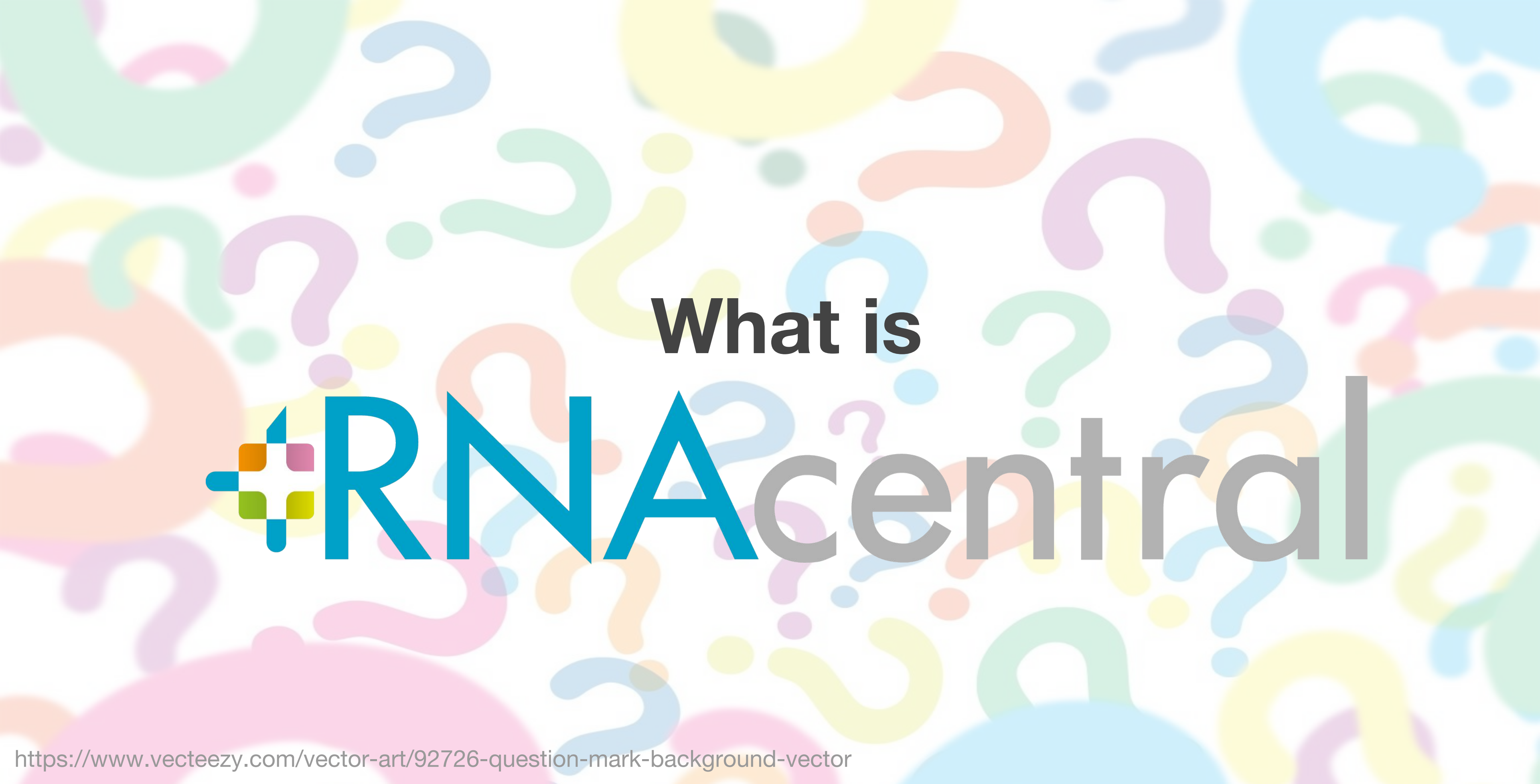
What’s new in RNAcentral and Rfam

Anton Petrov
apetrov@ebi.ac.uk

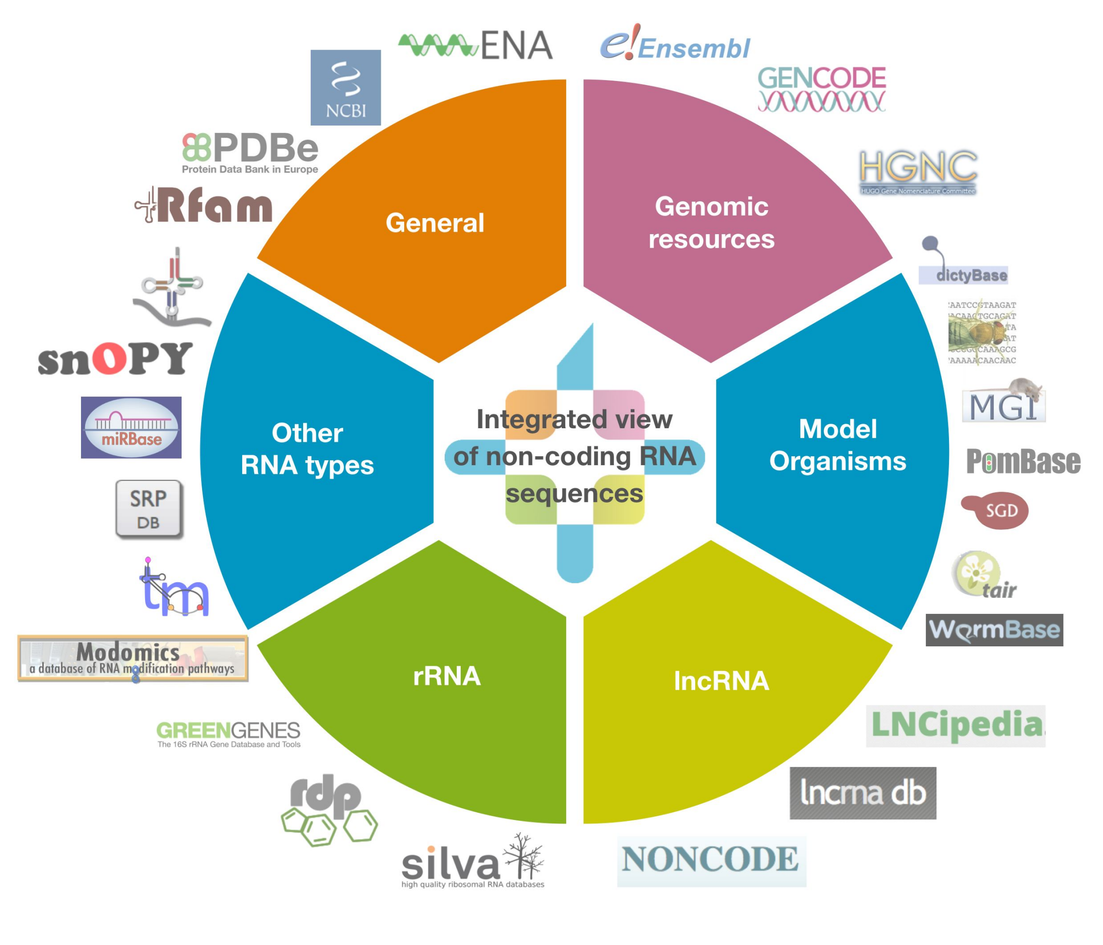
Benasque - July 20, 2018

EMBL-EBI



What is RNAcentral

https://www.vecteezy.com/vector-art/92726-question-mark-background-vector



Integrated view of non-coding RNA sequences

General

Genomic resources

Model Organisms

lncRNA

rRNA

Other RNA types

ENA
Ensembl
GENCODE
HGNC
HUGO Gene Nomenclature Committee
dictyBase
MGI
PomBase
SGD
tair
WormBase
LNCipedia
lncrna db
NONCODE
silva
high quality ribosomal RNA databases
rdp
GREENGENES
The 16S rRNA Gene Database and Tools
Modomics
a database of RNA modification pathways
tm
SRP DB
miRBase
snOPY
Rfam
PDBe
Protein Data Bank in Europe
NCBI

RNAcentral

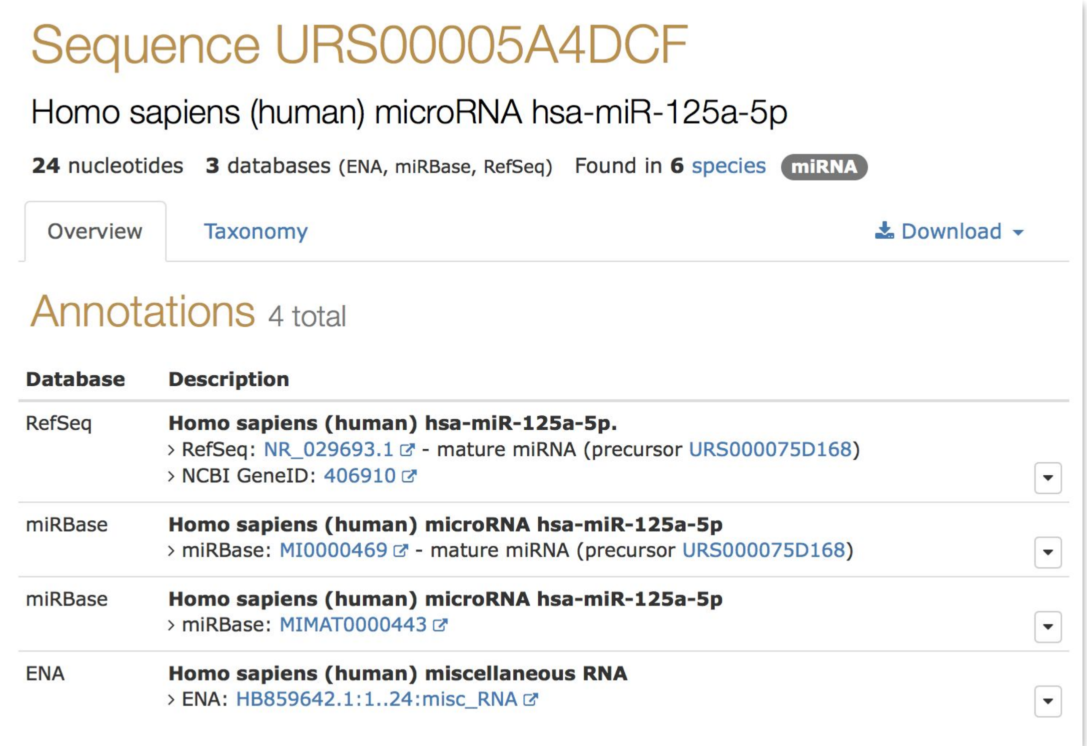
The non-coding RNA sequence database

rnacentral.org

- >10 million sequences
- 27 databases
- 800,000 species

RNAcentral has lots of useful data

- Sequence
- Description
- RNA type
- Links to other databases
- Genome locations
- Publications
- RNA modifications from MODOMICS and PDB



Sequence URS00005A4DCF

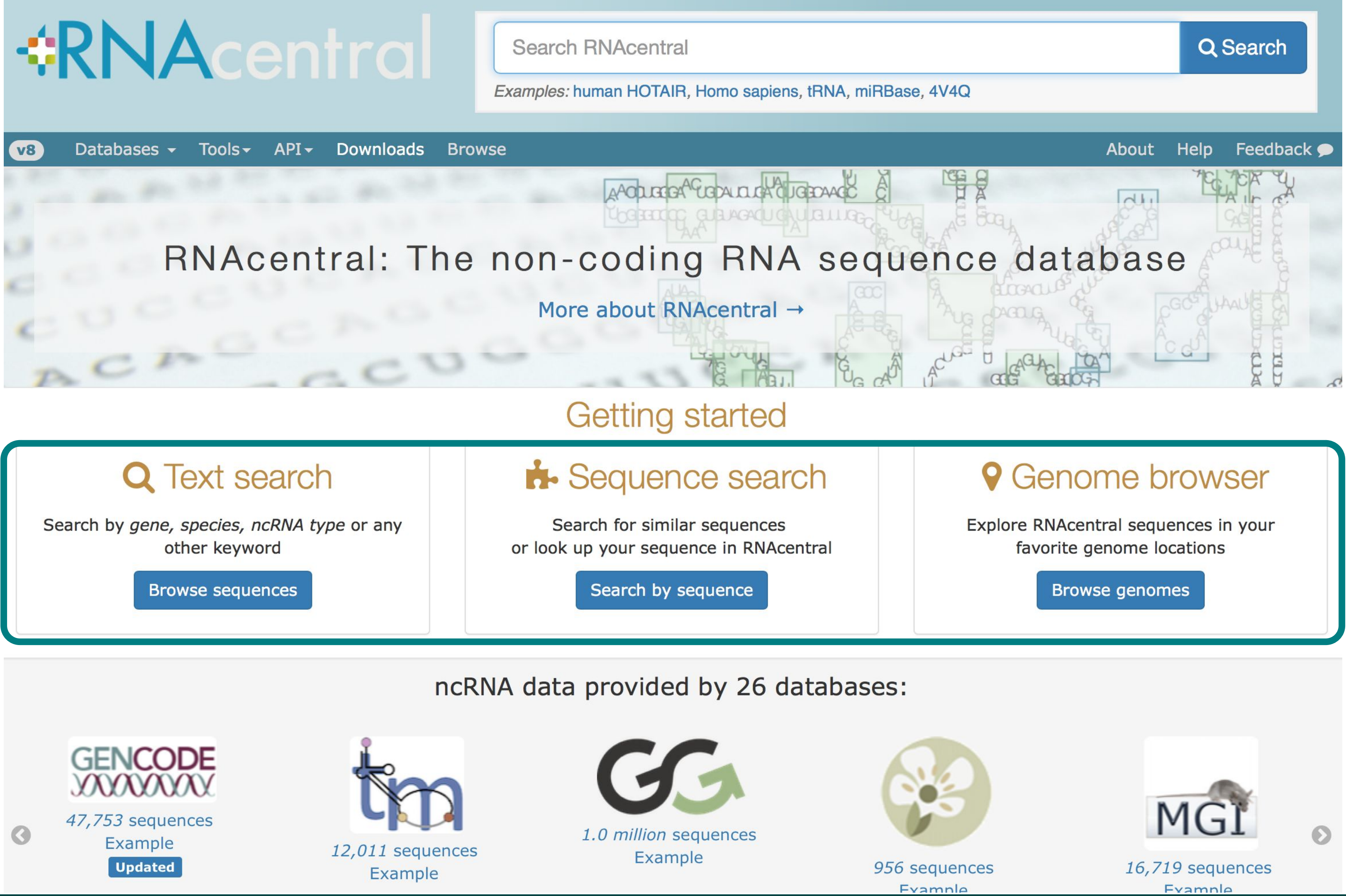
Homo sapiens (human) microRNA hsa-miR-125a-5p

24 nucleotides **3** databases (ENA, miRBase, RefSeq) Found in **6** species miRNA

Annotations 4 total

Database	Description
RefSeq	**Homo sapiens (human) hsa-miR-125a-5p.** › RefSeq: NR_029693.1 - mature miRNA (precursor URS000075D168) › NCBI GeneID: 406910
miRBase	**Homo sapiens (human) microRNA hsa-miR-125a-5p** › miRBase: MI0000469 - mature miRNA (precursor URS000075D168)
miRBase	**Homo sapiens (human) microRNA hsa-miR-125a-5p** › miRBase: MIMAT0000443
ENA	**Homo sapiens (human) miscellaneous RNA** › ENA: HB859642.1:1..24:misc_RNA

http://rnacentral.org/rna/URS00005A4DCF/9606



RNAcentral: The non-coding RNA sequence database

More about RNAcentral →

Getting started

Text search

Search by *gene, species, ncRNA type* or any other keyword

Sequence search

Search for similar sequences or look up your sequence in RNAcentral

Genome browser

Explore RNAcentral sequences in your favorite genome locations

ncRNA data provided by 26 databases:

Until recently

just data aggregation,

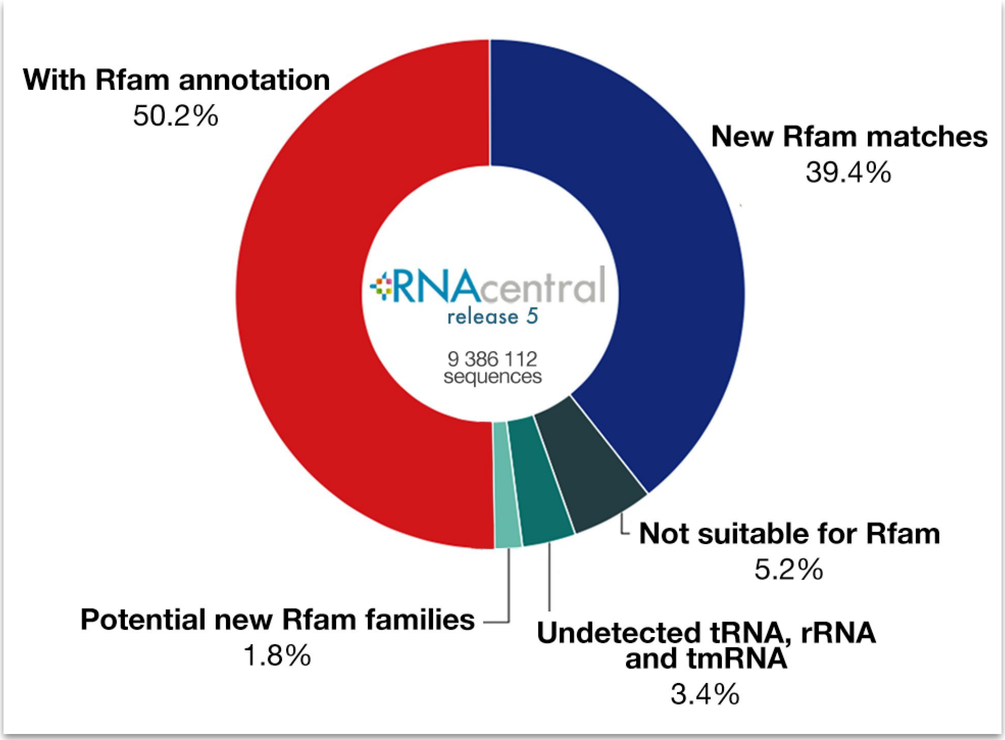
now additional analysis

Two important new features

1. **Quality control** using Rfam models

2. Comprehensive **genome mapping**

1. Rfam models are used to annotate RNAcentral

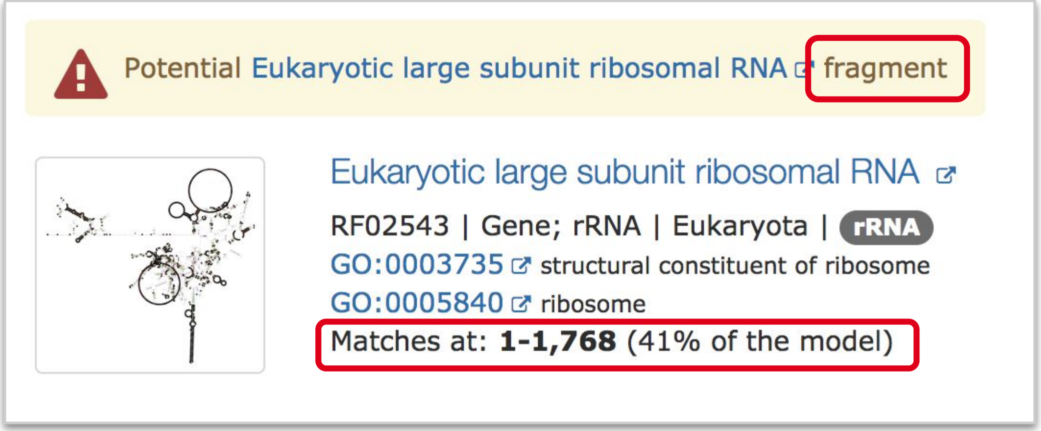


- **~90%** of RNAcentral sequences match Rfam models
- about **2%** of RNAcentral sequences can be used to build new Rfam models

Natalia Quiñones Olvera

Rfam annotations help detect:

- **truncated** sequences
- potential **contamination**
- **missing** annotations



⚠ Potential Eukaryotic large subunit ribosomal RNA fragment

Eukaryotic large subunit ribosomal RNA

RF02543 | Gene; rRNA | Eukaryota | rRNA

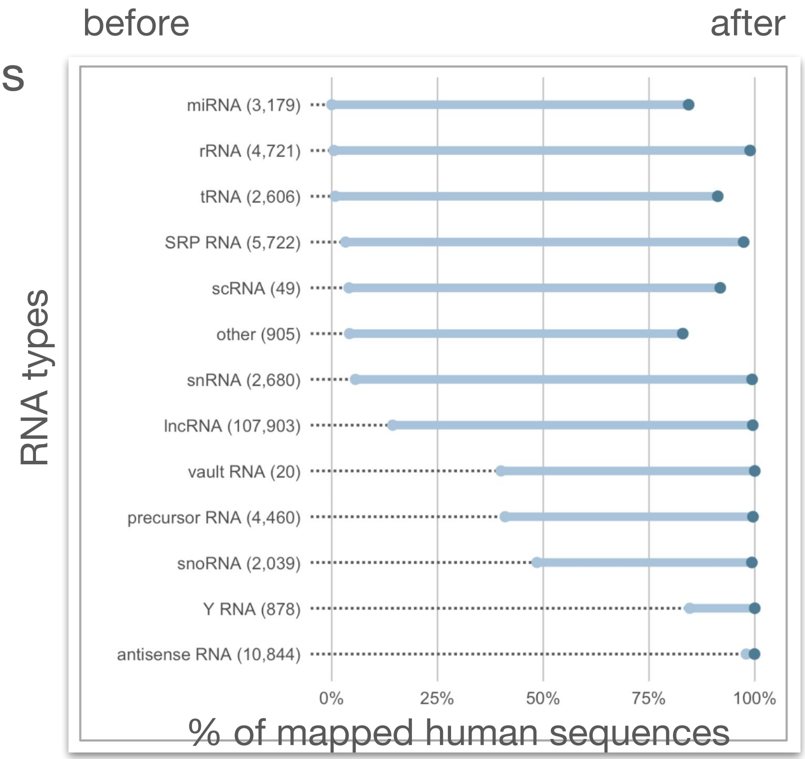
GO:0003735 structural constituent of ribosome

GO:0005840 ribosome

Matches at: **1-1,768** (41% of the model)

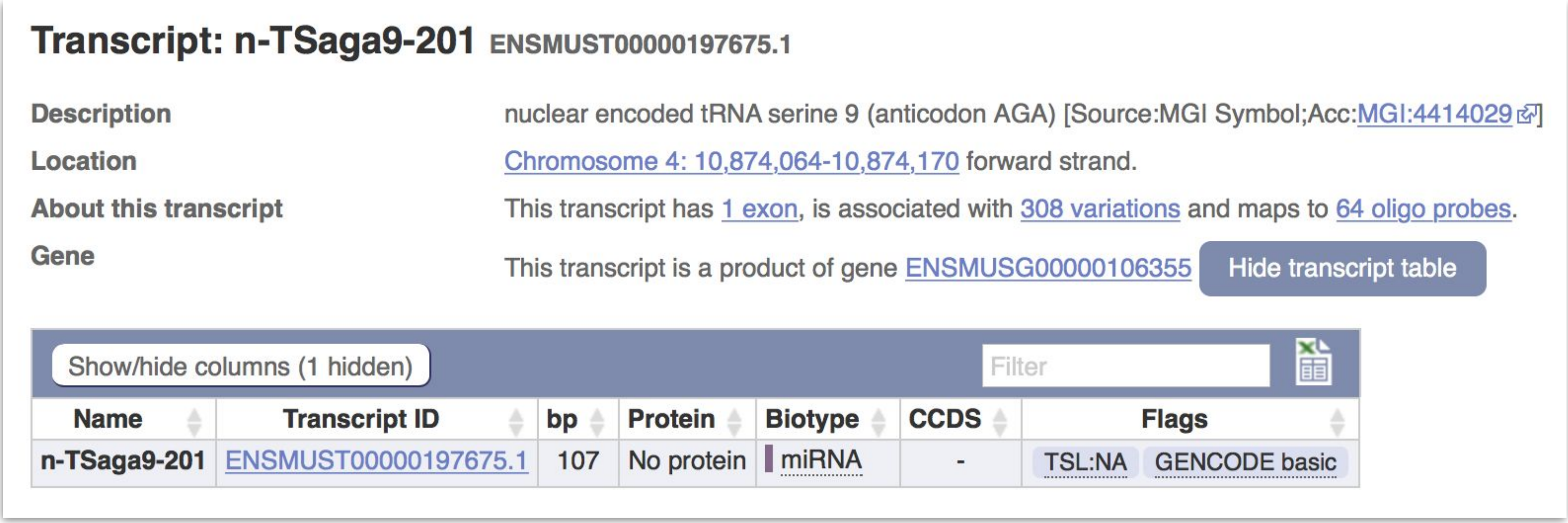
2. Comprehensive genome mapping for **>250 species**

- genome mapping Ensembl genomes and **blat**
- **>95% of sequences mapped** for human, mouse, and other key species
- one of the **largest collections** of ncRNA genome annotations



Example of Rfam quality control and genome mapping

Here is an Ensembl miRNA



Transcript: n-TSaga9-201 ENSMUST00000197675.1

Description nuclear encoded tRNA serine 9 (anticodon AGA) [Source:MGI Symbol;Acc:MGI:4414029]

Location Chromosome 4: 10,874,064-10,874,170 forward strand.

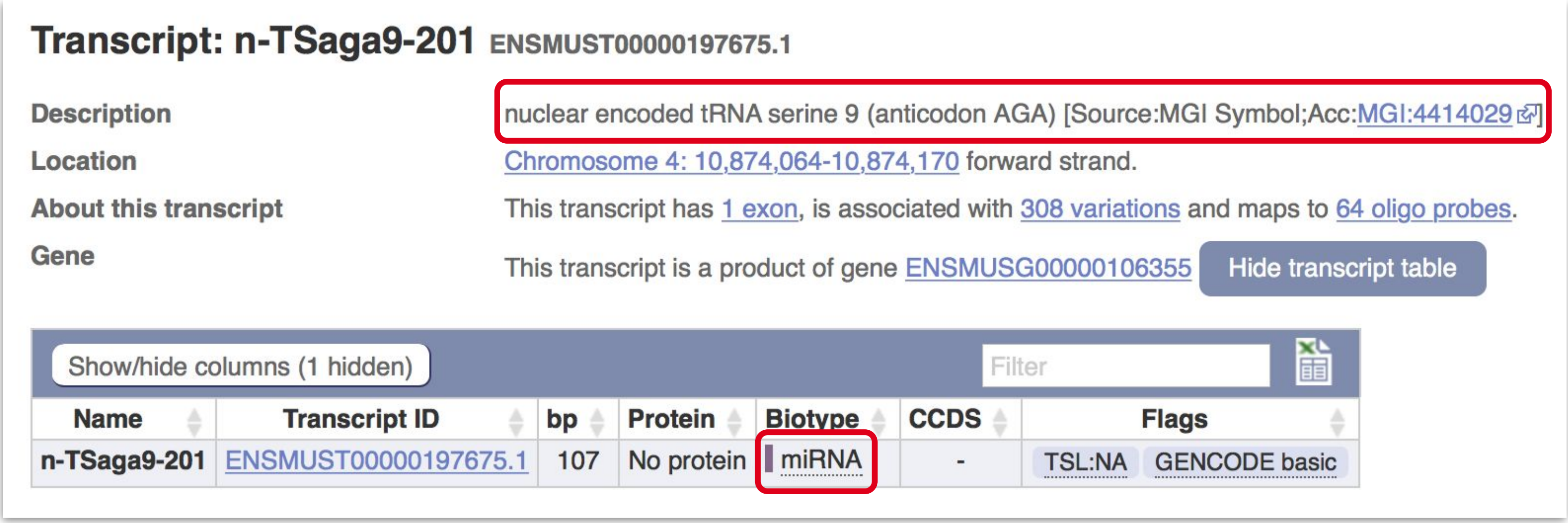
About this transcript This transcript has 1 exon, is associated with 308 variations and maps to 64 oligo probes.

Gene This transcript is a product of gene ENSMUSG00000106355

Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
n-TSaga9-201	ENSMUST00000197675.1	107	No protein	miRNA	-	TSL:NA GENCODE basic

http://www.ensembl.org/Mus_musculus/Transcript/Summary?g=ENSMUSG00000106355;r=4:10874064-10874170;t=ENSMUST00000197675

But is it a miRNA or a tRNA?



Transcript: n-TSaga9-201 ENSMUST00000197675.1

Description nuclear encoded tRNA serine 9 (anticodon AGA) [Source:MGI Symbol;Acc:MGI:4414029]

Location Chromosome 4: 10,874,064-10,874,170 forward strand.

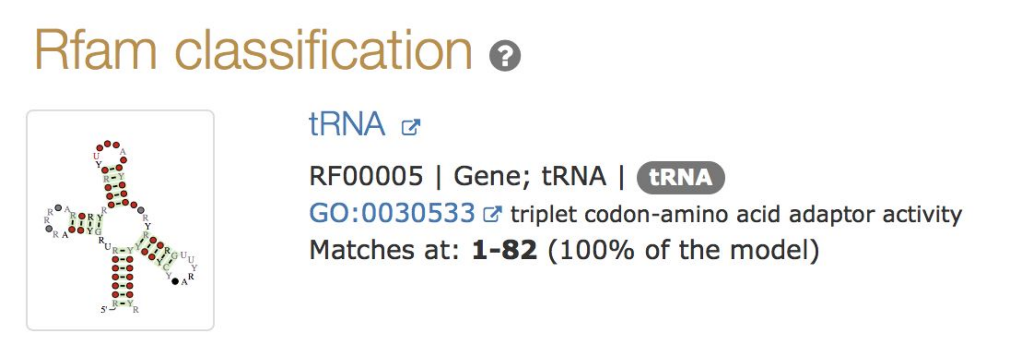
About this transcript This transcript has 1 exon, is associated with 308 variations and maps to 64 oligo probes.

Gene This transcript is a product of gene ENSMUSG00000106355

Name	Transcript ID	bp	Protein	Biotype	CCDS	Flags
n-TSaga9-201	ENSMUST00000197675.1	107	No protein	miRNA	-	TSL:NA GENCODE basic

http://www.ensembl.org/Mus_musculus/Transcript/Summary?g=ENSMUSG00000106355;r=4:10874064-10874170;t=ENSMUST00000197675

RNAcentral shows a match to a tRNA Rfam model



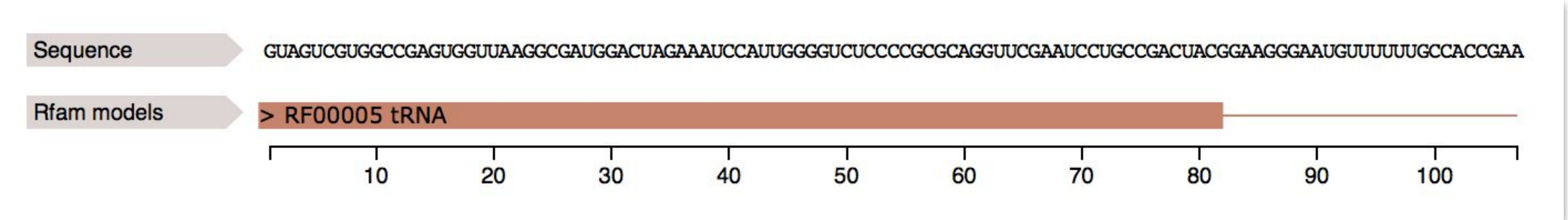
Rfam classification

tRNA

RF00005 | Gene; tRNA | tRNA

GO:0030533 triplet codon-amino acid adaptor activity

Matches at: **1-82** (100% of the model)

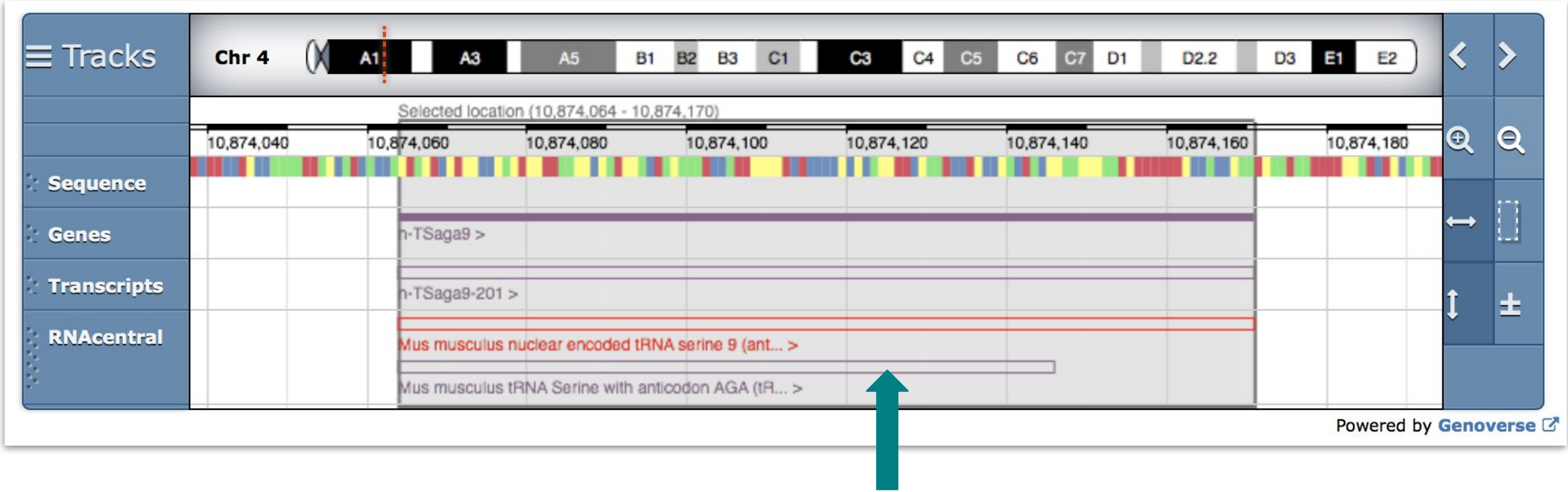


Sequence GUAGUCGUGGCCGAGUGGUUAAGGCGAUGGACUAGAAAUCCAUUGGGGUCUCCCCGCGCAGGUUCGAAUCCUGCCGACUACGGAAGGGAAUGUUUUUUGCCACCGAA

Rfam models > RF00005 tRNA

http://rnacentral.org/rna/URS0000A85A32/10090

… and other annotation in this genomic location



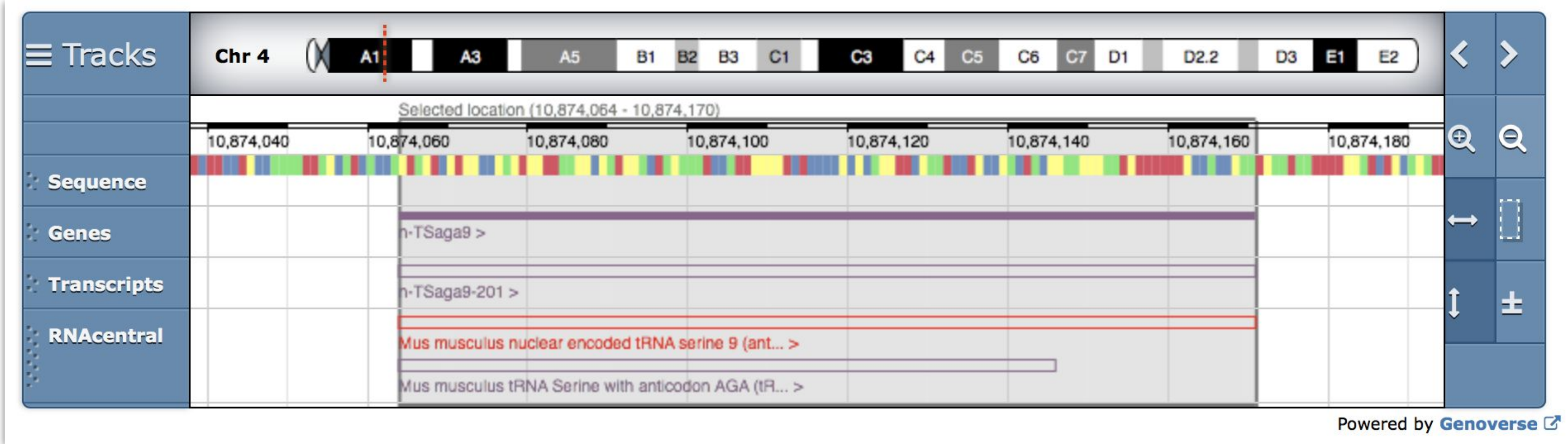
The other sequence is a well-annotated tRNA from GtRNAdb:
http://rnacentral.org/rna/URS000038D8D3/10090

RNAcentral makes data consistent across databases

- **Automatically reconcile** annotations for all sequences
- **Report** problems to member databases
- **Prioritise** sequences without inconsistencies

Overcoming important limitation of RNAcentral

- Can we **group** related sequences into “**genes**”
  using genomic location, Rfam annotations and sequence metadata?

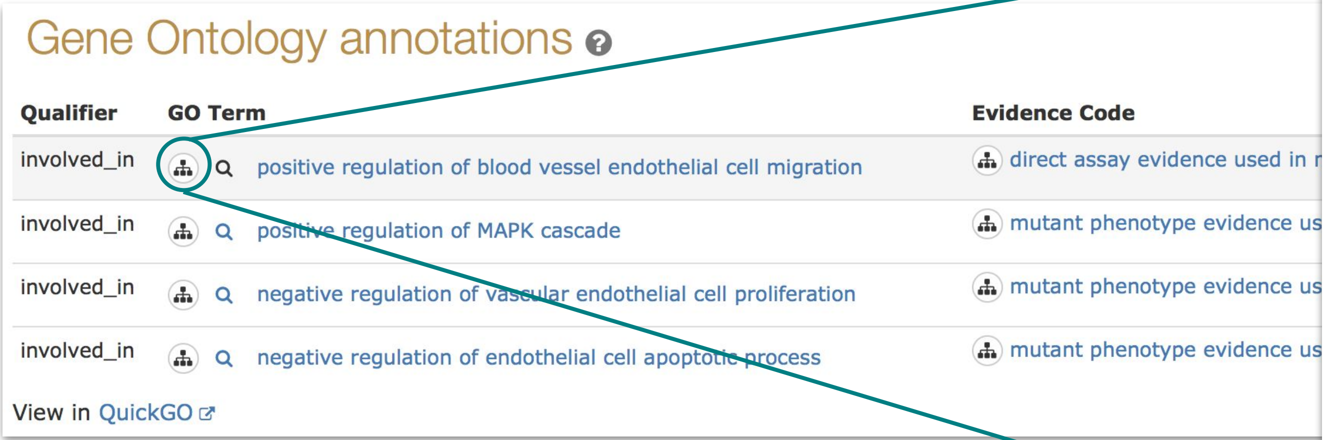


Powered by Genoverse

Adding new types of data to RNAcentral

Manually assigned **Gene Ontology terms**

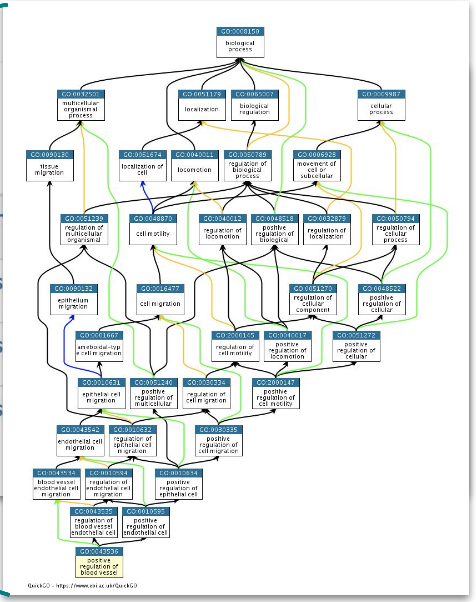
hsa-mir-126 involved in heart development:



Gene Ontology annotations ?

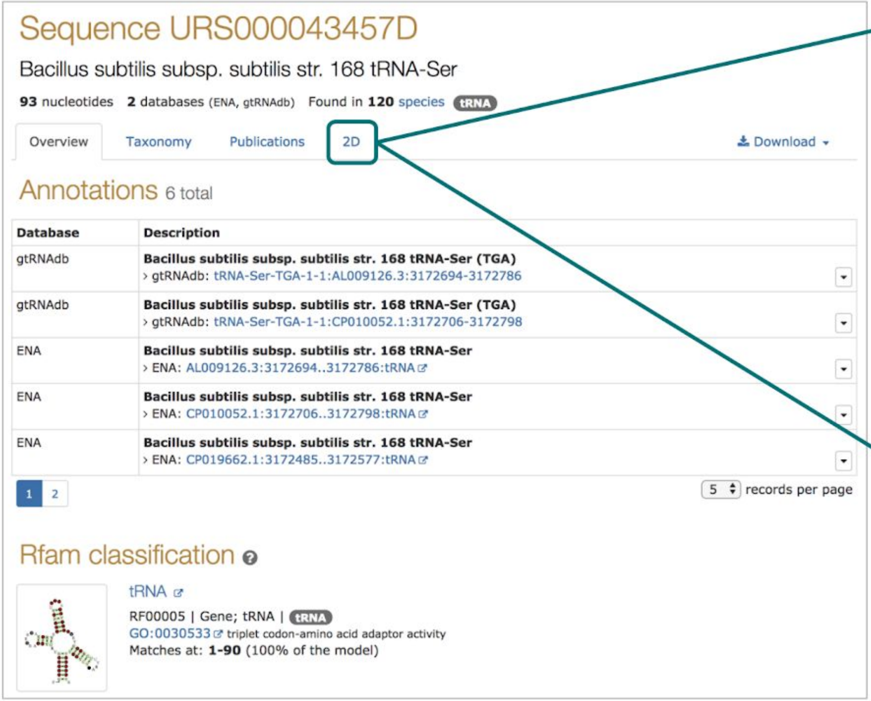
Qualifier	GO Term	Evidence Code
involved_in	positive regulation of blood vessel endothelial cell migration	direct assay evidence used in r
involved_in	positive regulation of MAPK cascade	mutant phenotype evidence us
involved_in	negative regulation of vascular endothelial cell proliferation	mutant phenotype evidence us
involved_in	negative regulation of endothelial cell apoptotic process	mutant phenotype evidence us

View in QuickGO


GO:0008150 biological process
GO:0032501 multicellular organismal process
GO:0051179 localization
GO:0065007 biological regulation
GO:0009987 cellular process
GO:0090130 tissue migration
GO:0051674 localization of cell
GO:0040011 locomotion
GO:0050789 regulation of biological process
GO:0006928 movement of cell or subcellular
GO:0051239 regulation of multicellular organismal
GO:0048870 cell motility
GO:0040012 regulation of locomotion
GO:0048518 positive regulation of biological
GO:0032879 regulation of localization
GO:0050794 regulation of cellular process
GO:0090132 epithelium migration
GO:0016477 cell migration
GO:0051270 regulation of cellular component
GO:0048522 positive regulation of cellular
GO:0001667 ameboidal-typ e cell migration
GO:2000145 regulation of cell motility
GO:0040017 positive regulation of locomotion
GO:0051272 positive regulation of cellular
GO:0010631 epithelial cell migration
GO:0051240 positive regulation of multicellular
GO:0030334 regulation of cell migration
GO:2000147 positive regulation of cell motility
GO:0043542 endothelial cell migration
GO:0010632 regulation of epithelial cell migration
GO:0030335 positive regulation of cell migration
GO:0043534 blood vessel endothelial cell migration
GO:0010594 regulation of endothelial cell migration
GO:0010634 positive regulation of epithelial cell
GO:0043535 regulation of blood vessel endothelial cell
GO:0010595 positive regulation of endothelial cell
GO:0043536 positive regulation of blood vessel
QuickGO - https://www.ebi.ac.uk/QuickGO

Huntley et al., 2018 http://rnajournal.cshlp.org/content/24/8/1005.long

Secondary structures from GtRNAdb



Sequence URS000043457D

Bacillus subtilis subsp. subtilis str. 168 tRNA-Ser

93 nucleotides **2** databases (ENA, gtRNAdb) Found in **120** species tRNA

Annotations 6 total

Database	Description
gtRNAdb	**Bacillus subtilis subsp. subtilis str. 168 tRNA-Ser (TGA)** › gtRNAdb: tRNA-Ser-TGA-1-1:AL009126.3:3172694-3172786
gtRNAdb	**Bacillus subtilis subsp. subtilis str. 168 tRNA-Ser (TGA)** › gtRNAdb: tRNA-Ser-TGA-1-1:CP010052.1:3172706-3172798
ENA	**Bacillus subtilis subsp. subtilis str. 168 tRNA-Ser** › ENA: AL009126.3:3172694..3172786:tRNA
ENA	**Bacillus subtilis subsp. subtilis str. 168 tRNA-Ser** › ENA: CP010052.1:3172706..3172798:tRNA
ENA	**Bacillus subtilis subsp. subtilis str. 168 tRNA-Ser** › ENA: CP019662.1:3172485..3172577:tRNA

5 records per page

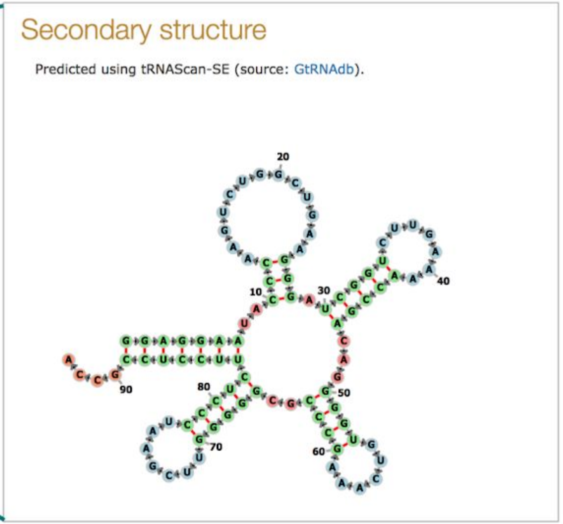
Rfam classification

tRNA

RF00005 | Gene; tRNA | tRNA

GO:0030533 triplet codon-amino acid adaptor activity

Matches at: **1-90** (100% of the model)



Secondary structure

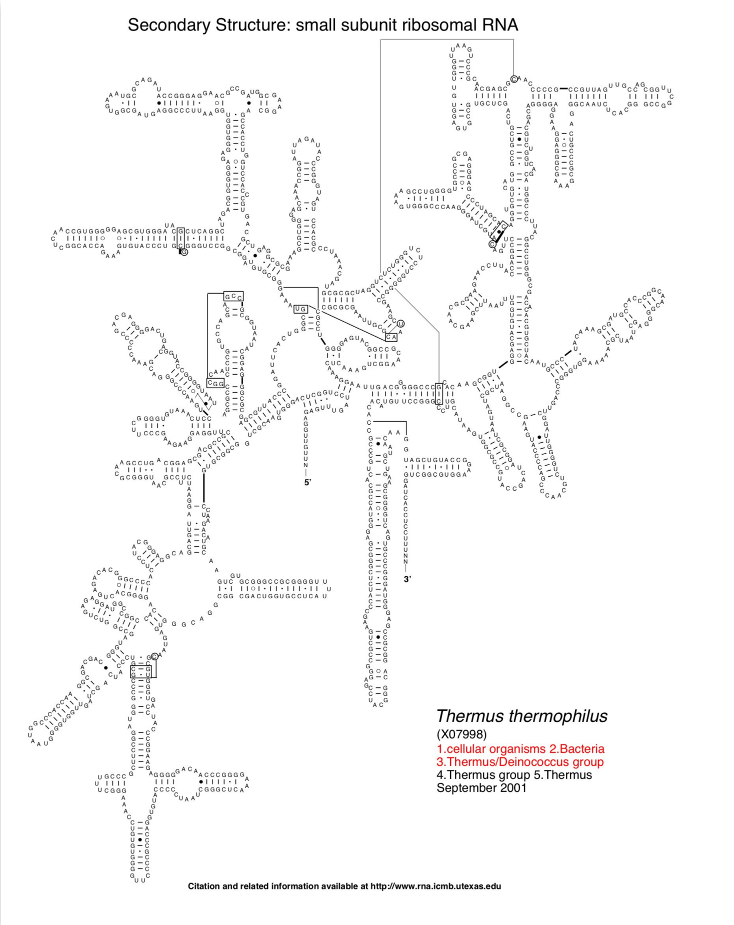
Predicted using tRNAScan-SE (source: GtRNAdb).

Benasque 2018

top-secret project:

display 2Ds for all rRNAs
in RNAcentral
using **standard layouts**

This is how RNA biologists want to see rRNA 2D

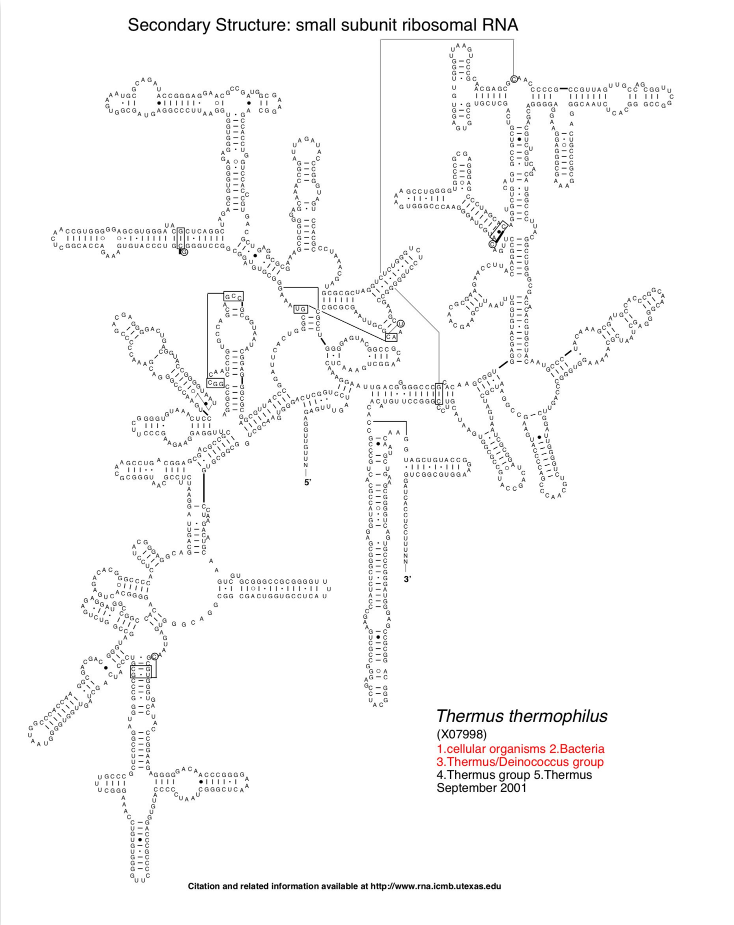


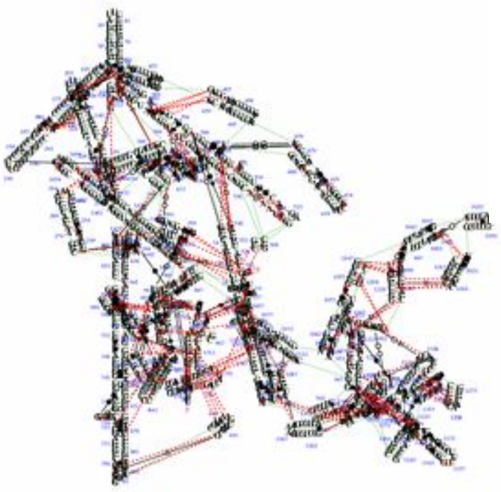
Secondary Structure: small subunit ribosomal RNA

Thermus thermophilus
(X07998)
1.cellular organisms 2.Bacteria
3.Thermus/Deinococcus group
4.Thermus group 5.Thermus
September 2001

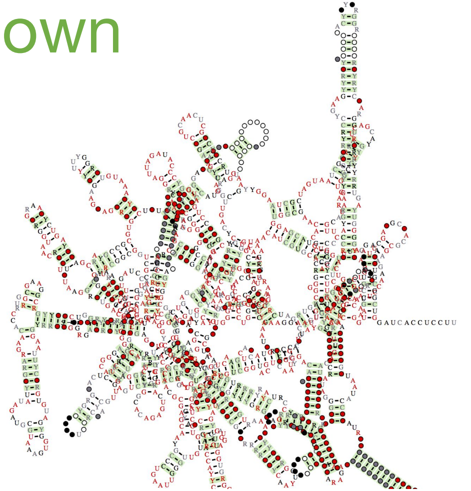
Citation and related information available at http://www.rna.icmb.utexas.edu

This is how it is shown in RNA databases

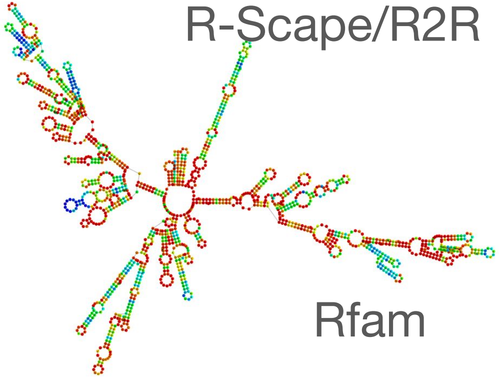

Secondary Structure: small subunit ribosomal RNA
5'
3'
Thermus thermophilus
(X07998)
1.cellular organisms 2.Bacteria
3.Thermus/Deinococcus group
4.Thermus group 5.Thermus
September 2001
Citation and related information available at http://www.rna.icmb.utexas.edu



NDB



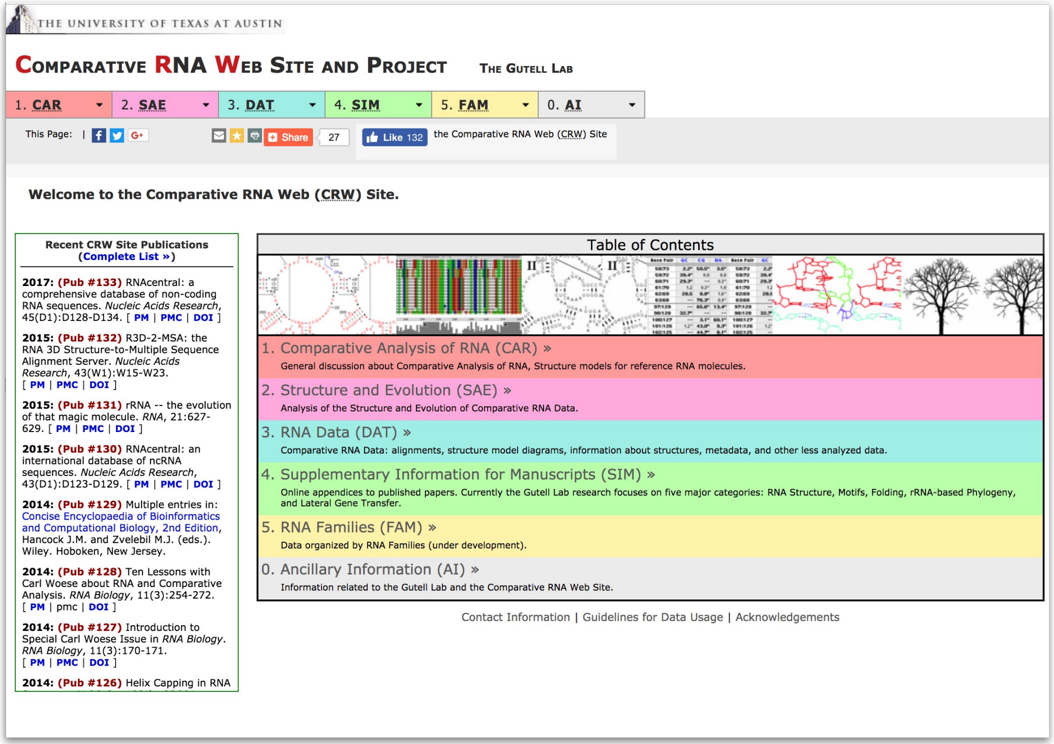
R-Scape/R2R



Rfam

1. **Comparative RNA Website** (Robin Gutell's lab)

- Contains ~1,000 of RNA secondary structures in standard layouts
- Authoritative source of rRNA data



THE UNIVERSITY OF TEXAS AT AUSTIN

COMPARATIVE RNA WEB SITE AND PROJECT THE GUTELL LAB

1. CAR ▾ | 2. SAE ▾ | 3. DAT ▾ | 4. SIM ▾ | 5. FAM ▾ | 0. AI ▾

This Page: | f t G+ | Share 27 | Like 132 the Comparative RNA Web (CRW) Site

Welcome to the Comparative RNA Web (CRW) Site.

Recent CRW Site Publications
(Complete List »)

2017: (Pub #133) RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Research*, 45(D1):D128-D134. [**PM** | **PMC** | **DOI**]

2015: (Pub #132) R3D-2-MSA: the RNA 3D Structure-to-Multiple Sequence Alignment Server. *Nucleic Acids Research*, 43(W1):W15-W23. [**PM** | **PMC** | **DOI**]

2015: (Pub #131) rRNA -- the evolution of that magic molecule. *RNA*, 21:627-629. [**PM** | **PMC** | **DOI**]

2015: (Pub #130) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Research*, 43(D1):D123-D129. [**PM** | **PMC** | **DOI**]

2014: (Pub #129) Multiple entries in: Concise Encyclopaedia of Bioinformatics and Computational Biology, 2nd Edition, Hancock J.M. and Zvelebil M.J. (eds.). Wiley. Hoboken, New Jersey.

2014: (Pub #128) Ten Lessons with Carl Woese about RNA and Comparative Analysis. *RNA Biology*, 11(3):254-272. [**PM** | pmc | **DOI**]

2014: (Pub #127) Introduction to Special Carl Woese Issue in *RNA Biology*. *RNA Biology*, 11(3):170-171. [**PM** | **PMC** | **DOI**]

2014: (Pub #126) Helix Capping in RNA

Table of Contents

1. Comparative Analysis of RNA (CAR) »
General discussion about Comparative Analysis of RNA, Structure models for reference RNA molecules.

2. Structure and Evolution (SAE) »
Analysis of the Structure and Evolution of Comparative RNA Data.

3. RNA Data (DAT) »
Comparative RNA Data: alignments, structure model diagrams, information about structures, metadata, and other less analyzed data.

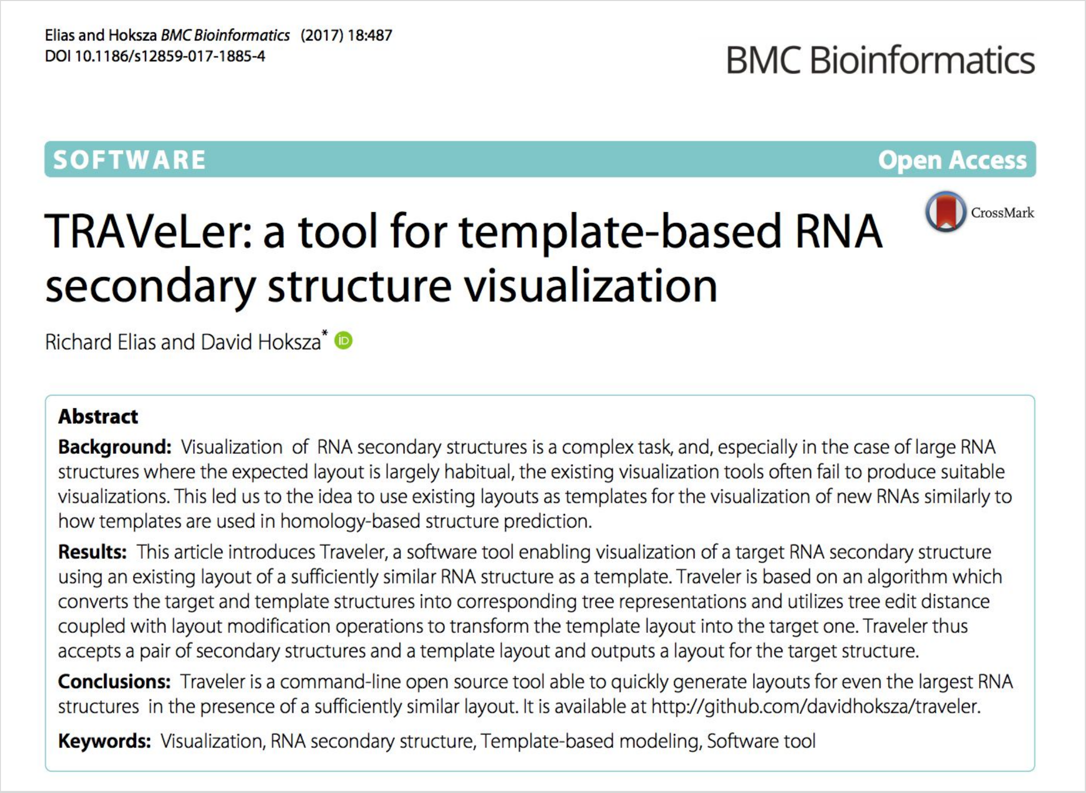
4. Supplementary Information for Manuscripts (SIM) »
Online appendices to published papers. Currently the Gutell Lab research focuses on five major categories: RNA Structure, Motifs, Folding, rRNA-based Phylogeny, and Lateral Gene Transfer.

5. RNA Families (FAM) »
Data organized by RNA Families (under development).

0. Ancillary Information (AI) »
Information related to the Gutell Lab and the Comparative RNA Web Site.

Contact Information | Guidelines for Data Usage | Acknowledgements

2. **Traveler** software



Elias and Hoksza *BMC Bioinformatics* (2017) 18:487
DOI 10.1186/s12859-017-1885-4

BMC Bioinformatics

SOFTWARE Open Access

CrossMark

TRAVeLer: a tool for template-based RNA secondary structure visualization

Richard Elias and David Hoksza*

Abstract

Background: Visualization of RNA secondary structures is a complex task, and, especially in the case of large RNA structures where the expected layout is largely habitual, the existing visualization tools often fail to produce suitable visualizations. This led us to the idea to use existing layouts as templates for the visualization of new RNAs similarly to how templates are used in homology-based structure prediction.

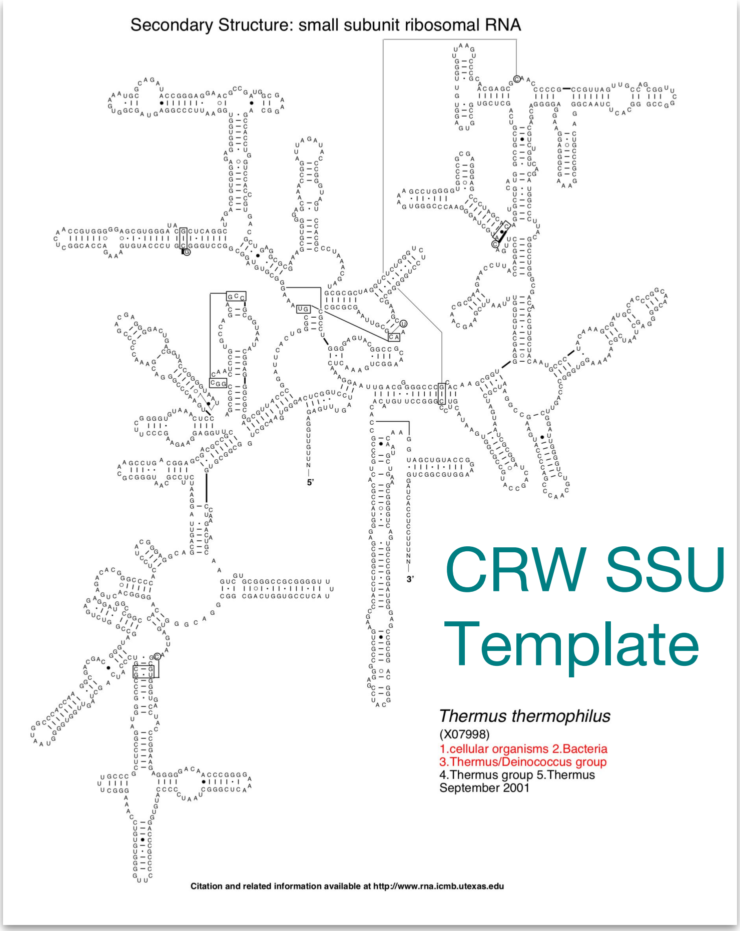
Results: This article introduces Traveler, a software tool enabling visualization of a target RNA secondary structure using an existing layout of a sufficiently similar RNA structure as a template. Traveler is based on an algorithm which converts the target and template structures into corresponding tree representations and utilizes tree edit distance coupled with layout modification operations to transform the template layout into the target one. Traveler thus accepts a pair of secondary structures and a template layout and outputs a layout for the target structure.

Conclusions: Traveler is a command-line open source tool able to quickly generate layouts for even the largest RNA structures in the presence of a sufficiently similar layout. It is available at http://github.com/davidhoksza/traveler.

Keywords: Visualization, RNA secondary structure, Template-based modeling, Software tool

David Hoksza

Assistant Professor
Charles University, Prague



Secondary Structure: small subunit ribosomal RNA

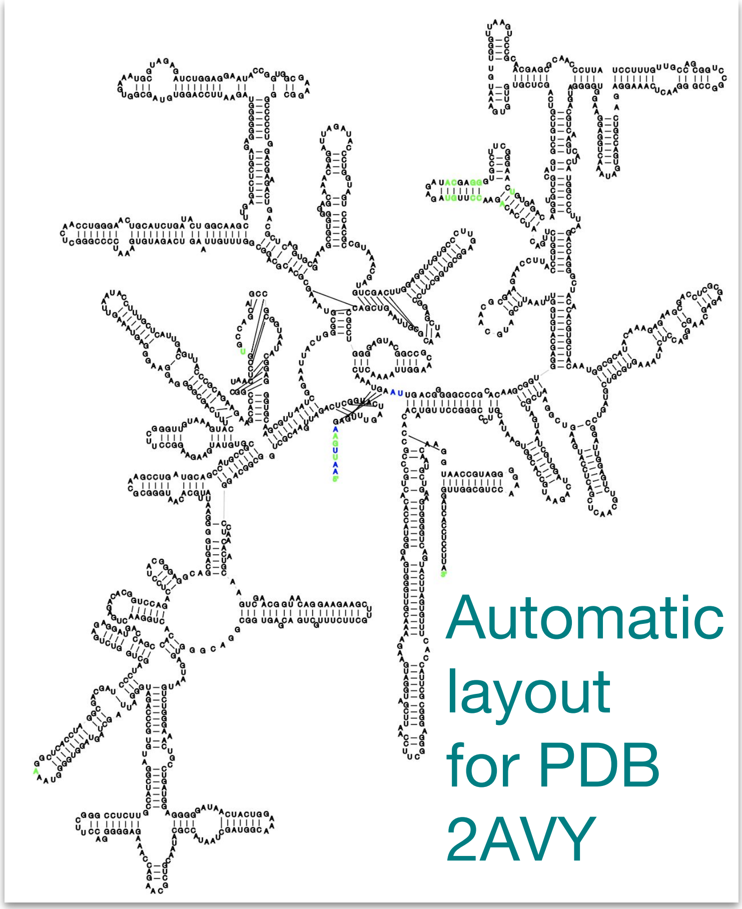
Thermus thermophilus

(X07998)

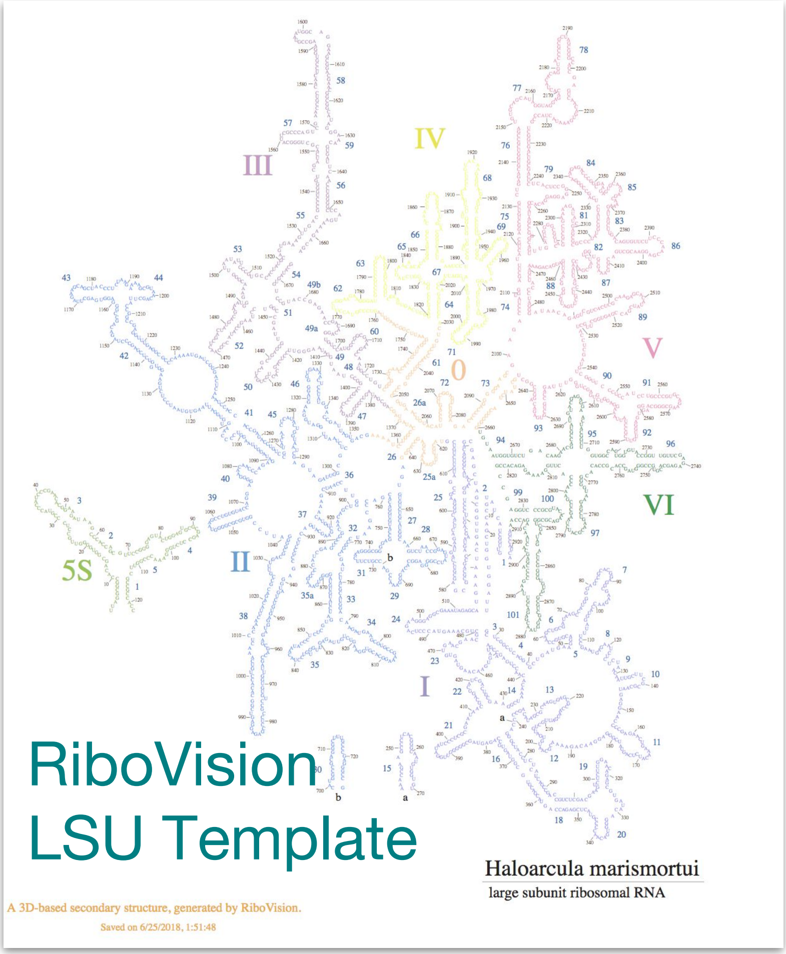
1.cellular organisms 2.Bacteria
3.Thermus/Deinococcus group
4.Thermus group 5.Thermus
September 2001

Citation and related information available at http://www.rna.icmb.utexas.edu

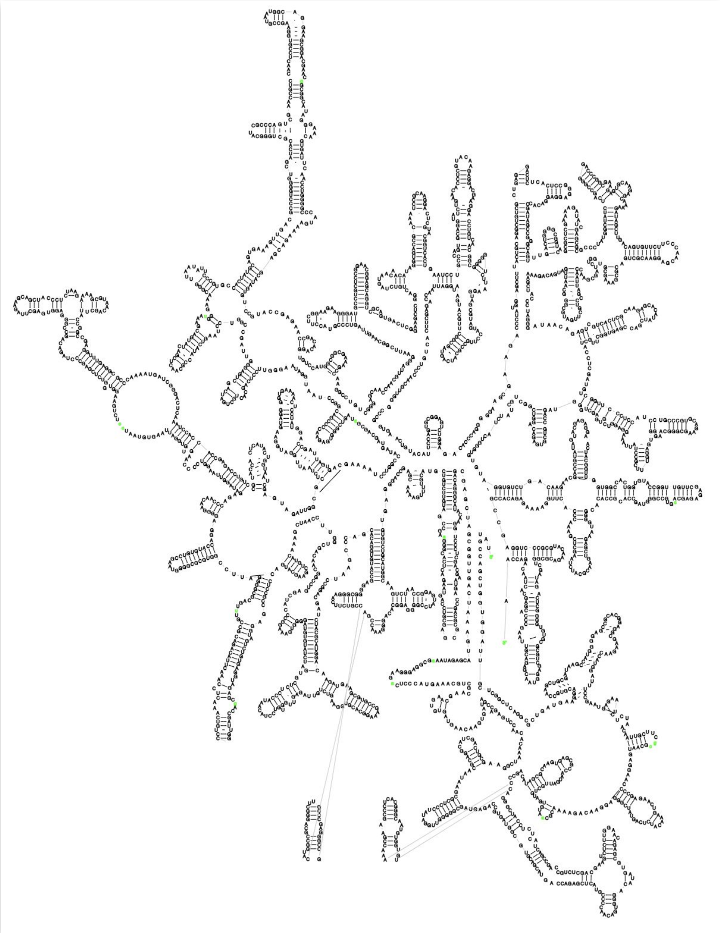
CRW SSU Template



Automatic layout for PDB 2AVY


III
IV
V
0
VI
II
I
5S
Haloarcula marismortui
large subunit ribosomal RNA
A 3D-based secondary structure, generated by RiboVision.
Saved on 6/25/2018, 1:51:48

RiboVision LSU Template



Automatic layout for PDB 1S72

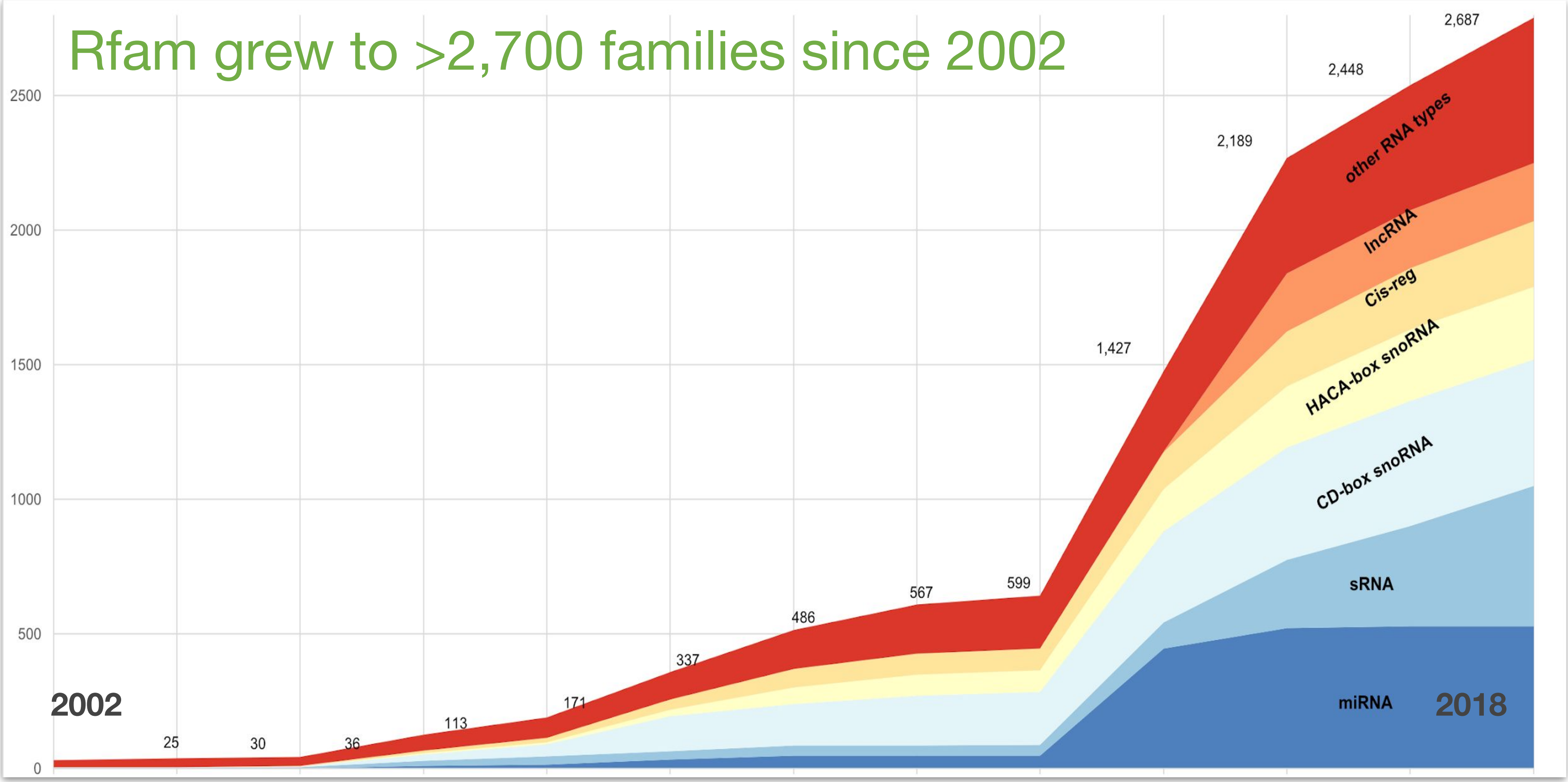
RNAcentral keeps evolving

- Try it and send us your **feedback**
- Help us **improve**

Rfam

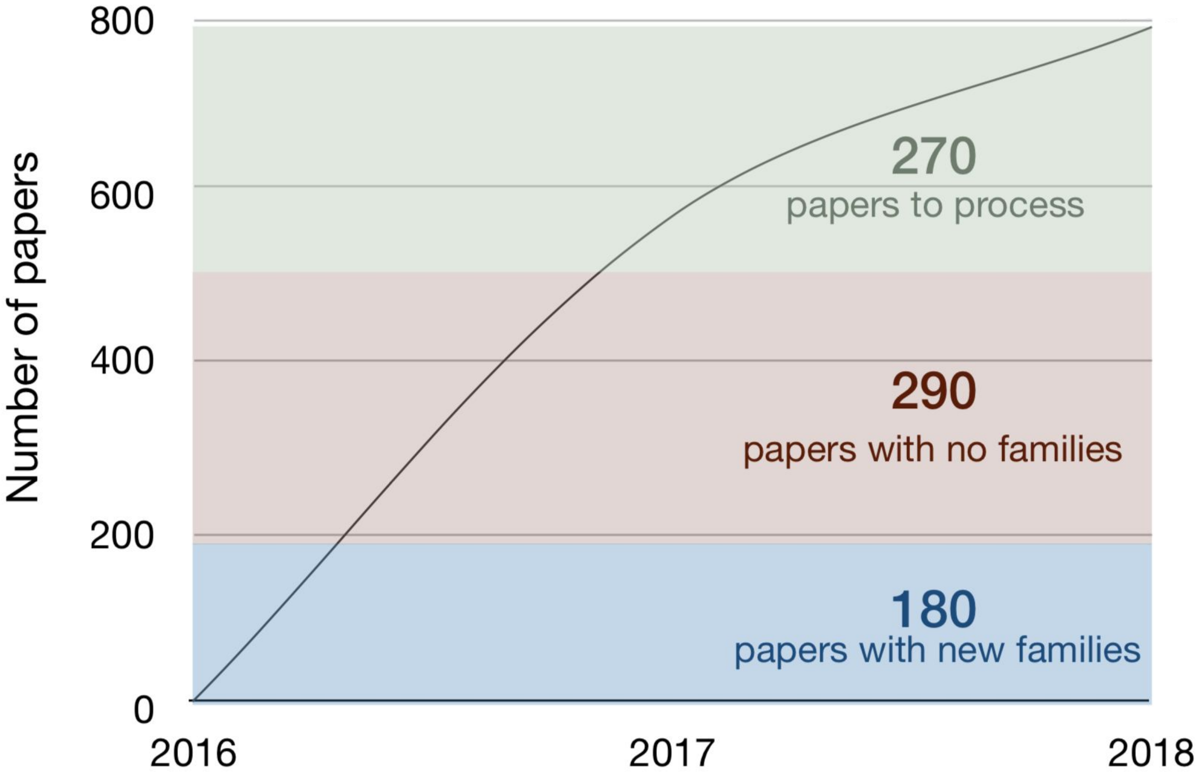
The transition to a genome-centric database of ncRNA families

Rfam grew to >2,700 families since 2002



EMBL-EBI

There are still more RNA families to be added



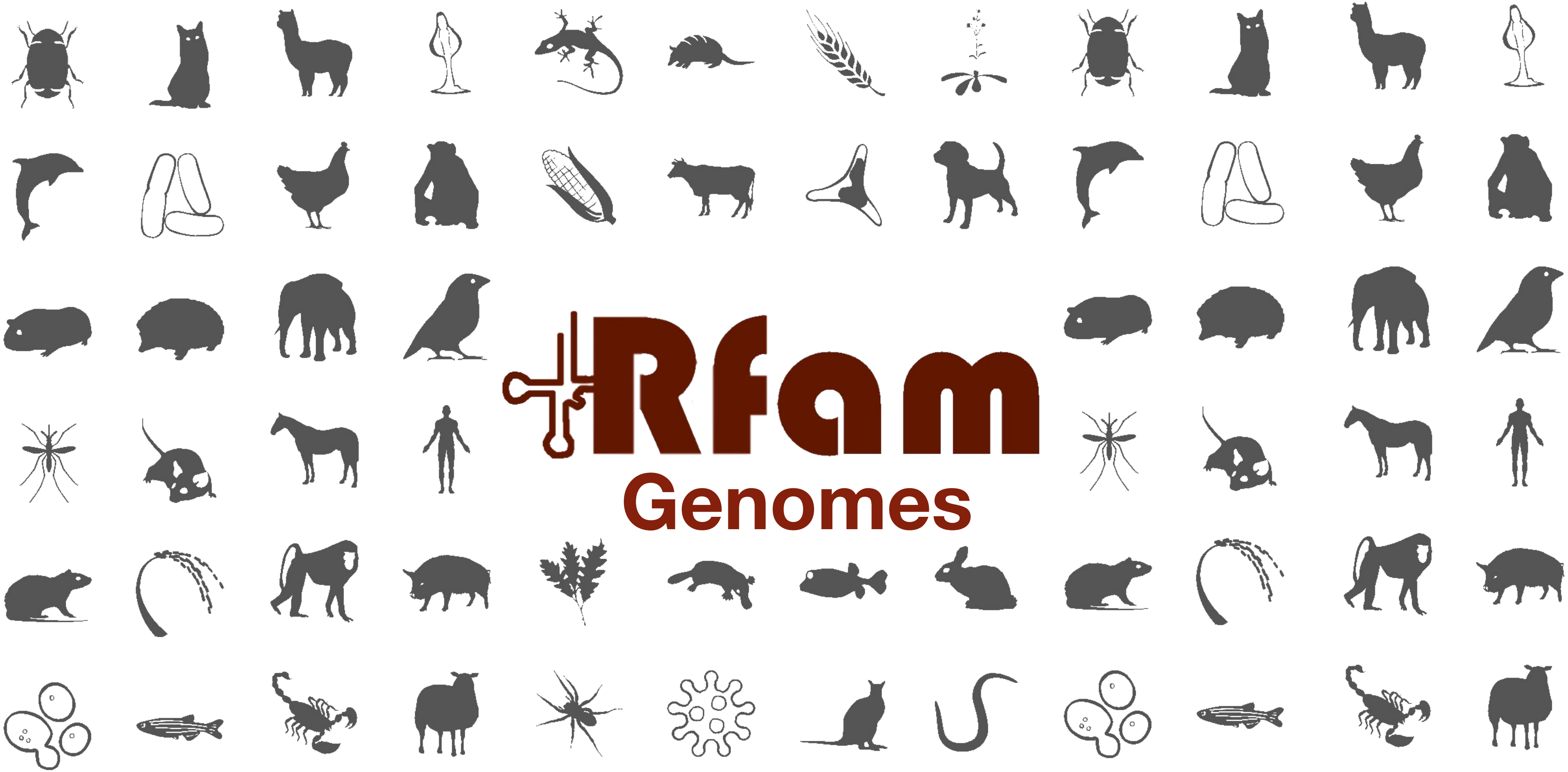
- **350** new families created since 2016
- **30%** of papers still waiting to be curated

What’s new in Rfam

Can you guess

the **most common user request**

in Rfam 12.*?



Rfam

Genomes

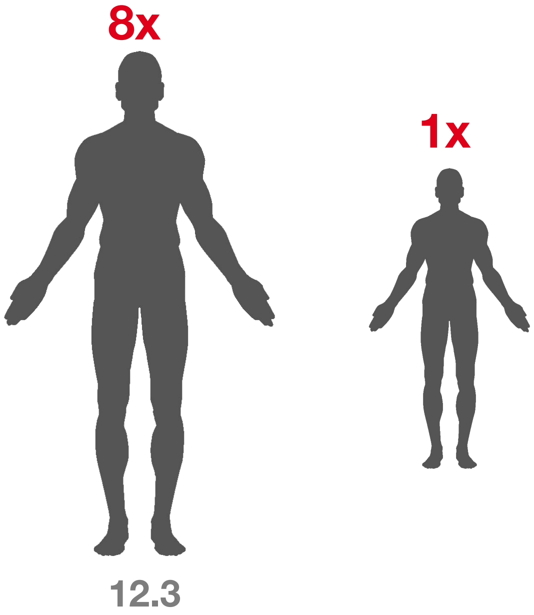
Previously Rfam analysed WGS and STD sequences from ENA and GenBank

- The data were **redundant**
- Taxonomic comparisons were **difficult**

Redundant strains

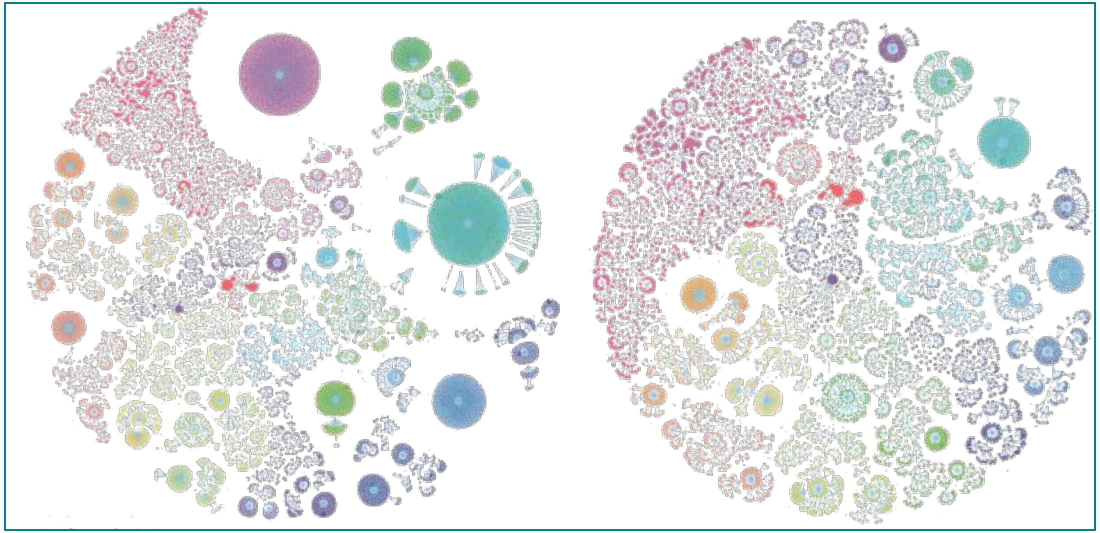
	Strains
Mycobacterium tuberculosis	77
Escherichia coli	348

Over-represented genomes



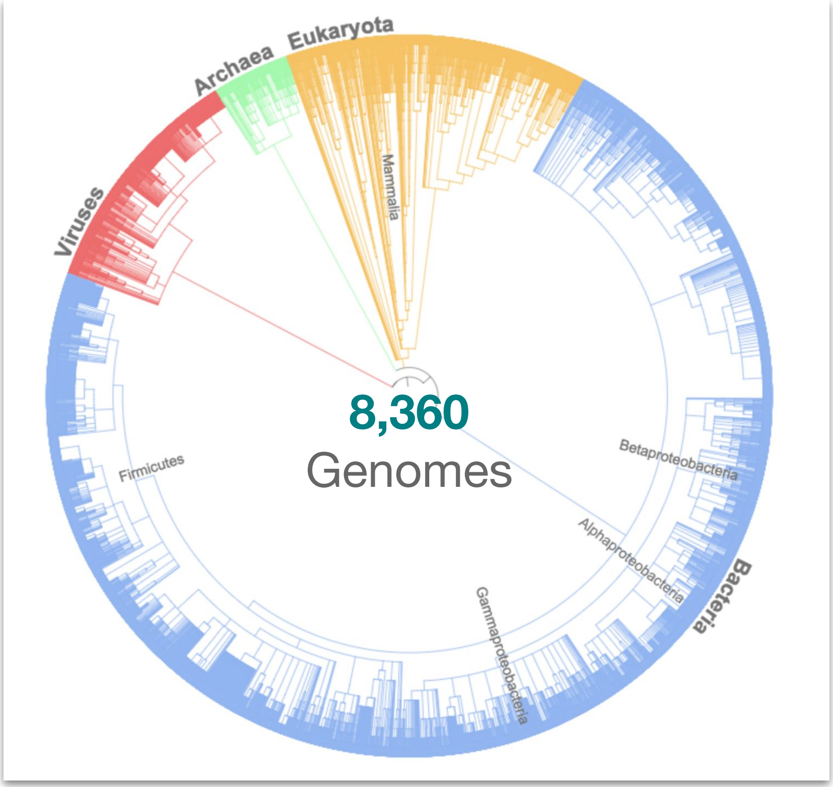
Rfam now annotates **complete**, **representative**, and **non-redundant** genomes

Based on the UniProt Reference Proteome collection



https://doi.org/10.1093/database/baw139

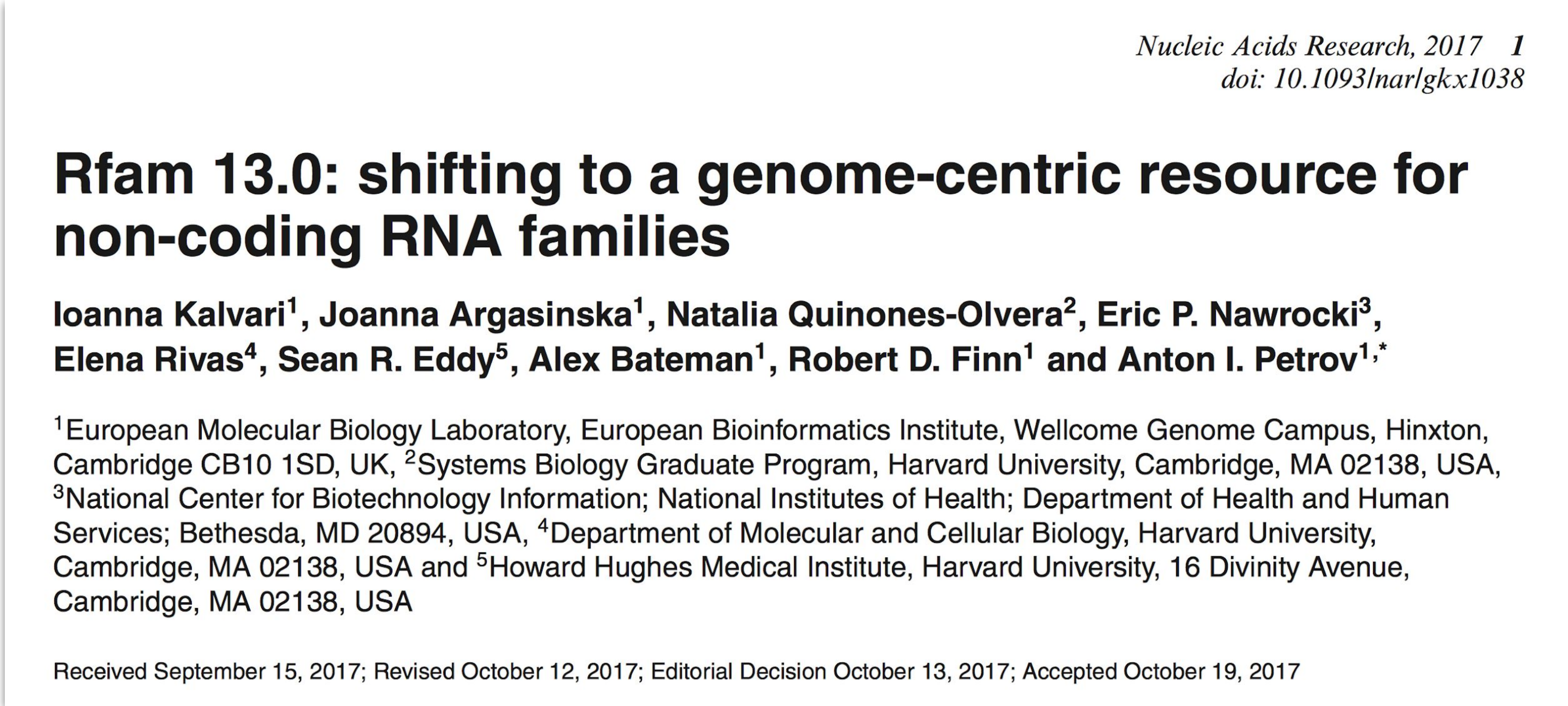
Rfam 13.0 is based on **8,360 genomes**



Need to further expand Rfam sequence database

- Viruses
- Metagenomes
- 3D structures
- RNAcentral sequences that do not match Rfam families

Find out more about genome-centric Rfam



Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx1038

Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families

Ioanna Kalvari[1], Joanna Argasinska[1], Natalia Quinones-Olvera[2], Eric P. Nawrocki[3], Elena Rivas[4], Sean R. Eddy[5], Alex Bateman[1], Robert D. Finn[1] and Anton I. Petrov[1,*]

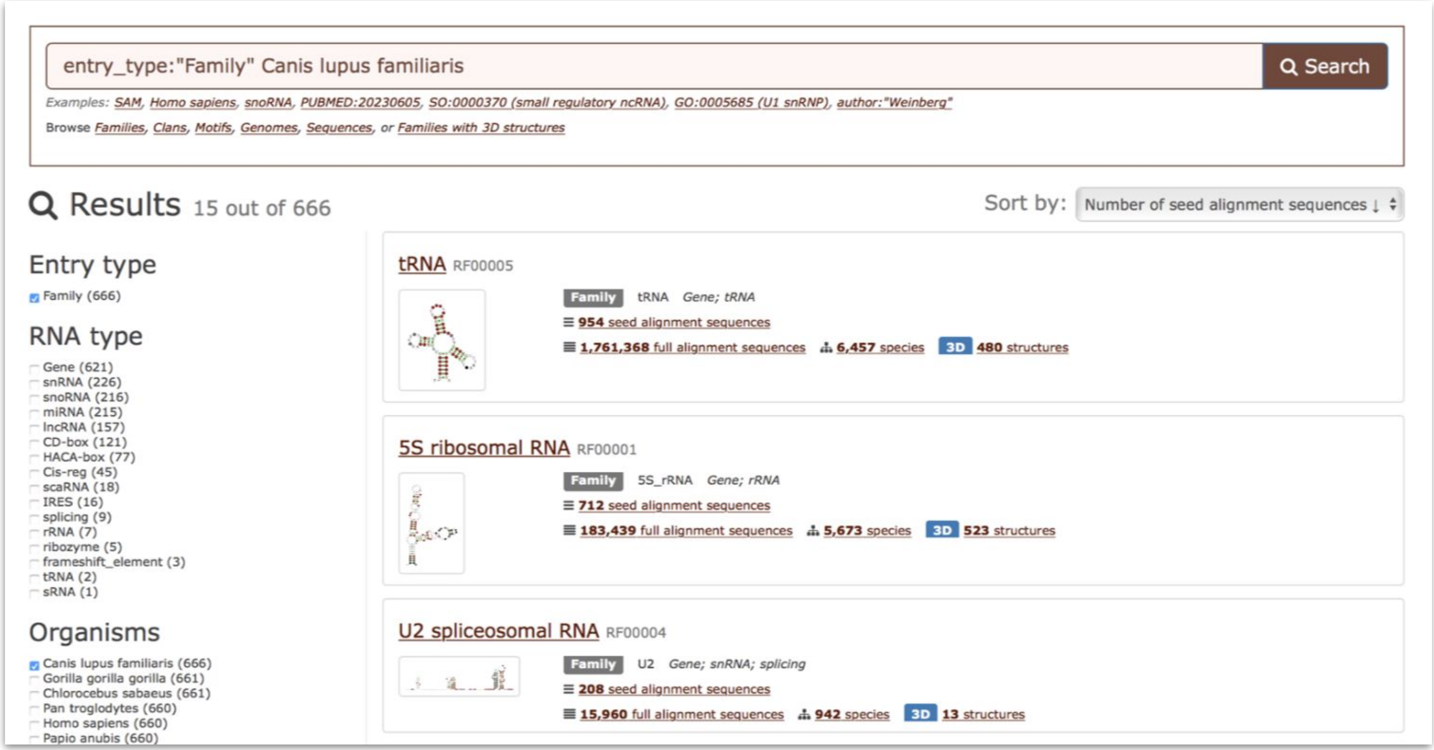
[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Systems Biology Graduate Program, Harvard University, Cambridge, MA 02138, USA, [3]National Center for Biotechnology Information; National Institutes of Health; Department of Health and Human Services; Bethesda, MD 20894, USA, [4]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA and [5]Howard Hughes Medical Institute, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

Received September 15, 2017; Revised October 12, 2017; Editorial Decision October 13, 2017; Accepted October 19, 2017

https://academic.oup.com/nar/article/4588106

New website functionality

New **faceted** text search and **search API**



entry_type:"Family" Canis lupus familiaris

Examples: *SAM*, *Homo sapiens*, *snoRNA*, *PUBMED:20230605*, *SO:0000370 (small regulatory ncRNA)*, *GO:0005685 (U1 snRNP)*, *author:"Weinberg"*

Browse *Families*, *Clans*, *Motifs*, *Genomes*, *Sequences*, *or* *Families with 3D structures*

Results 15 out of 666

Sort by: Number of seed alignment sequences ↓

Entry type

☑ Family (666)

RNA type

- □ Gene (621)
- □ snRNA (226)
- □ snoRNA (216)
- □ miRNA (215)
- □ lncRNA (157)
- □ CD-box (121)
- □ HACA-box (77)
- □ Cis-reg (45)
- □ scaRNA (18)
- □ IRES (16)
- □ splicing (9)
- □ rRNA (7)
- □ ribozyme (5)
- □ frameshift_element (3)
- □ tRNA (2)
- □ sRNA (1)

Organisms

- ☑ Canis lupus familiaris (666)
- □ Gorilla gorilla gorilla (661)
- □ Chlorocebus sabaeus (661)
- □ Pan troglodytes (660)
- □ Homo sapiens (660)
- □ Papio anubis (660)

tRNA RF00005

Family tRNA *Gene; tRNA*

954 seed alignment sequences

1,761,368 full alignment sequences 6,457 species 3D 480 structures

5S ribosomal RNA RF00001

Family 5S_rRNA *Gene; rRNA*

712 seed alignment sequences

183,439 full alignment sequences 5,673 species 3D 523 structures

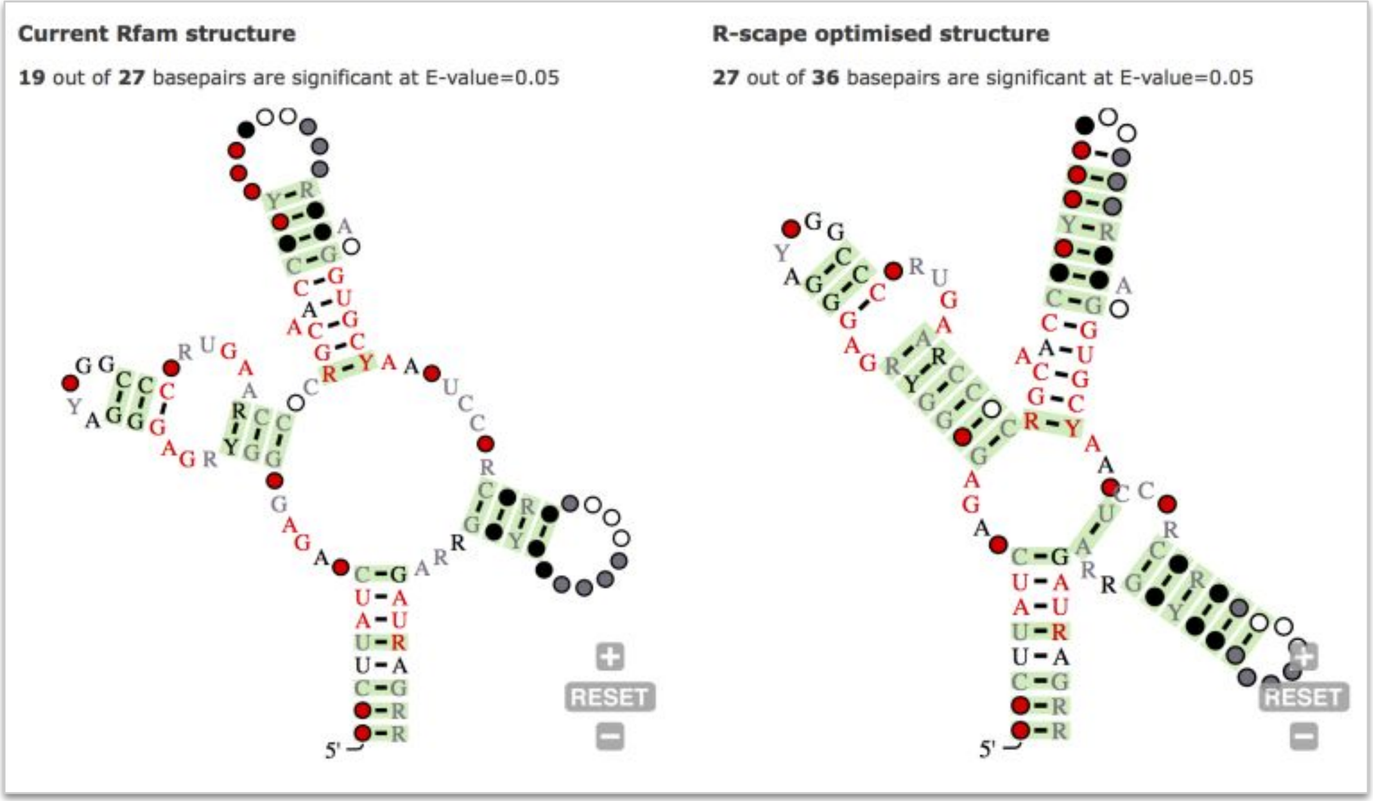
U2 spliceosomal RNA RF00004

Family U2 *Gene; snRNA; splicing*

208 seed alignment sequences

15,960 full alignment sequences 942 species 3D 13 structures

R-Scape visualisations



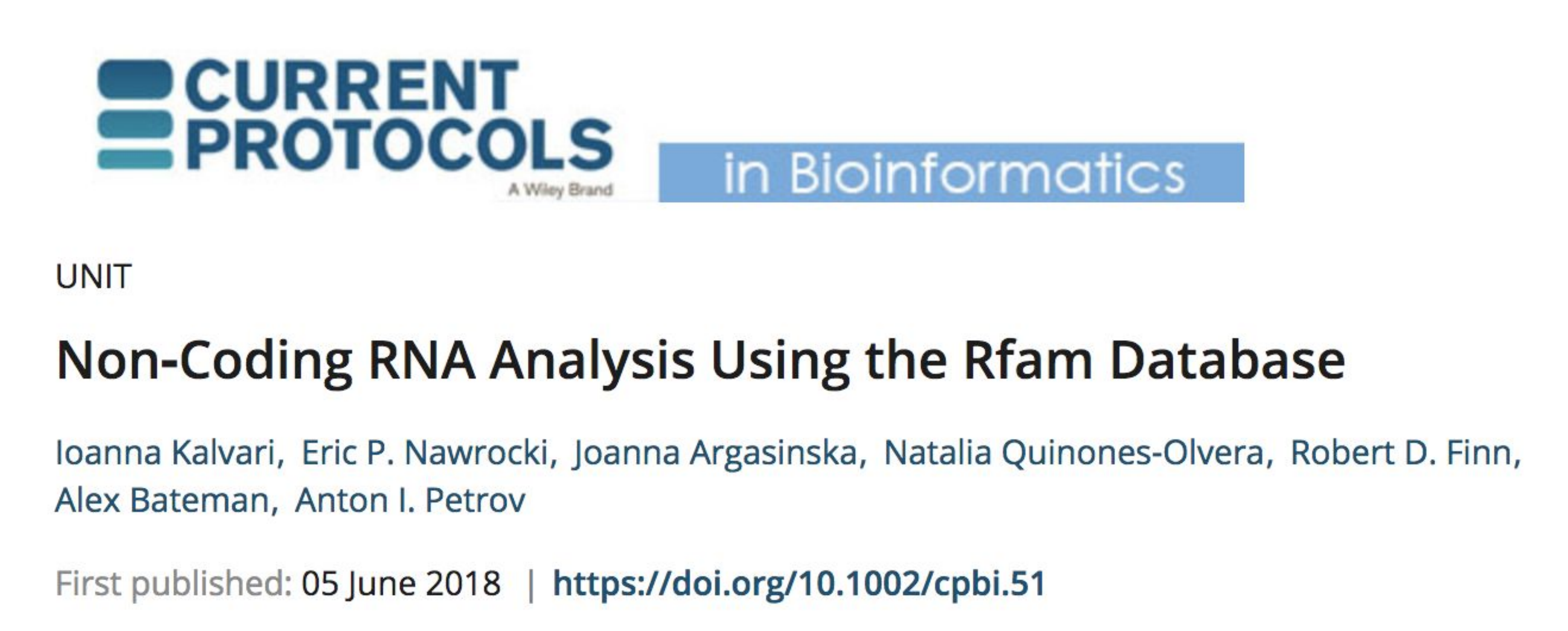
Current Rfam structure

19 out of **27** basepairs are significant at E-value=0.05

R-scape optimised structure

27 out of **36** basepairs are significant at E-value=0.05

Find how to search Rfam, query public MySQL database, and more



CURRENT PROTOCOLS
A Wiley Brand
in Bioinformatics

UNIT

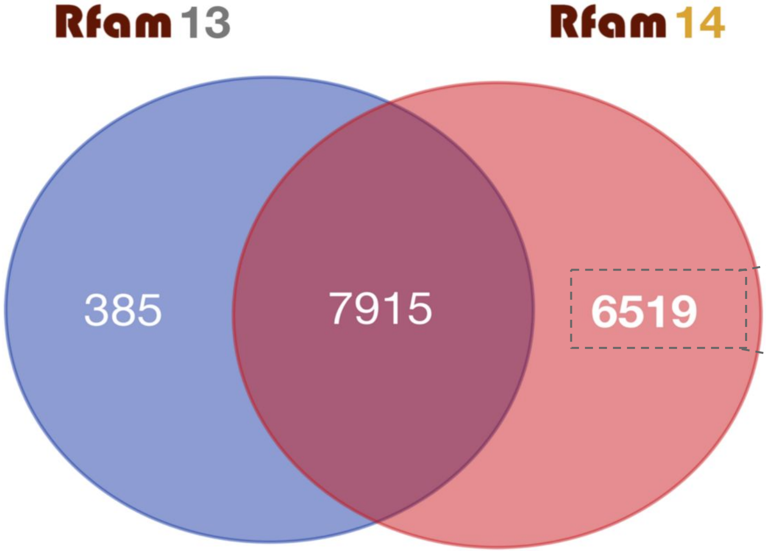
Non-Coding RNA Analysis Using the Rfam Database

Ioanna Kalvari, Eric P. Nawrocki, Joanna Argasinska, Natalia Quinones-Olvera, Robert D. Finn, Alex Bateman, Anton I. Petrov

First published: 05 June 2018 | https://doi.org/10.1002/cpbi.51

https://doi.org/10.1002/cpbi.51

Upcoming Rfam 14.0 annotates **60% more genomes**



New species

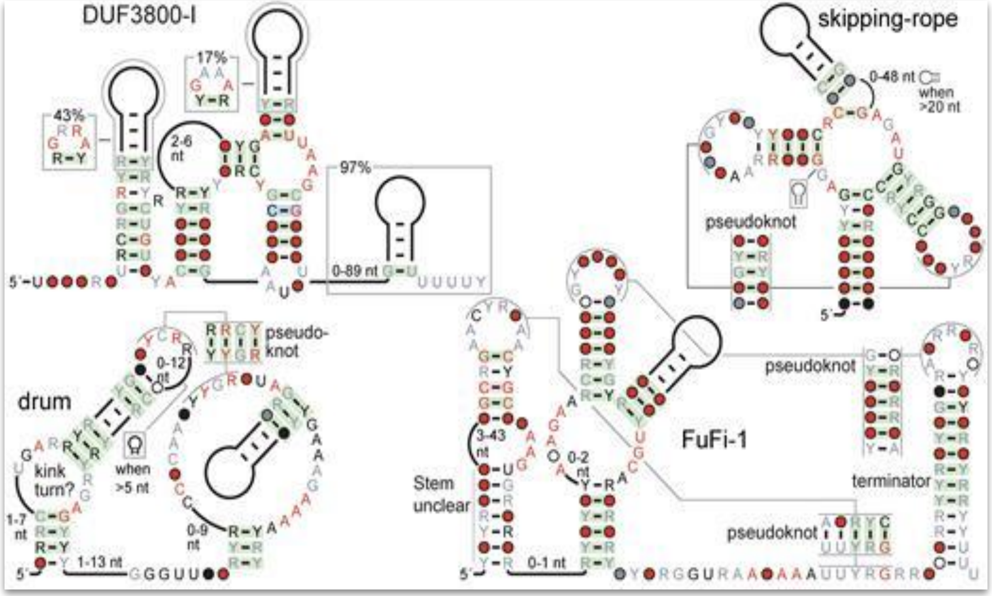
180	Eukaryotes
206	Archaea
3160	Bacteria
2973	Viruses

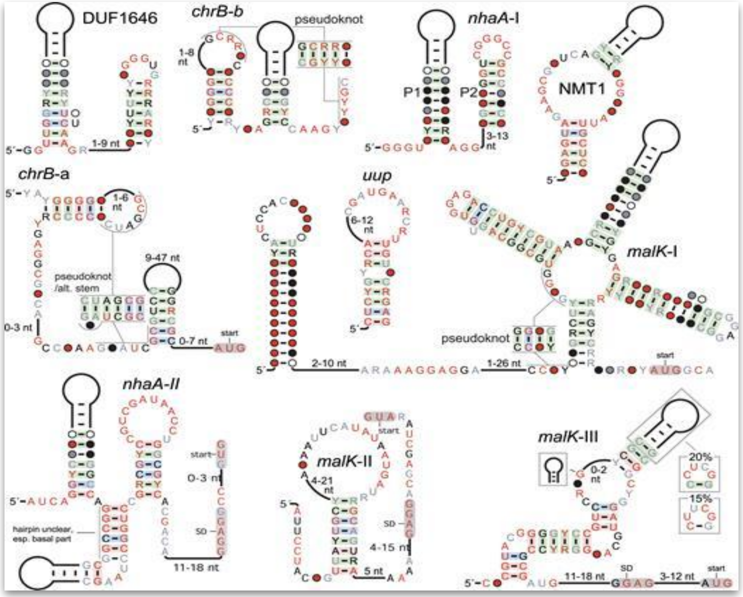
Preview at **rfamlive.xfam.org**

8,360 genomes → **14,434** genomes

Expect Rfam 14.1 later this year

No new genomes but lots of new families from Zasha Weinberg





https://doi.org/10.1093/nar/gkx699

Do you want to build Rfam families?

- Family curation by approved **experts**
- **Cloud-based** Rfam pipeline
- Command line or **Galaxy** access

Special session

Wednesday 3pm

Acknowledgements

BBSRC
bioscience for the future

EMBL-EBI

Alex Bateman
Robert Finn
Ioanna Kalvari
Joanna Argasinska
Blake Sweeney
Boris Burkov

EMBL-EBI

NCBI

Eric Nawrocki

NIH NLM

Harvard University

Sean Eddy
Elena Rivas
Natalia Quinones-Olvera

VE RI TAS
HARVARD