Webinar Q&A Session

The RNAcentral webinar was presented by Anton Petrov
as part of the EBI Training webinar series on October 28th, 2015

Useful links

- Slides from the webinar
- Webinar recording on YouTube
- This document

If two transcripts differ by one nucleotide, are they considered different molecules?

If two RNA sequences are different even by a single nucleotide, they get different RNAcentral identifiers. This is useful for many bioinformatic applications where it is important to know exactly what sequence is being referred to. Currently such sequences can be identified by examining their genomic neighborhood using the embedded genome browser (where available).

Here is an example where two sequences URS0000759BE2 (from RefSeq) and URS0000621DCB (from Vega) are different by one nucleotide but are found in the same genomic location. In the future versions of RNAcentral, such sequences will be identified automatically and presented to the user in the web interface.

Will FlyBase data be incorporated in the next release?

FlyBase is a member of the RNAcentral Consortium and will be imported in one the coming RNAcentral releases, possibly in release 5, so stay tuned!

What is an MD5 value?

For each sequence in RNAcentral we calculate an MD5 checksum, which is a 32-character string that uniquely corresponds to the input sequence. Using MD5s makes looking up sequences much faster because instead of searching the database for the entire sequences we can look for much shorter MD5 values.

Here is a use case for MD5 checksums. The RNAcentral FTP Archive contains a mapping file with correspondences between RNAcentral identifiers and MD5 checksums of their sequences. If you have a large number of sequences that you would like to get RNAcentral identifiers for,

the quickest way to do it is to calculate MD5 values for your sequences and check if any of them are found in the mapping file.

How does RNAcentral deal with CDS and UTRs?

RNAcentral does not import sequences that code for proteins or the untranslated regions of mRNAs. However, there is growing evidence that many sequences that were previously thought to be non-coding can actually be translated into short peptides, and these sequences might be present in RNAcentral. Here is a recent review about smORFs by Juan Pablo Couso.

Is there a way to know how the ncRNA was experimentally detected?

Each RNAcentral sequence is linked to publications which describe how the sequence was obtained. In the future we plan to use evidence codes to facilitate searches for sequences with certain levels of experimental support.

Does RNAcentral have data on transposons?

No, RNAcentral does not import transposon sequences, although it is possible that some transposon sequences were misannotated and made their way into RNAcentral.

Are there any plans to include information regarding functional annotation of the ncRNAs?

Indeed, we are planning to import new types of functional annotations into RNAcentral, such as information about modified nucleotides, interaction partners, and high quality secondary structures. This will be the focus of development in the coming years.

In future releases will you be able to find ncRNA with similar 2D or 3D structures, even if they have different sequences? Is there any way to do this now?

In the future we plan to import high quality secondary structures into RNAcentral and we may implement a search by secondary structure. Other databases currently provide similar functionality. For example, Rfam is a database of non-coding RNA families, and if some sequences are found in the same Rfam family then they share the same consensus secondary structure. Also RNA Frabase 2.0 provides a web-based search for secondary structure fragments found in RNAs with experimentally determined 3D structures found in PDB.

Does RNAcentral currently perform curation steps for the sequences imported from expert databases or are all sequences imported from the expert databases?

RNAcentral does not manually curate any of the data provided by the Expert Databases. However, we perform various quality control steps during data import, and we work closely with Expert Databases to resolve any issues that are detected.

Some Expert Databases submit to RNAcentral only a subset of their data, for example, at the time of writing RDP provided only the high-quality subset of ribosomal RNAs. In general we plan to import as much sequence data as possible and develop efficient search tools to enable the users to explore and filter the data.