 8,607,919 sequences

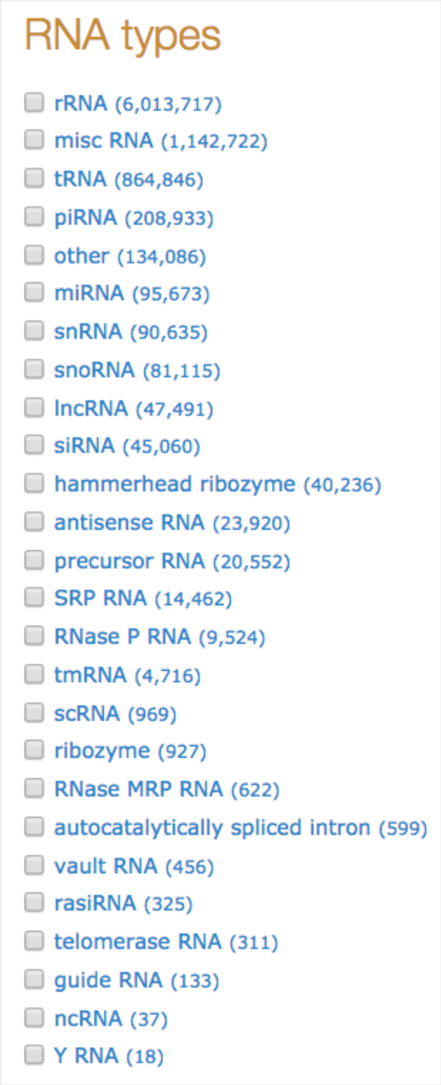
Load more results on demand

http://rnacentral.org/search?q=RNA

Search facets

- **Facets** allow to quickly filter search results while also exposing the kinds of data that are available
- For example, the **RNA types** facet shows how many sequences of each type are present in RNAcentral.

Keyboard shortcut: hitting “/” puts the cursor in the search box



RNA types

- □ rRNA (6,013,717)
- □ misc RNA (1,142,722)
- □ tRNA (864,846)
- □ piRNA (208,933)
- □ other (134,086)
- □ miRNA (95,673)
- □ snRNA (90,635)
- □ snoRNA (81,115)
- □ lncRNA (47,491)
- □ siRNA (45,060)
- □ hammerhead ribozyme (40,236)
- □ antisense RNA (23,920)
- □ precursor RNA (20,552)
- □ SRP RNA (14,462)
- □ RNase P RNA (9,524)
- □ tmRNA (4,716)
- □ scRNA (969)
- □ ribozyme (927)
- □ RNase MRP RNA (622)
- □ autocatalytically spliced intron (599)
- □ vault RNA (456)
- □ rasiRNA (325)
- □ telomerase RNA (311)
- □ guide RNA (133)
- □ ncRNA (37)
- □ Y RNA (18)

Example search: HOTAIR lncRNA

1. Search for HOTAIR http://rnacentral.org/search?q=HOTAIR
2. Exclude HOTAIRM1 entries (type “not hotairm1”)

   http://rnacentral.org/search?q=HOTAIR%20not%20hotairm1
3. Focus on sequences from Vega

   using the Expert Databases facet:

   http://rnacentral.org/search?q=HOTAIR%20not%20hotairm1%20AND%20expert_db:%22VEGA%22
4. Get just the human sequences

   using the Organisms facet:

   http://rnacentral.org/search?q=HOTAIR%20not%20hotairm1%20AND%20expert_db:%22VEGA%22%20AND%20TAXONOMY:%229606%22

Programmatic access

- **Learn just one API** instead instead of dozens
- Browse the API to get a sense of how it works

  http://rnacentral.org/api/v1/
- Documentation: http://rnacentral.org/api
- Example Python script:

  http://rnacentral.org/api#v1-example-script

FTP archive

For each release the archive contains:

- **sequences** in FASTA format
- **mapping** between RNAcentral and Expert Database identifiers
- **MD5** values for all sequences that can be used to match large numbers of sequences to RNAcentral IDs
- **genomic coordinates** in multiple formats

ftp://ftp.ebi.ac.uk/pub/databases/RNAcentral

RNA modifications (coming soon)

RNAcentral will import modification information from PDBe and Modomics.

For example, position **628** in URS000080DFCD is **1MA**

(6-HYDRO-1-METHYLADENOSINE)

Positional IDs:

URS000080DFCD.628 = 1MA

Example: http://test.rnacentral.org/rna/URS000080DFCD

Future plans for RNAcentral

- import **more data**
- do more **genomic mapping**
- organise **related sequences** from the same species
- import **2D and 3D** structure information
- **integrate** closer with Rfam and other resources (Ensembl, EuropePMC, PDBe, others)

New RNA positions at EBI

RNAcentral and Rfam:

- RNA project leader (EBI_00582)

Rfam:

- Software developer (EBI_00551)
- Database biocurator (EBI_00552)

Getting in touch

- by email: apetrov@ebi.ac.uk or helpdesk@rnacentral.org
- via the website: http://rnacentral.org/contact
- on Twitter: @rnacentral
- on GitHub: https://github.com/RNAcentral/rnacentral-webcode/issues

Acknowledgements

EMBL-EBI

- Alex Bateman
- Paul Kersey
- Guy Cochrane
- Simon Kay
- Richard Gibson
- Dan Staines
- Rob Finn
- Elspeth Bruford
- Mathew Wright
- Sameer Velankar

- DBAs
- Systems
- Web Production

and other teams

Acknowledgements - Expert Databases

Sam Griffiths-Jones

Jennifer Harrow

Christian Zwieb

Todd Lowe
Patricia Chan

Kelly Williams
Corey Hudson

Michael Clark
Camelia Quek

Mike Cherry

Paul Sternberg

James Cole
Benli Chai

Kim Pruitt

Naoya Kenmochi

Tanya Berardini

All RNAcentral Consortium members:
http://rnacentral.org/expert-databases

Thank you!

@RNAcentral

http://blog.rnacentral.org/