RNAcentral: Exploring non-coding RNA sequences

Anton Petrov [1]
DNA & RNA
Beginner
0.5 hour

RNAcentral [2] is a database of non-coding RNA (ncRNA) sequences that aggregates data from multiple specialised resources and provides a single entry point for accessing ncRNA sequences of all types from all organisms. This course provides a brief overview of the data and the RNAcentral website. You can also watch our webinar [3] for an overview and demo of RNAcentral.

Learning objectives:

- Describe RNAcentral and the data it contains
- Be able to navigate and search RNAcentral
- Be able to access the data via the FTP archive and REST API

What is RNAcentral?

[4]RNAcentral [2] is a database of non-coding RNA [5] (ncRNA) sequences that aims to create a comprehensive set of sequences of all ncRNA types from all species. RNAcentral integrates information from a Consortium of established databases, such as miRBase or GENCODE [6].

Why RNAcentral?

ncRNA sequences are distributed across a number of databases that usually specialise only on certain ncRNA types or focus only on certain organisms. Before RNAcentral there was no single starting point for searching and accessing this data. There were no standard sequence identifiers for RNA sequences, so comparing data across different resources or performing sequence searches was challenging.

Launched in 2014 [7], RNAcentral aims to make it easier for researchers to find ncRNA sequences and ultimately become an RNA equivalent of UniProt [8].
Key features

- **unified access** to all non-coding RNA types from all organisms
- faceted **text search** that enables quick data filtering [9] and comparison
- **sequence similarity search** against a comprehensive set of ncRNA sequences
- **genome annotations** for select species
- single **API** [10] for accessing data from multiple resources
- **FTP** [11] download archive
- **stable identifiers** for distinct RNA sequences
- **cross-references** to specialised RNA databases

Explore the interactive timeline [12] to see how RNAcentral developed over time

Where do the data come from?

**Expert Databases**

RNAcentral presents a unified interface for sequences and annotations supplied by a Consortium of specialised RNA databases, which are referred to as **Expert Databases**.

The list of the imported databases keeps growing. For example, RNAcentral release 5 [13] contained **22** Expert Databases (Figure 1).
Figure 1 The 22 expert databases in RNAcentral release 5.

Work is underway to import ~20 more databases, and the up-to-date list of Expert Databases [14] is available on the RNAcentral website.

RNAcentral architecture

RNAcentral Expert Databases submit their sequences and annotations to RNAcentral, and the data become discoverable through the RNAcentral website. RNAcentral also provides quality control feedback to the Expert Databases, which in many instances led to improved data quality. The users can submit the data to Expert Databases or work directly with RNAcentral (see the section How to submit data to RNAcentral [15] for more information) (Figure 2).
Find out more about how RNAcentral works in a recent paper [17] in Nucleic Acids Research [2 [16]].

**Data overview**

**How many sequences are in RNAcentral?**

In release 3 [18], RNAcentral contained more than 8.6 million unique ncRNA sequences and more than 18 million database cross-references to 15 external databases (Figure 3).
RNAcentral: Exploring non-coding RNA sequences

Figure 3 The number of sequences in RNAcentral is growing. View an interactive version of the graph [19] on the RNAcentral website.

Which species are represented in RNAcentral?
The sequences come from a wide range of species covering most of the taxonomic space (Figure 4).

Figure 4 Species distribution of RNAcentral sequences [216].

Which ncRNA types are found in RNAcentral?
RNAcentral imports most known types of non-coding RNAs, for example:

- tRNA [20]
- rRNA [21]
- miRNA [22]
- IncRNA [23]
- snoRNA [24]
- piRNA [25]
- SRP RNA [26]
- vault RNA [27]
- and many others

The sequences are annotated with ncRNA types from a controlled vocabulary [28] maintained by INSDC [29]. Some classification categories, such as other ncRNA [30] or misc_RNA [31], serve as umbrella terms for other ncRNA classes that were either not well described or poorly annotated, and are subject to refinement in the future.

RNAcentral focuses on ncRNA transcripts. Cis-regulatory RNAs, such as structured RNA elements found in mRNAs, are not included.

**How often is the data updated?**

New RNAcentral releases are made available every 3-4 months, and the website is updated more often. Provisional release dates are available on the public RNAcentral events [32] calendar.

Never miss another release!
Subscribe to the RNAcentral blog [33] and follow RNAcentral on Twitter [34] to stay in the loop.

**RNAcentral website**

Let's have a look at the RNAcentral homepage [2] (Figure 5). There are three main ways of exploring the data which will be covered in detail later in the course:

1. Keyword search
2. Sequence search
3. Genome browser

You can also see the list of participating databases and access example entries.
Figure 5 The RNAcentral homepage.

Tip: Click the 'Take a tour' button to take an interactive guided tour of the website in action!

Viewing RNAcentral sequences

Each ncRNA sequence has a page similar to this example [35] (Figure 6):
Genomic mapping

Let’s explore some other features of the sequence pages.

Viewing entries in their genomic context

Many entries in RNAcentral come from reference genomes. These entries can be viewed in their genomic context using an embedded genome browser [37] and their coordinates can be downloaded in GFF [38]/GFF3 [39]/BED [40] formats (Figure 7).
The up-to-date list of species with genome mapping [41] can be found in the RNAcentral help centre [42].

Figure 7: Viewing entries in their genomic context. Click here [43] to view the example above.

Tip: Entries with genomic mapping can be found using RNAcentral search by selecting the Genomic mapping facet in the RNAcentral search as shown below (Figure 8).

Figure 8: The genomic mapping facet can be used to view entries with genomic mapping.

Literature references
Viewing citations for each cross-reference

For each cross-reference in the **Annotations table** you can view literature citations (Figure 9):

![Annotations table with cross-reference citations](image)

**Figure 9** Viewing citations in the annotations table.

**Viewing all citations for a sequence**

Publications from all cross-references are collected in the **Publications tab** (Figure 10):
RNAcentral: Exploring non-coding RNA sequences

Publications are searchable

Publication titles, authors names, and other metadata are searchable, so you can quickly look up ncRNA sequences mentioned in the literature.

For example, you can find braveheart lincRNA [44] because the sequence is associated with the publication where it was described (Figure 11).
Exploring Expert Databases

Each Expert Database has a page similar to the miRBase page [45] (Figure 5) where you can find what database version is currently imported into RNAcentral and explore taxonomic coverage and sequence length distribution using interactive graphs.

![miRBase Expert Database](image)

**Figure 12** The miRBase page in RNAcentral.

**Sequence length distribution graph**

The sequence length distribution graph can provide important insights about the data.

For example, the graph below (Figure 13) shows that most sequences from miRBase are about 21 nucleotides long, which is consistent with the length of mature miRNAs. However, miRBase also provides the information about precursor miRNAs, which are more varied in length.
**Figure 13** Sequence length distribution graph.

**Tip:** Explore the data by clicking on the graph to launch a search for sequences of specific length. View an example [45].

**Species sunburst**

The sunburst diagram shows what taxonomic groups the sequences are coming from (Figure 14). You can hover your mouse over the graph to explore the taxonomic distribution of species. See an example [45].
**Figure 14** Species sunburst diagram.

**Searching RNAcentral**

RNAcentral has a powerful search interface that lets you discover ncRNA sequences from multiple Expert Databases in one place.

**Search facets**

RNAcentral search has **facets** that can filter search results and expose the structure of the data. For example, you can quickly find all human IncRNAs or get all ncRNAs from any organism (mouse, worm, fly, cow and others).

Here is an example of using facets. If you click the **Browse** button in the menu or on the homepage, you should see a page similar to this (Figure 15):
Figure 15 Example search results page.

The facets on the left show how many entries of each kind are present in the database. You can also filter the entries by source, ncRNA type, organism, and more.

💡 **Tip: Enjoy using keyboard shortcuts?** Use "/" (forward slash) to focus keyboard in the search box and start typing your query faster.
Example search

In this example we want to find HOTAIR IncRNA from humans that are in the VEGA [47] database:

1. Let's search for HOTAIR [48] IncRNA by typing "HOTAIR" in the search box. You will notice that this query will match many HOTAIR sequences and also several HOTAIRM1 sequences, which is a different IncRNA (Figure 16). To exclude unwanted items you can use a "not" operator.

![Search results for HOTAIR](image)

**Figure 16** Search results for HOTAIR

2. To exclude HOTAIRM1 entries [49] add "not hotairm1" to the query. Notice that this search finds a lot fewer entries than the one above.

![Tip: Another way of achieving the same goal is to use double quotes](image)

**Tip:** Another way of achieving the same goal is to use **double quotes** (so you would type "HOTAIR" instead of HOTAIR), which ensures that the search results contain the query as a full word.

**Try for yourself** [50].

3. To focus on sequences from the VEGA data base [51] tick the checkbox in the Expert Databases facet. At the time of writing, this query retrieved six sequences, five from human and one from mouse (Figure 17).
Figure 17 Filtering the HOTAIR search results using the 'not' operator and the 'Expert Databases' facet.

4. To view only human sequences tick 'Homo sapiens' in the 'Organisms' facet [52].

Figure 18: HOTAIR lncRNAs from humans that are in the VEGA database.

5. Now you can explore the alternative transcripts one by one or download the sequences for further analysis.

**Exporting search results**

**Downloading search results**

Sometimes it is useful to download search results, for example, if you would like to create a multiple sequence alignment for a set of sequences matching your search criteria.
To download the results:

1. Click the **Download** button to choose a file format (Figure 19):

   ![Figure 19](image19.png)

   **Figure 19** Using the download button to save search results.

2. You will be redirected to a page with a unique URL where the results will be available for download (Figure 20):

   ![Figure 20](image20.png)

   **Figure 20**: Results can be downloaded from a unique URL.

   - The export functionality works best for search results with less than 100,000 entries. For larger results, you might want to use the [FTP archive](#) or the [API](#).
Advanced queries

RNAcentral search supports advanced query syntax that makes it easy to compose very precise queries. You can find examples and documentation in the RNAcentral help centre, but here is a quick review of some advanced search features.

Tip: Don't worry about remembering the syntax - you can always look it up on any search results page or in RNAcentral help (55) (Figure 21).

Field-specific searches

When you type a simple query in the search box, it will match all entries that contain that word anywhere in that entry.

For example, if you type "Vega" in the query box when trying to find sequences from the Vega Expert Database [57], you will also find sequences from publications authored by people whose last name is Vega.

To perform a more specific search, you can either use facets or rewrite your query like this: expert_db:"Vega" [58], which means that you are only interested in the field called expert_db. This search will match only sequences from the Vega Expert Database.

Field-specific searches have the following format: field_name:"query term"

Note that the query term is surrounded by quotation marks.

You can also run field-specific searches for:

- NCBI Taxonomy identifiers;
- common species names (for example, mouse, cow etc);
- ncRNA type (rRNA, tRNA, snoRNA etc);
- gene name (HOTAIR, TSIX etc);
- sequence length;
RNAcentral: Exploring non-coding RNA sequences

- publication metadata: author, paper title, PubMed identifiers, DOI;

- and more [55]!

**Logical operators**

Keywords can be grouped with parentheses and connected with **AND**, **OR**, and **NOT** logical operators to express relations between query terms.

By default, all keywords are joined using the **AND** operator, even if you don't type it.

For example, if you would like to download sequences from PDB [59] structures 4V4Q and 1S72, you can search for **4V4Q or 1S72** [60] and get sequences from both 3D structures.

Note that a search for **4V4Q and 1S72** [61] won't find any results because these 3D structures don't have any sequences in common.

**Exercises**

Try out the [RNAcentral search][2] by completing the following example tasks and check the answers on the next page:

1. Find the [RefSeq][62] identifier of braveheart IncRNA.

2. Find the IncRNAs from the following paper:
   
   A cytoplasmic NF-kB interacting long noncoding RNA blocks IκB phosphorylation and suppresses breast cancer metastasis [63]
   

3. Download [FASTA][64]-formatted [rRNA][21] sequences that have experimentally determined 3D structures deposited in Protein Data Bank.

4. Search for your favorite ncRNA and see if it is in RNAcentral.

**Check your answers**

1. A search for braveheart [44] should find this [RNAcentral entry][65] which has RefSeq
identifier NR_045420.2.

2. You can search for a Pubmed ID [66] of the paper and find the sequence of the human NKILA IncRNA [67].

3. Click Browse, then pick rRNA in the RNA type facet and PDB in the Expert database facet, which should give you this search [68]. Then click the Download button and choose FASTA format.

Hope you found all the answers! If you need help completing any of the exercises, feel free to get in touch [69].

Sequence search

RNAcentral provides a convenient way of running sequence searches [70] against a comprehensive set of ncRNA sequences.

Sequence search features

- Sequence search is powered by nhmmer [71] [3 [16]].
- Results are stored for 7 days and can be accessed using unique URLs which can be shared or saved.
- Results are sortable (for example, you can compare the best and the worst alignments by sorting the results by E-value [72] in increasing and decreasing order).

Sequence search results

Here is an example sequence search (Figure 22):
**Figure 22** Example of a sequence search.

**Instant retrieval of exact sequence matches**

Sometimes all you want is to find whether a certain sequence is found in RNAcentral. When you submit a query, the exact sequence match is retrieved **instantly** (if it exists), and you can **cancel** the search if you are not interested in similar sequences (Figure 23).
Figure 23 Instant retrieval of sequence matches.

Give it a try [70] by launching one of the example searches.

Browsing by genome location

If you would like to examine ncRNAs found in a specific genomic location, you can use a lightweight

**genome browser** [73] that visualises RNAcentral sequences alongside genes and transcript from

Ensembl [74] and Ensembl Genomes [75] (Figure 24). If you need more powerful tools, you can
easily switch to Ensembl [74] or UCSC [76] genome browsers using the provided links:
RNAcentral: Exploring non-coding RNA sequences

Figure 24 Viewing human HOTAIR transcripts in the RNAcentral genome browser.

RNAcentral has genome mapping for several important species, including human, mouse, fly, worm, yeast, and others.

Not every sequence in RNAcentral is mapped to a genome. For example, some sequences come from older genome assemblies and are not carried over to the latest assembly. Other sequences do not match the reference genome and may represent sequence variation. There is also a possibility of misannotation or contamination (for example, a bacterial sequence incorrectly annotated as coming from human), and such cases could be detected by checking Taxonomic distribution of a sequence or by running sequence similarity searches.

RNAcentral identifiers

One of the goals of RNAcentral is to assign unique and stable identifiers to all ncRNA sequences, which may simplify many tasks, such as referring to an RNA sequence in a paper or comparing two sets of sequences.

RNAcentral identifiers

RNAcentral identifiers are like barcodes that uniquely identify each RNA sequence. Each
distinct sequence is assigned an **Unique RNAcentral Sequence identifier** \[77\] (URS) regardless of what species the sequence is from. RNAcentral identifiers have the following format: **URS + 10-digit hexadecimal number**

For example, sequence \texttt{AACCGUAGAUCCGAACUUGUG} has been assigned the following identifier: \texttt{URS000040D674} \[78\]. All Expert Databases that annotate this sequence will be linked to this identifier.

This format can accommodate more than 1 trillion sequences. At the time of writing about 8.6 million identifiers have been assigned.

**Species-specific identifiers**

RNAcentral also provides **species-specific identifiers** that are useful when you are interested in information about a certain species. When you search RNAcentral, you are searching within these species-specific entries.

Species-specific identifiers have the following format: **URS identifier** \[77\] / NCBI taxid.

Example: \texttt{URS00003B7674} \[79\]/\texttt{9606} \[79\]

An underscore can be used instead of a slash (for example, the identifier above is equivalent to \texttt{URS00003B7674} \[80\] \_\texttt{9606} \[80\]).

These identifiers can be used to link to annotations from specific species. For example \[81\], the RNAcentral species-specific identifiers are used for annotating ncRNAs with Gene Ontology \[82\] terms.

**Switching between species**

You can switch between species using the **Taxonomy tab** (Figure 25):
How are RNAcentral identifiers assigned?

Knowing how the identifiers are assigned can help you to take advantage of several useful features of the RNAcentral API [10] and the FTP [11] archive.

Assigning RNAcentral identifiers

1. First, the sequence is normalised:

   Each sequence is uppercased and all U's are replaced with T's, thus the RNA sequence is converted to its DNA form. The DNA form is used for consistency with the sequence archives, such as ENA.

2. MD5 hash of the normalised sequence is computed:

   An MD5 hash [84] is a string of 32 characters that can be uniquely generated based on a full sequence. Instead of comparing sequences directly, RNAcentral compares their MD5 hashes because it can be much faster, especially for longer sequences.

3. Finally, the MD5 hash is checked against the database to see if it is already present:

   If the hash is not present in the database, it means that the sequence is new, and then a new URS identifier [77] is created and permanently stored along with the sequence and the MD5 hash.
If the hash is already in the database, then the sequence has been seen before and there is no need to generate a new identifier.

Now you can take any RNA sequence, get its MD5 hash and check if it is found in RNAcentral using the RNAcentral API [54] or the FTP archive [53].

All sequences in RNAcentral are at least 10 nucleotides long because shorter sequences are not likely to represent biologically relevant ncRNAs.

**Distinct sequences get distinct identifiers**

Every distinct ncRNA sequence gets its own identifier, so if two sequences are even slightly different, they still get distinct RNAcentral identifiers. This is similar to how UniParc [85] assigns unique identifiers to protein sequences.

For example, sequence **URS0000759BE2** [86] (2,547 nucleotides long) is the same as **URS0000621DCB** [87] (2,546 nucleotides) except that it has one more nucleotide on its 3' end. Although these sequences are almost identical and come from the same genomic location in the same species, they still get distinct identifiers to recognise the fact that the sequences are distinct.

**How to get data from RNAcentral**

**FTP archive**

RNAcentral data, including previous releases, can be downloaded for local processing from the FTP archive [88].

**Example FTP use cases**

- Looking up RNAcentral identifiers for a large number of sequences. This can be achieved by downloading the MD5 file [89] and checking if the file contains MD5 hashes of your sequences;
- Getting genome coordinates [90] of RNAcentral sequences from the reference genomes.

**API**

Most data in RNAcentral can be accessed programmatically using a RESTful API allowing for integration with other resources.

To get started with the API, have a look at the API documentation [54] and explore the API [91] directly in your browser.
Thanks to [content negotiation](#), when the API URLs are opened in a browser, they are rendered as [human-friendly](#) documents. However, when the same pages are requested programmatically, the response is returned in a [machine-readable](#) format.

**Example API use cases**

- [Forna](#) can import sequences from RNAcentral on the fly to predict and visualise their [secondary structure](#).
- [Assemble](#) uses the RNAcentral API to [load sequence data](#) for further analysis.

**Should I use API or FTP?**

The API and the FTP archive are [complementary](#) resources. Choosing the right one can help you accomplish your computational tasks in the most efficient way.

**When to use the API?**

- For random access to a small number of entries
- For interactive user interfaces
- If you need JSON-formatted data

**When to use the FTP archive?**

- To perform computationally-intensive analyses
- For parallelisable applications (e.g. using a cluster)

**Tip: Fewer than 100,000 entries?**

Consider using search results export!

**How to submit data to RNAcentral**

RNAcentral takes advantage of the data sharing infrastructure maintained by [INSDC](#), which is a collaboration between EMBL-EBI ([ENA](#)), [DDBJ](#), and NCBI ([GenBank](#)). This means that ncRNA sequences submitted to any of these resources will automatically flow into RNAcentral.
A **good starting place** [101] for submitting your data to RNAcentral is the **European Nucleotide Archive** [102] (ENA).

You can also **contact RNAcentral** [69] directly if you would like to provide annotations on sequences that are already in INSDC.

![ENA](image1.png) ![NCBI](image2.png) ![DDBJ](image3.png)

Your data may already be in RNAcentral!

Try searching RNAcentral [2] for your name or your paper title.

---

**Summary**

**What is RNAcentral?**

RNAcentral is a database of ncRNA sequences of **all ncRNA types from all organisms**. It provides a single entry point for accessing data from numerous specialised RNA databases [14].

**How to access the data?**

There are **three main ways** to explore RNAcentral:

- **Text search** [46] and browsing the data using the search interface
  - Try searching RNAcentral by gene name, RNA type, database or species name, paper title or **other keywords** [55].

- **Sequence search** [103]
  - Search against a comprehensive set of ncRNA sequences using a web interface.

- **Genome browser** [73]
  - Explore ncRNAs in reference genomes from select species.

**Want to learn more?**

Watch a recording of RNAcentral webinar [3]:

---

**Where to get help?**

If you have any questions about RNAcentral, do not hesitate to get in touch [69]!
Quiz: RNAcentral

Questions: 4
Attempts allowed: Unlimited
Available: Always
Pass rate: 75%
Backwards navigation: Allowed

Your Feedback

Tell us what you thought about this course!

Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

How to get help

If you have questions or feedback please feel free to get in touch using the Feedback form located in the site menu (Figure 26):

Figure 26 Get help using the feedback form available on any page.

You can also reach us in other ways:

- email us using the Contact us [69] form on the RNAcentral website
RNAcentral: Exploring non-coding RNA sequences

- send your feedback on Twitter [104]
- report issues or make feature requests on GitHub [105].

Also consider searching for an answer in the help centre [106] or on the RNAcentral blog [107].

References


Acknowledgements

RNAcentral is developed at the European Bioinformatics Institute [110] and is funded by BBSRC [111], the UK Biotechnology and Biological Sciences Research Council.

Expert Databases

RNAcentral would not be possible without collaboration from the Expert Databases [14] and without thousands of scientists around the world who submit their ncRNA data to the international sequence archives.

Contributors
Anton Petrov [1]

EMBL-EBI
Project Leader RNA Resources

Anton Petrov joined EMBL-EBI in 2013 to help develop RNAcentral, a public database of non-coding RNA sequences. In 2015 he became Project Leader of RNA Resources, including RNAcentral and Rfam, a database that classifies non-coding RNAs into families. Originally from St.Petersburg, Russia, Anton completed his PhD with Prof. Neocles Leontis in Bowling Green State University, Ohio, USA, where he built databases and web applications for RNA 3D structure analysis.

The RNAcentral team

RNAcentral is produced by a large team of people at the European Bioinformatics Institute and around the world who write code, keep the servers running, and most importantly - submit their ncRNA sequences and annotations.

Have a look at the list of authors [112] of the latest RNAcentral paper in Nucleic Acids Research.

All course materials in Train online are free cultural works licensed under a Creative Commons Attribution-ShareAlike 4.0 International license

Source URL: http://www.ebi.ac.uk/training/online/course/rnacentral-exploring-non-coding-rna-sequences

Links
[1] http://www.ebi.ac.uk/training/online/trainers/apetrov
[9] http://www.ebi.ac.uk/training/online/glossary/filtering
[10] http://www.ebi.ac.uk/training/online/glossary/api
[68] http://rnacentral.org/search?q=rna_type:%22rRNA%22%20AND%20expert_db:%22PDB%22
[69] http://rnacentral.org/contact
[70] http://rnacentral.org/sequence-search/
[72] http://www.ebi.ac.uk/training/online/glossary/e-value
[73] http://rnacentral.org/genome-browser
[74] http://ensembl.org
[75] http://ensemblgenomes.org
[76] http://genome.ucsc.edu
[77] http://www.ebi.ac.uk/training/online/glossary/identifier
[80] http://rnacentral.org/rna/URS00003B7674_9606
[82] http://www.ebi.ac.uk/training/online/glossary/gene-ontology
[85] http://www.ebi.ac.uk/training/online/glossary/uniparc
[92] http://www.ebi.ac.uk/training/online/glossary/content-negotiation
[93] https://github.com/pkerpedjiev/forna/
[94] http://www.ebi.ac.uk/training/online/glossary/secondary-structure
[95] http://bioinformatics.org/s2s/
[96] https://assemble2.wordpress.com/2015/02/04/load-data-directly-from-rna-databases-into-assemble2/
[98] http://www.ebi.ac.uk/ena
[99] http://www.ddbj.nig.ac.jp/
[101] http://www.ebi.ac.uk/ena/submit
[102] http://www.ebi.ac.uk/training/online/glossary/european-nucleotide-archive
[104] https://twitter.com/rnacentral
[105] https://github.com/RNAcentral/rnacentral-webcode/issues
[109] http://europepmc.org/articles/PMC4384043
[110] http://www.ebi.ac.uk/
[111] http://www.bbsrc.ac.uk/