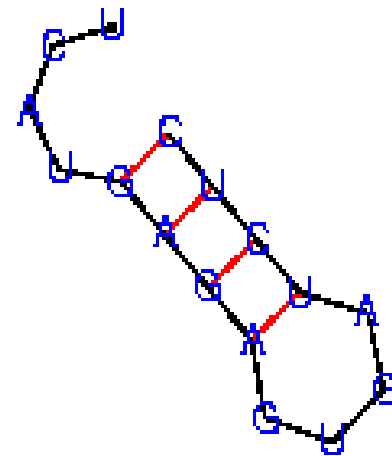


Prediction & annotation of ncRNA

C. Gaspin

Unité de Mathématique et Informatique Appliqués
Toulouse

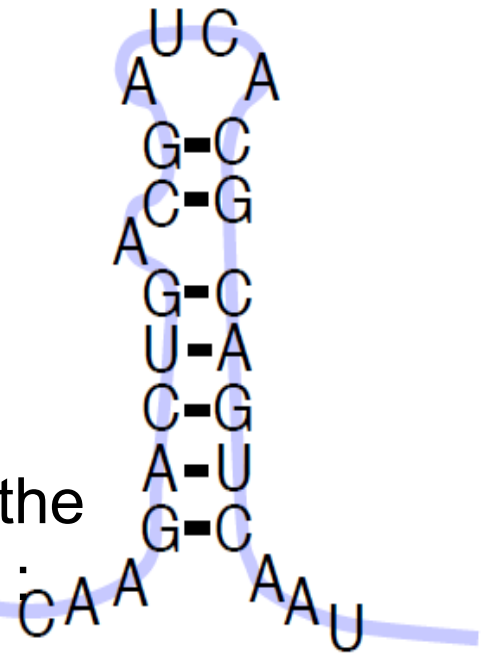


RNAspace.org

ncRNA background

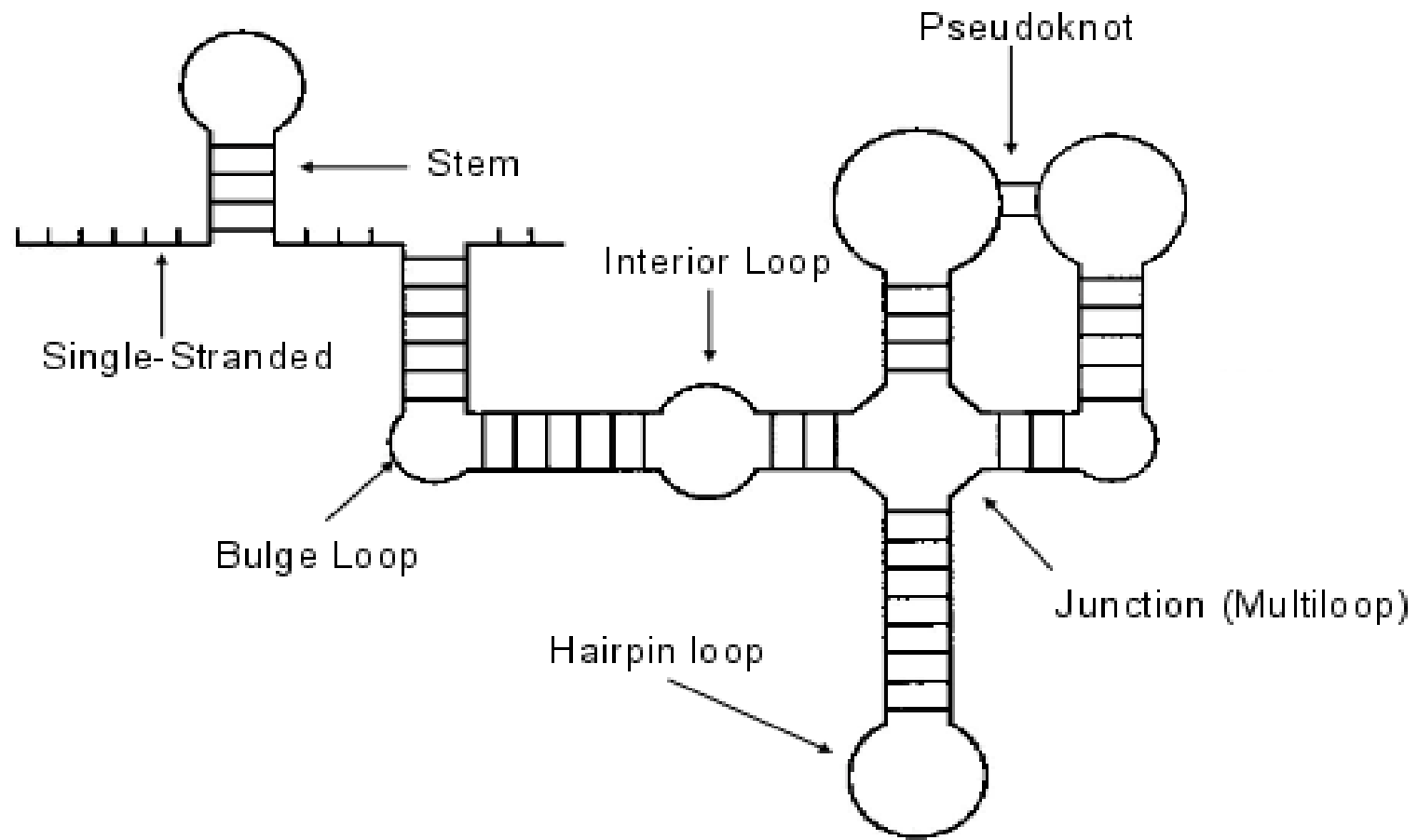
RNA background

- RNA folds on itself by base pairing :
 - A with U : A-U, U-A
 - C with G : G-C, C-G
 - Sometimes G with U : U-G, G-U
- Folding = Secondary structure
- Structure related to function : ncRNA of the same family have a conserved structure : the family signature
- Sequence less conserved



RNA background

Different elementary motifs



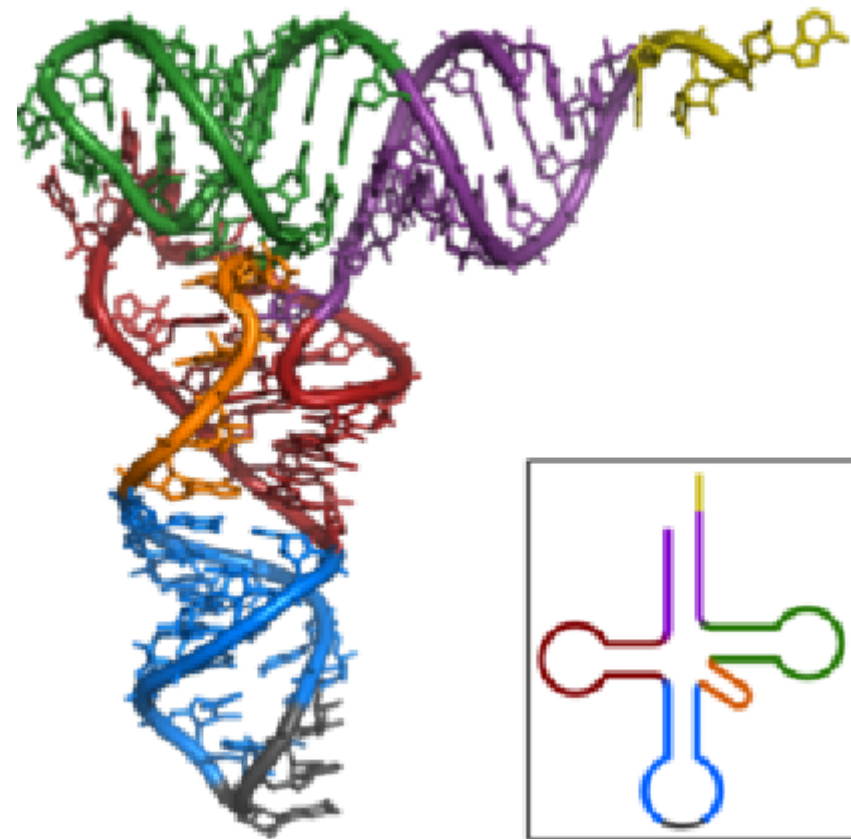
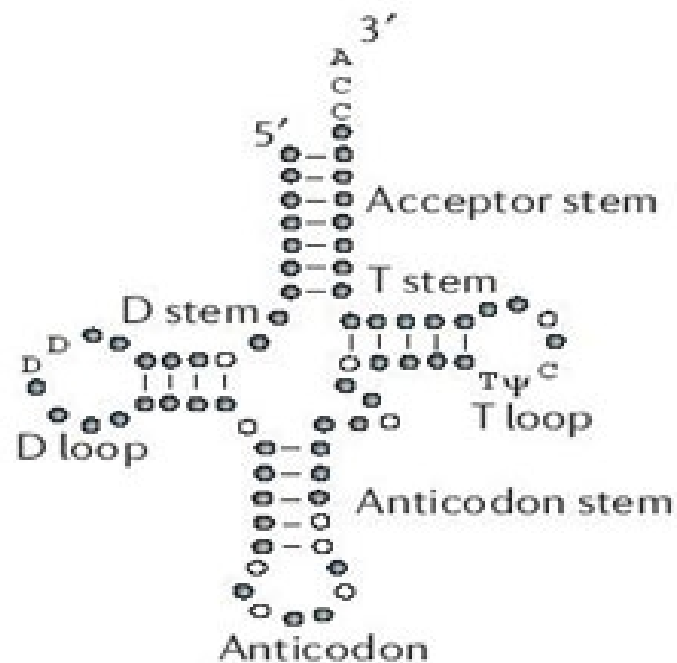
The non coding protein RNA world

A high diversity in size

- **Large non coding protein RNA**
 - >300 nt : rRNA, Xist, H19, ...
 - Genome structure & expression
- **Small non coding protein RNA**
 - >30 nt : tRNA, snoRNA, snRNA...
 - mRNA maturation, translation
- **Micro non coding protein RNA**
 - 18-30 nt : miRNA, hc-siRNA, ta-siRNA, nat-siRNA, piRNA...
 - PTGS, TGS, Genome stability, defense...

RNA background

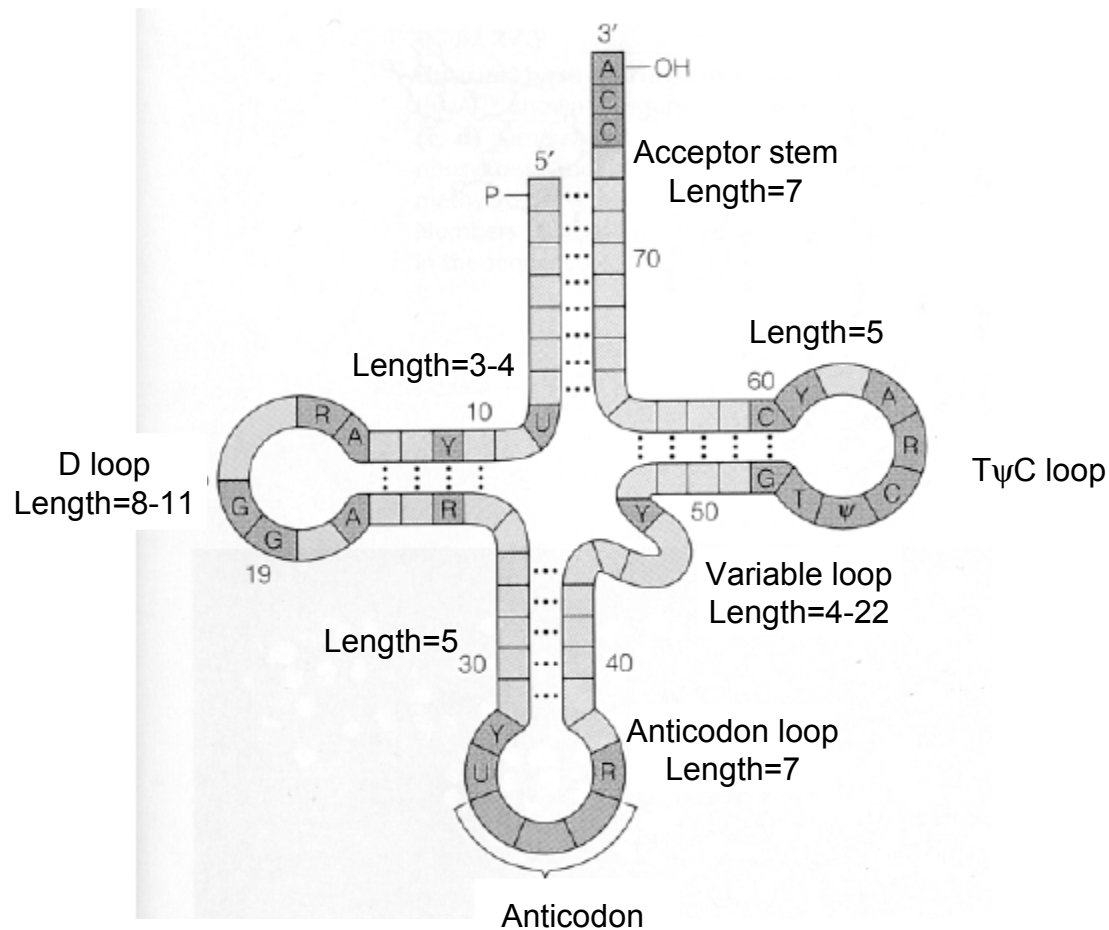
Example: the tRNA family



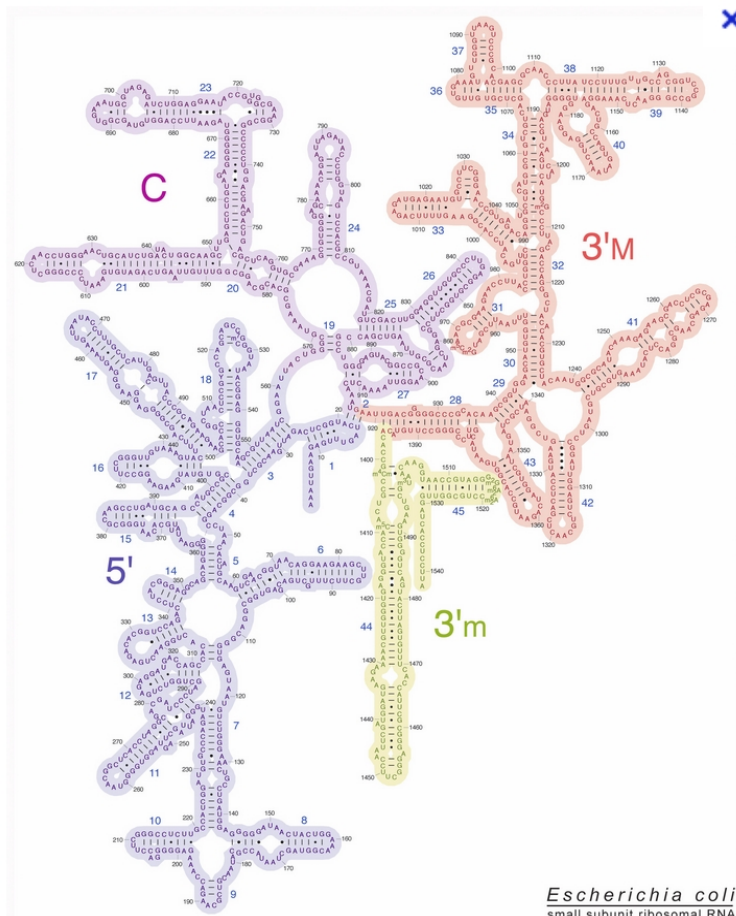
Copyright © 2006 Nature Publishing Group
Nature Reviews | Microbiology

RNA background

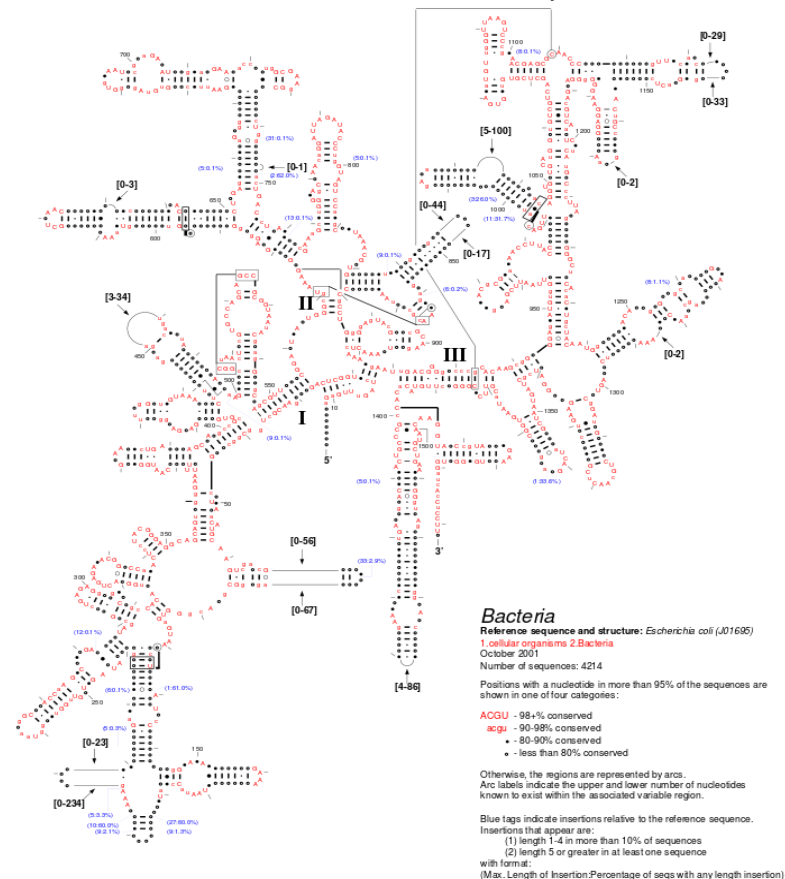
Example: the tRNA family



RNA background Example: 16S rRNA family

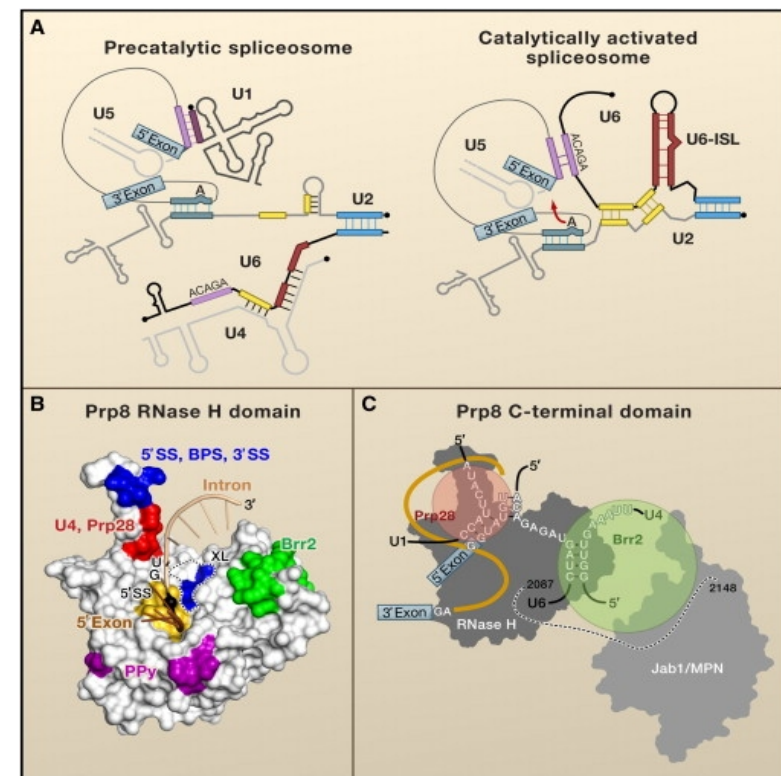
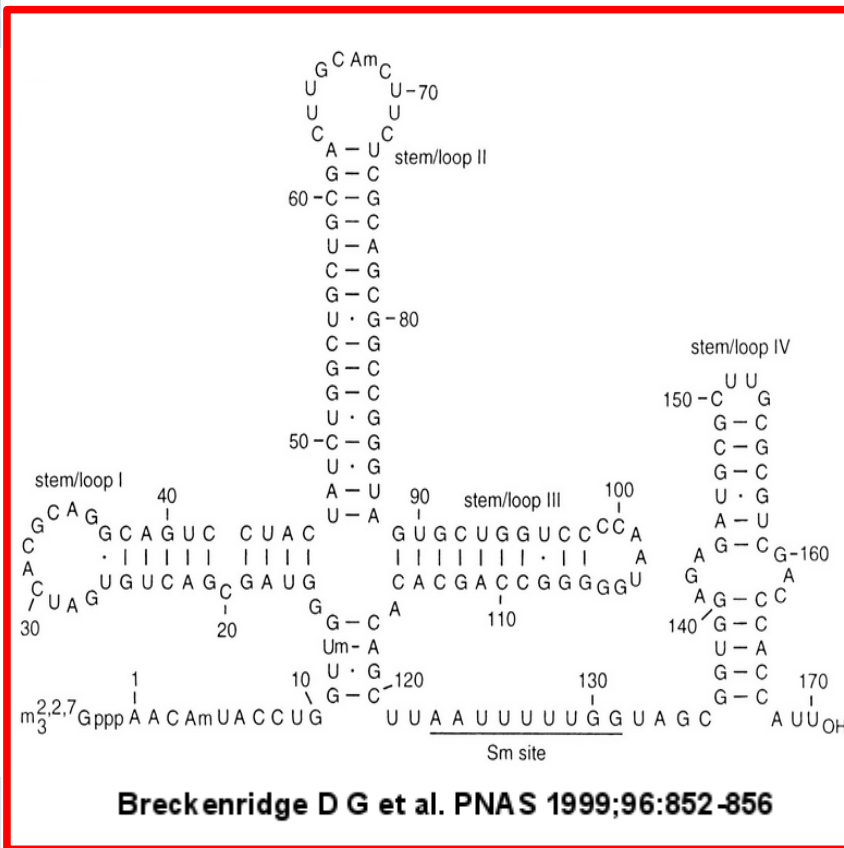


Phylogenetic conservation superimposed onto the
Escherichia coli Small Subunit rRNA secondary structure



RNA background

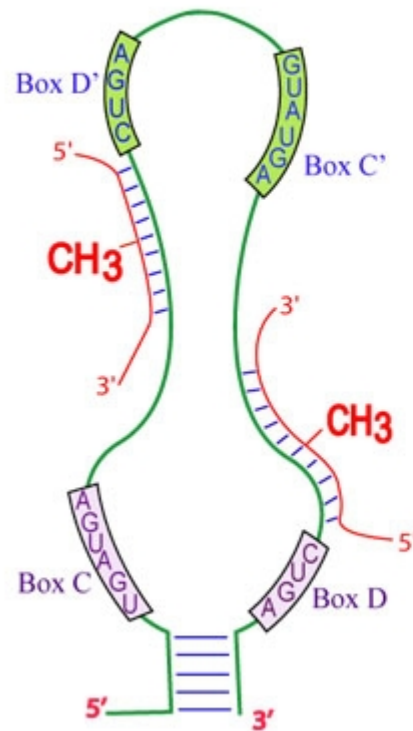
Example: snRNA family



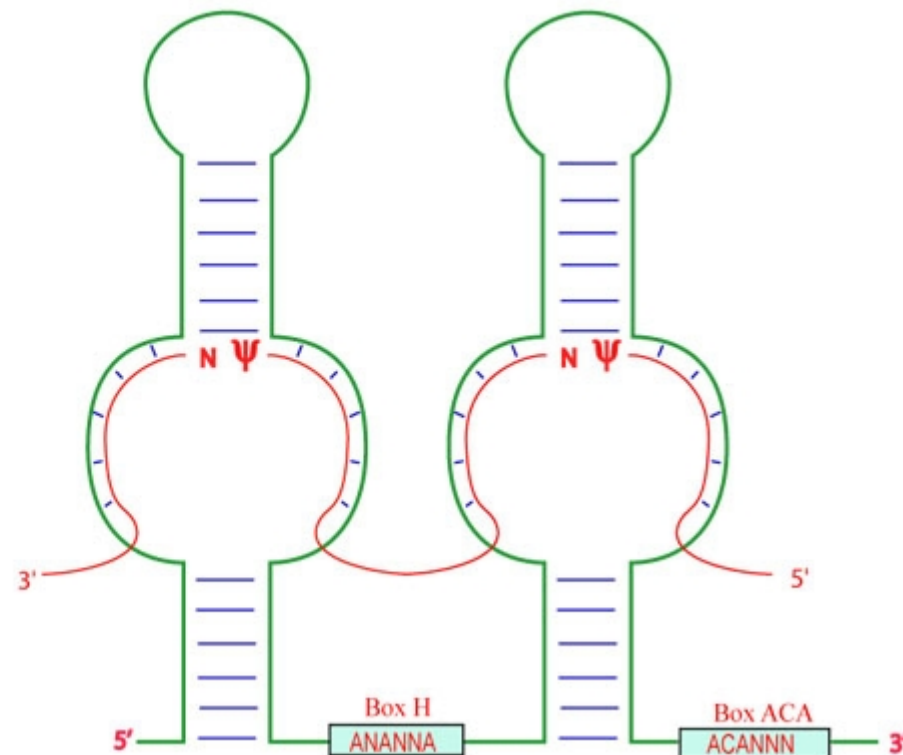
RNA background

Example: snoRNA families

Box C/D snoRNA



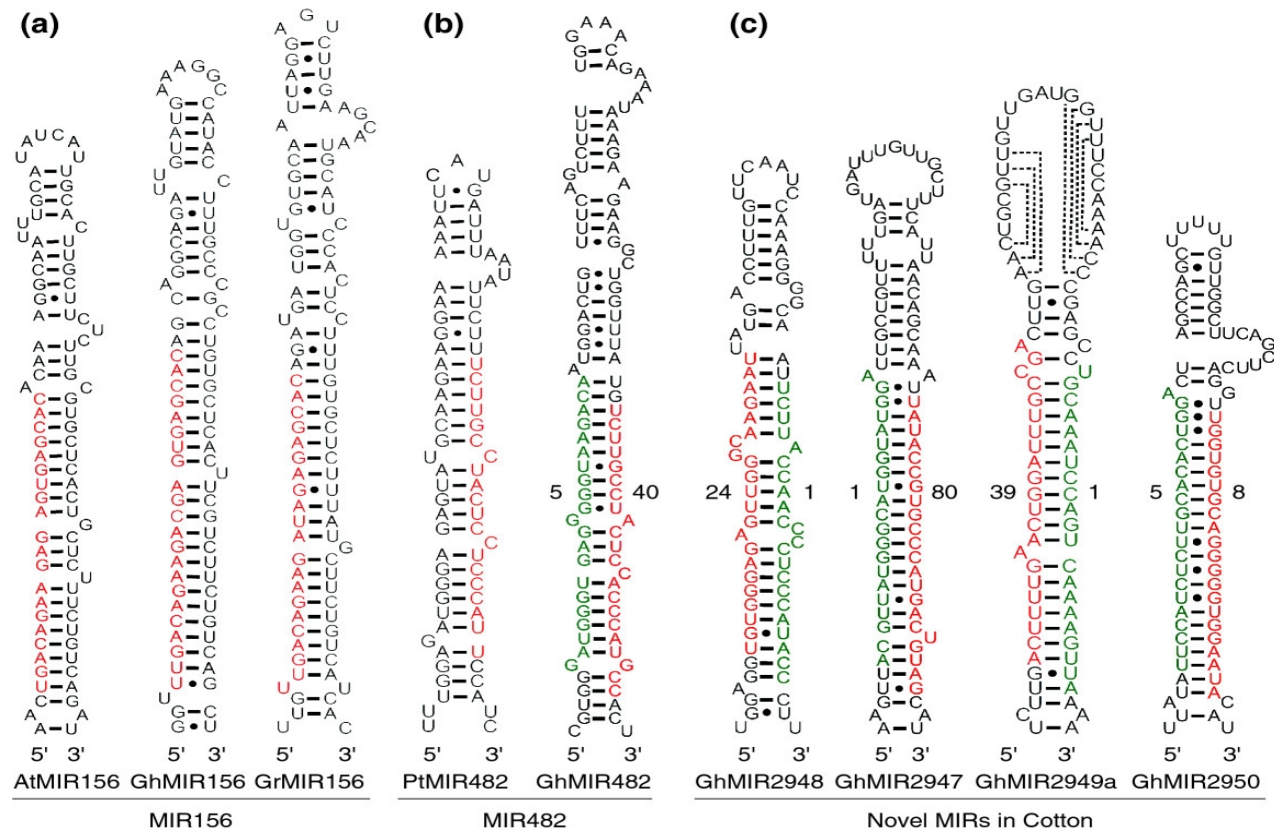
Box H/ACA snoRNA



<http://biochem.ncsu.edu/faculty/maxwell/snoRNA.jpg>

H. sapiens chr22

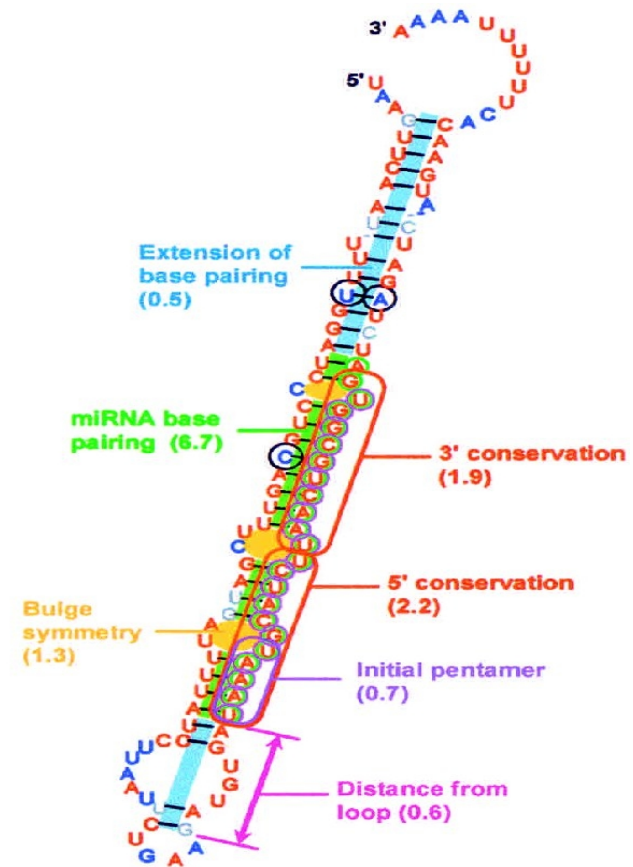
RNA background Example: miRNA family



RNA background

Example: miRNA family

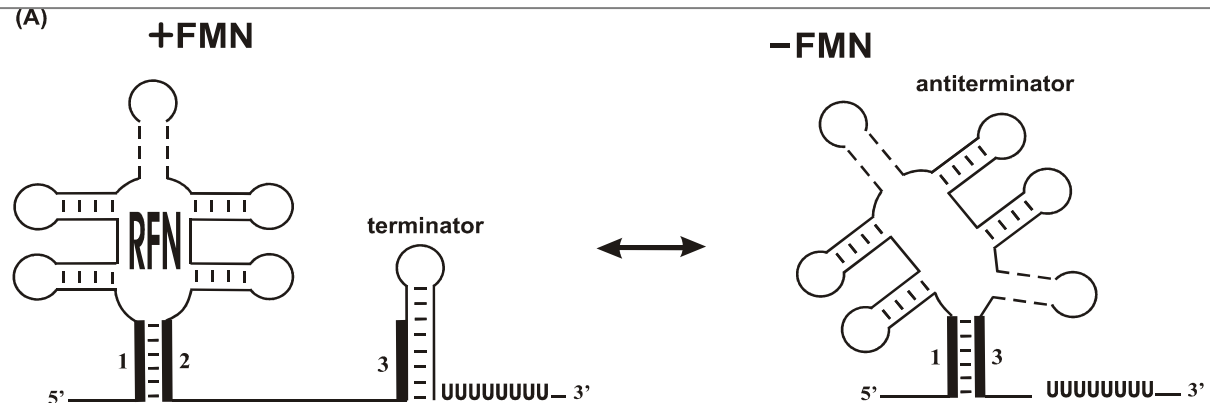
- MiRscan examines several features and computes a score
- The score is the sum of the evidence scores, computed independantly for each feature



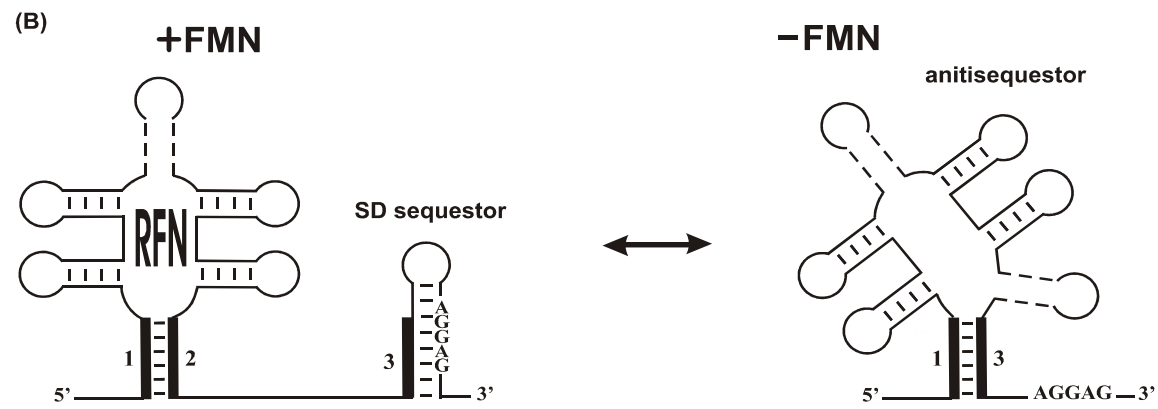
RNA background

Example: RFN riboswitch family

**Transcription
attenuation**



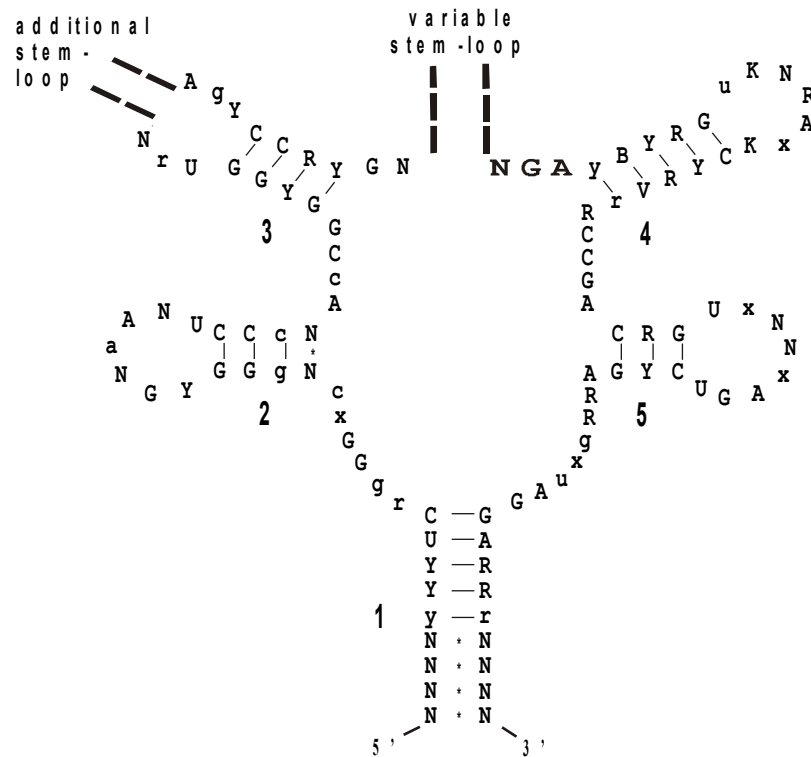
**Translation
attenuation**



From Gelfand, Paris, 2004

RNA background

Example: RFN riboswitch family



Capitals: invariant positions.

Lower case: strongly conserved positions.

Dashes and stars: obligatory and facultative base pairs

Degenerate positions:

R = A or G

Y = C or U

K = G or U

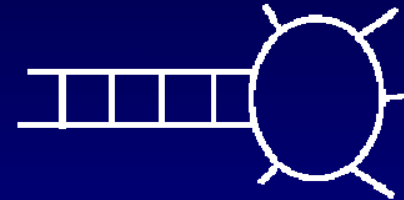
B = not A

V = not U

N: any nucleotide

X: any nucleotide or deletion

RNA background ncRNA and free energy

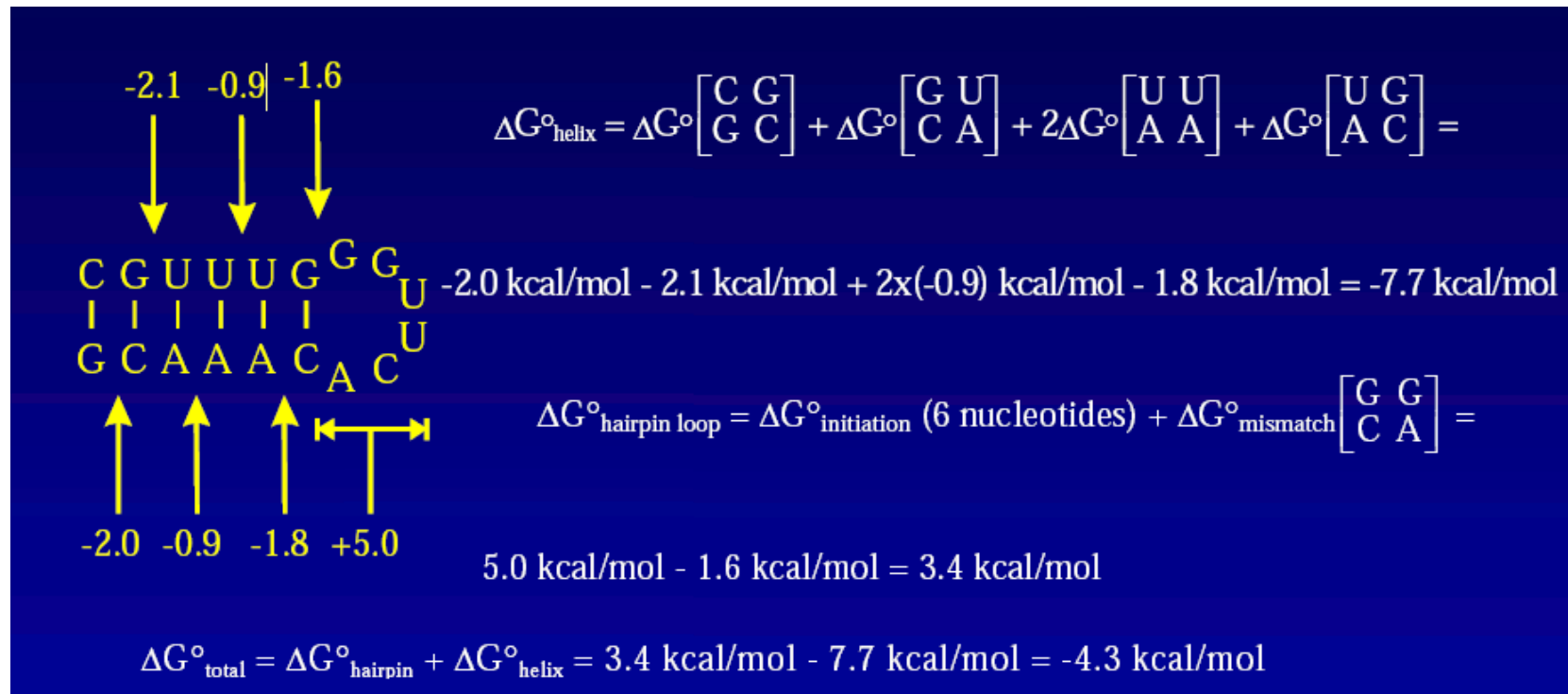


Unpaired State \rightleftharpoons Structure i

$$K_i = \frac{[\text{Structure } i]}{[\text{Unpaired State}]} = e^{-\Delta G_i^0/RT}$$

- K=equilibrium constant giving the ratio of concentrations for folded, S, and unfolded, U, species at equilibrium
- ΔG^0 = standard free energy difference between S and U
- R = gas constant
- T = temperature in kelvins

RNA background ncRNA and free energy



RNA background ncRNA and free energy

- **It is admitted**
 - The right secondary structure is that minimizing the free energy
 - 18^N possible secondary structures for a sequence of length N
 - For $N=100$: 3×10^{25} structures to compute
- **Efficient software to do that**
 - RNAfold (Hofacker, 2003)
<http://www.tbi.univie.ac.at/~ivo/RNA/>
 - Mfold (Zucker, Science, 1989)
<http://frontend.bioinfo.rpi.edu/zukerm/export/mfold-3.html>

ncRNA background

Where are ncRNA located ?

ncRNA background

Where are they in eucaryotes ?

- **Generally in non coding regions**
- **But also in :**
 - Inter-ORF
 - Introns
 - snoRNA, miRNA, tRNA
 - Coding regions
 - Anti-sens of (non) coding regions

ncRNA background

Where are they in bacteria and archea ?

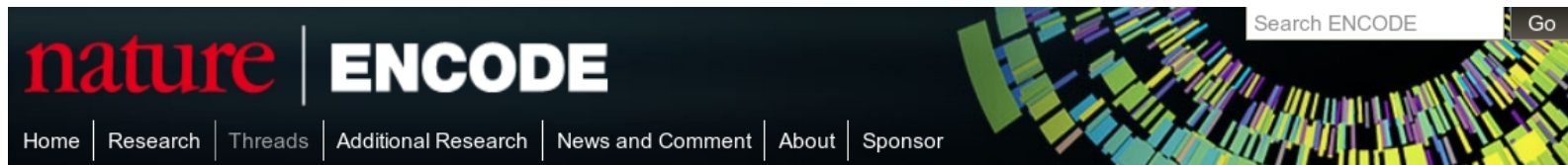
- **Generally in non coding regions**
 - All organisms, all families
- **But also in :**
 - Cis-reg : 5'UTR of mRNA
 - Antisens of (non) coding regions

ncRNA background

Where are they in Human ?

- **ncRNA in the ENCODE project**

<http://www.nature.com/encode/threads/non-coding-rna-characterization>



An advertisement for Illumina's MiSeq next-generation sequencing technology. It features the Illumina logo at the top right. The text reads 'MiSeq' in a large, stylized font, followed by 'Next-generation sequencing for all you seek.' Below this is a 'LEARN MORE' button with a play icon. At the bottom, there is an image of the MiSeq sequencing machine with orange lines representing data flow.

A thread is a way to follow a specific scientific theme that brings together relevant sections extracted from the co-published ENCODE papers

06

Non-coding RNA characterization

Many novel and previously known non-coding RNA species are characterized in ENCODE

nature

An Integrated Encyclopedia of DNA Elements in

In addition, we annotated 8,801 automatically derived small RNAs and 9,640 manually curated long non-coding RNA (lncRNA) loci¹⁷. Comparing lncRNAs to other ENCODE data indicates

ncRNA background

Where are they in Human ?

- **lncRNA**

chromatin marks have been identified for 13.9% (Derrien *et al.* 2012). These lncRNAs can be further reclassified into the following locus biotypes based on their location with respect to protein-coding genes:

1. **Antisense RNAs:** Locus that has at least one transcripts that intersect any exon of a protein-coding locus on the opposite strand, or published evidence of antisense regulation of a coding gene.
2. **LincRNA:** Locus is intergenic non-coding RNA loci.
3. **Sense overlapping:** Locus contains a coding gene within an intron on the same strand.
4. **Sense intronic:** Locus resides within intron of a coding gene, but does not intersect any exons on the same strand.
5. **Processed transcript:** Locus where non of its transcripts contain an open reading frame (ORF) and cannot be placed in any of the other categories because of complexity in their structure.

In summary the lncRNAs data set in GENCODE 7 consists of 5,058 lincRNA loci, 3,214 antisense loci, 378 sense intronic loci and 930 processed transcripts loci. Manually evaluating

ncRNA background

Where are they in eucaryotes ?

- miRNA

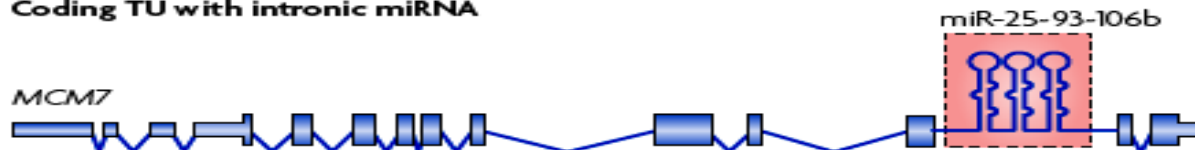
a Non-coding TU with intronic miRNA



b Non-coding TU with exonic miRNA



c Coding TU with intronic miRNA



d Coding TU with exonic miRNA



RNAspace.org

ncRNA annotation

Existing in International Repositories

Genomes and ncRNA annotation

- The reference : Genbank

<http://www.ncbi.nlm.nih.gov/genbank/genomesubmit/>

Annotation

While annotation is optional for incomplete WGS submissions, complete genome submissions must be annotated. You can annotate the genome yourself, following the instructions on these pages. Alternatively, you can request that your genome submission be annotated by NCBI's [Prokaryotic Genomes Annotation Pipeline](#) (PGAAP) that is available for genomes being submitted to GenBank. **New in Dec. 2012: Generate files for genome submission to GenBank and request PGAAP annotation in the Private Comments box during the submission of that genome. (There is no longer a separate PGAAP submission process.)**

If you annotate yourself, several features are the minimal required annotation, but there are many additional features that can be included. It is our hope that the annotation present on any genome will evolve over time as more is known about the biology. In reviewing bacterial genome annotation, NCBI strives to ensure that the annotation is consistent throughout the submission and when compared to other genome submissions. We also strive to present information that is an accurate representation of the known biology. To do this we need your help. Please pay careful attention to the annotation instructions presented here and please review all your annotation before submitting your genome. Many genomes are annotated by automatic prediction programs and since these programs do make mistakes, it is up to all of us to try and ensure the information being presented is as accurate as possible. A summary of the required annotation is presented below, however please also refer to our [detailed annotation instructions](#) for our annotation expectations.

Required Annotation

1. Genes
 - locus_tag
2. Coding regions of known proteins
 - product (protein) names
 - protein_id
3. structural RNAs (tRNAs and ribosomal RNAs)


structural RNAs

rRNA, tRNA, misc_RNA, and ncRNA are features used to annotate the various structural RNA genes. All RNA features must include a corresponding gene feature with a locus_tag qualifier. Only ribosomal RNAs (rRNA) and transfer RNAs (tRNA) are required.

Genomes and ncRNA annotation

- Another reference : Ensembl

<http://www.ensembl.org/info/docs/genebuild/index.html>


[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

[Login/Register](#)

[Help & Documentation](#) > [Gene Annotation](#)

The Ensembl Annotation Process

Genome assemblies

The [Genome Assemblies](#) page gives more information on where we get our genome assemblies from, how the sequence data for these genome assemblies are structured, and how we represent these data in Ensembl.

Protein-coding gene annotation

Protein-coding genes are automatically annotated using Ensembl's genebuild pipeline. All transcripts are based on mRNA and proteins in public scientific databases.

The human gene set is used as the [GENCODE](#) gene set. The human and mouse gene sets include all [CCDS](#) transcripts.

See the [annotation article](#) for more about the Ensembl genebuild pipeline, gene names and annotation.

[Low-coverage genomes](#) are annotated using a modified pipeline which attempts to locate genes across multiple scaffolds.

More genes

The Ensembl gene set also includes automatically-annotated pseudogenes and **non-coding RNAs**. For human and mouse, we include annotation from [IMGT](#) for [Ig genes](#).

EST-based genes are predicted and displayed on the website but are not included in the final gene set.

Paired-end Illumina [RNA-seq data](#) have been used to generate transcript models for many species including human, zebrafish and pig.


Alternative Splicing

Ensembl includes automatically-annotated [Alternative splicing events](#) for model organisms.

Ensembl release 71 - April 2013 © [WTSI](#) / [EBI](#)
[About Ensembl](#) | [Privacy Policy](#) | [Contact Us](#)

Genomes and ncRNA annotation

- **Another reference : Ensembl**
<http://www.ensembl.org/info/docs/genebuild/ncrna.html>


[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

[Login/Register](#)

[Help & Documentation](#) > [Gene Annotation](#) > [Annotation of Non-Coding RNAs](#)

Annotation of Non-Coding RNAs

Non-coding RNA Overview

Non-coding RNAs (ncRNAs) are involved in many biological processes and are increasingly seen as important. As is the case with proteins, it is the overall structure of the molecule which imparts function. However, while similar protein structures are often reflected in a conserved amino acid sequence, sequences underlying RNA secondary structure are very variable; this makes ncRNAs difficult to detect using sequence alone.

Because of this, we use a variety of techniques to detect ncRNAs. First, a combination of sensitive BLAST searches are used to identify likely targets, then a covariance model search is used to measure the probability that the targets can fold into the structures required. Other ncRNAs are added as part of the raw compute stage.

The following non-coding RNA gene types are annotated, along with pseudogenes

- tRNA**
transfer RNA
- Mt-tRNA**
transfer RNA located in the mitochondrial genome
- rRNA**
ribosomal RNA
- scRNA**
small cytoplasmic RNA
- snRNA**
small nuclear RNA
- snoRNA**
small nucleolar RNA
- miRNA**
microRNA precursors
- misc_RNA**
miscellaneous other RNA
- lincRNA**
Long intergenic non-coding RNAs

Genomes and ncRNA annotation

- **Another reference : Ensembl**
<http://www.ensembl.org/info/docs/genebuild/ncrna.html>

Annotation Details

Most **ncRNAs** are annotated by aligning genomic sequence against [RFAM](#) using [BLASTN](#). The BLAST hits are clustered and filtered by E value and are used to seed Infernal searches of the locus with the corresponding RFAM covariance models. The purpose of this is to reduce the search space required, as to scan the entire genome with all the RFAM covariance models would be extremely CPU-intensive. The resulting BLAST hits are then used as supporting evidence for ncRNA genes.

miRNAs are predicted by BLASTN of genomic sequence slices against [miRBase](#) sequences. All species are used. The BLAST hits are clustered and filtered by E value and the aligned genomic sequence is then checked for possible secondary structure using RNAFold. If evidence is found that the genomic sequence could form a stable hairpin structure, the locus is used to create a miRNA gene model. The resulting BLAST hit is used as supporting evidence for the miRNA gene.

Note: The miRNA identifier and name are only associated to the resulting Ensembl miRNA if they are of the same species.

tRNAs are annotated as part of the raw compute process using [tRNAscan-SE](#).

lincRNA (Long intergenic non-coding RNAs) Ensembl gene annotation, cDNA alignments and chromatin-state map data from the Ensembl regulatory build are used to predict lincRNAs for human and mouse. We do not import the lincRNAs identified by Guttman et al [1], but their publication guided us to our current approach for automatically annotating lincRNAs. First, regions of chromatin methylation (H3K4me3 and H3K36me3) outside known protein-coding loci are identified. Next, cDNAs which overlap with H3K4me3 or H3K36me3 features are identified as candidate lincRNAs. A final evaluation step investigates if each candidate lincRNA has any protein-coding potential. Any candidate lincRNA containing a substantial open reading frame (ORF) covering 35% or more of its length and containing PFAM/tigfam protein domains will be rejected. Candidate lincRNAs that pass the final evaluation step are included in the human or mouse gene set as lincRNA genes.

Genomes and ncRNA annotation

- Another reference : Ensembl
<http://www.ensembl.org/info/data/ftp/index.html>

Single species data

Popular species are listed first. You can customise this list via our [home page](#).

Show 10 entries		Show/hide columns															Filter	
★	Species	DNA (FASTA)	cDNA (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (EMF)	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM		
Y	Human <i>Homo sapiens</i>	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	EMF	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM		
Y	Mouse <i>Mus musculus</i>	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	EMF	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	-		
Y	Zebrafish <i>Danio rerio</i>	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	GVF	VCF	VEP	-	-	BAM		
	Alpaca <i>Vicugna pacos</i>	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-	-		
	Anole lizard <i>Anolis carolinensis</i>	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-	BAM		
	Armadillo <i>Dasypus novemcinctus</i>	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-	-		
	Bushbaby <i>Oryzomys flavescens</i>	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	-	-	-	-	-		

ncRNA prediction

Versus

Coding RNA prediction

ncRNA prediction and annotation

- **Not predicted by gene prediction tools**
 - No specific signal (start, stop, splicing sites...)
 - Multiple location (intergenic, intronic, coding, antisens)
 - Variable size
 - No strong sequence conservation in general
- **A variety of existing approaches not always easy to integrate**
 - Known family: Homology prediction
 - New family: *De novo* prediction

The non coding protein RNA world

- **Protein Approaches**
 - Statistically biased (codon triplets)
 - Open Reading Frames
- **ncRNA Approaches**
 - High CG content (hyperthermophiles archaea)
 - Orphan promoter/Terminator identification (bacteria)

The non coding protein RNA world

Comparative analysis: similarity searching

- **Proteins**

- BLAST, Sequence Alignment (DP)
- Genes that code for proteins are conserved across genomes (e.g. low rate of mutation)

- **ncRNA**

- Low sequence conservation
- Secondary structure usually conserved
- Alignment scoring based on structure can be imperative

The non coding protein RNA world

Comparative analysis: similarity searching

A U
 G A
 C<-> G-C<->G
 U-A
 A-U
 G-C
 A<->C-G<->U
 S1

S2	AGAUCGAAAGAUCU
	* * * *
S1	CGAUGGAUACAUCG
	* *
S3	CCAUGGAUAGUUCG

A U
 G A
 G*G
 U*U
 A-U
 C*C
 C-G
 S3

RNAspace.org

ncRNA annotation

ncRNA databases

The non coding protein RNA world Databases

- **Generalist databases**
 - No organism specificity
 - No family specificity
- **Specific databases**
 - Groups of organisms : Plants, Animals, Human...
 - ncRNA families: rRNA, tRNA, miRNA, snRNA, snoRNA, tmRNA...
 - Both

The non coding protein RNA world

Generalist databases

- RFAM

<http://www.sanger.ac.uk/Software/Rfam/>

- NonCode

<http://noncode.bioinfo.org.cn/index4.htm>

- RNAdb

<http://jsm-research.imb.uq.edu.au/rnadb/>

- fRNAdb

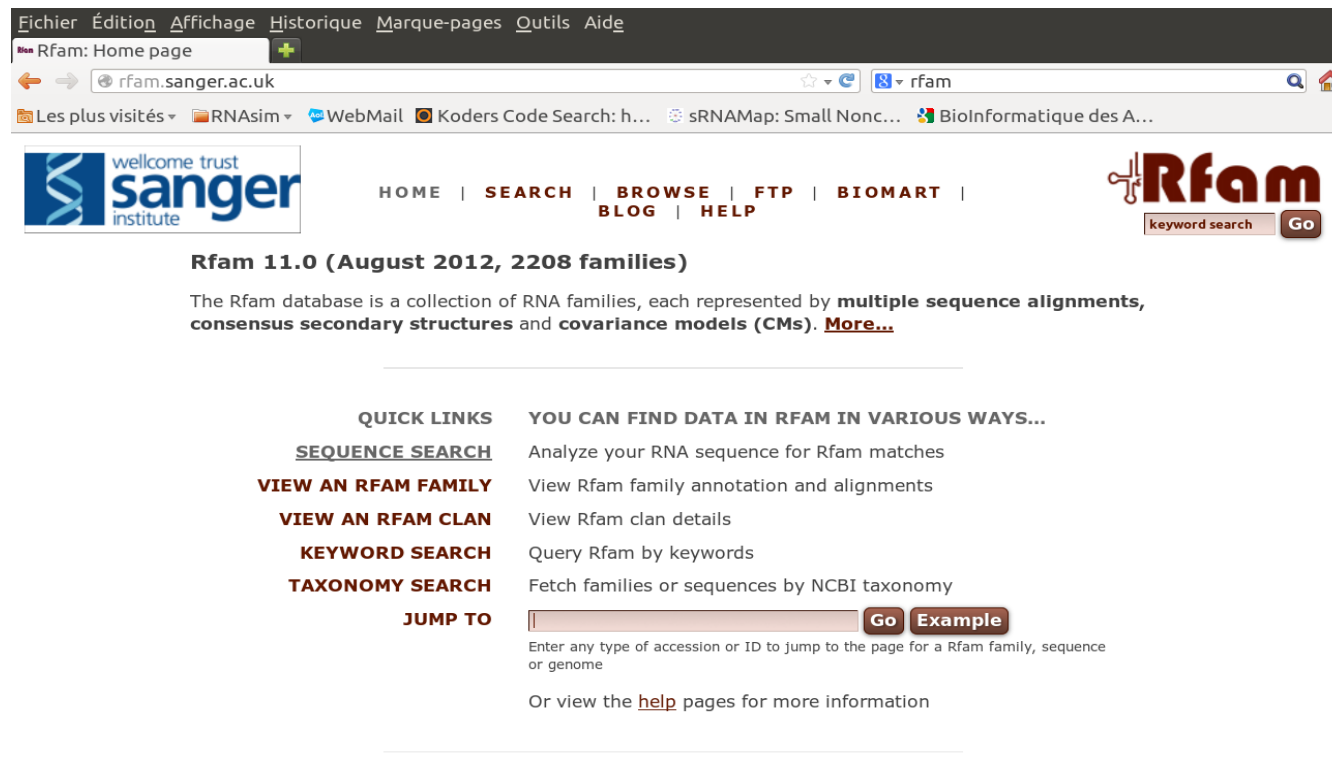
<http://www.ncrna.org/frnadb/>

- ncRNA

<http://biobases.ibch.poznan.pl/ncRNA/>

The non coding protein RNA world Generalist databases

- RFAM



The screenshot shows the Rfam database homepage as it appeared in August 2012. The browser window displays the URL `rfam.sanger.ac.uk`. The page features the Wellcome Trust Sanger Institute logo on the left and the Rfam logo on the right, which includes a keyword search bar. A navigation menu in the center lists links for HOME, SEARCH, BROWSE, FTP, BIOMART, BLOG, and HELP. Below the navigation menu, the text "Rfam 11.0 (August 2012, 2208 families)" is displayed, followed by a description of the database as a collection of RNA families represented by multiple sequence alignments, consensus secondary structures, and covariance models (CMs). A "More..." link is provided for further information. The page is divided into two main sections: "QUICK LINKS" on the left and "YOU CAN FIND DATA IN RFAM IN VARIOUS WAYS..." on the right. The "QUICK LINKS" section includes links for SEQUENCE SEARCH, VIEW AN RFAM FAMILY, VIEW AN RFAM CLAN, KEYWORD SEARCH, TAXONOMY SEARCH, and JUMP TO. The "YOU CAN FIND DATA IN RFAM IN VARIOUS WAYS..." section provides instructions on how to analyze RNA sequences, view family annotations, query by keywords, and fetch families by NCBI taxonomy. A search bar with "Go" and "Example" buttons is also present.

Eichier Édition Affichage Historique Marque-pages Outils Aide

Rfam: Home page

rfam.sanger.ac.uk

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc... Bioinformatique des A...

wellcome trust
sanger
institute

HOME | SEARCH | BROWSE | FTP | BIOMART |
BLOG | HELP

Rfam
keyword search Go

Rfam 11.0 (August 2012, 2208 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](#)

VIEW AN RFAM FAMILY

VIEW AN RFAM CLAN

KEYWORD SEARCH

TAXONOMY SEARCH

JUMP TO

YOU CAN FIND DATA IN RFAM IN VARIOUS WAYS...

Analyze your RNA sequence for Rfam matches

View Rfam family annotation and alignments

View Rfam clan details

Query Rfam by keywords

Fetch families or sequences by NCBI taxonomy

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

Go Example

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

The non coding protein RNA world Generalist databases

- RFAM

Fichier Édition Affichage Historique Marque-pages Outils Aide

Rfam: Browse Rfam fa...

rfam.sanger.ac.uk/Families#M

Les plus visités RNAsim WebMail Kodors Code Search: h... sRNAMap: Small Nonc... BioInformatique des A...

HOME | SEARCH | BROWSE | FTP | BIOMART | BLOG | HELP keyword search

Browse all 2208 Rfam families

The table may be sorted by clicking on the column titles, or restored to the original order [here](#). **Please note** that sorting large tables can be slow. Go [back](#) to the browse index.

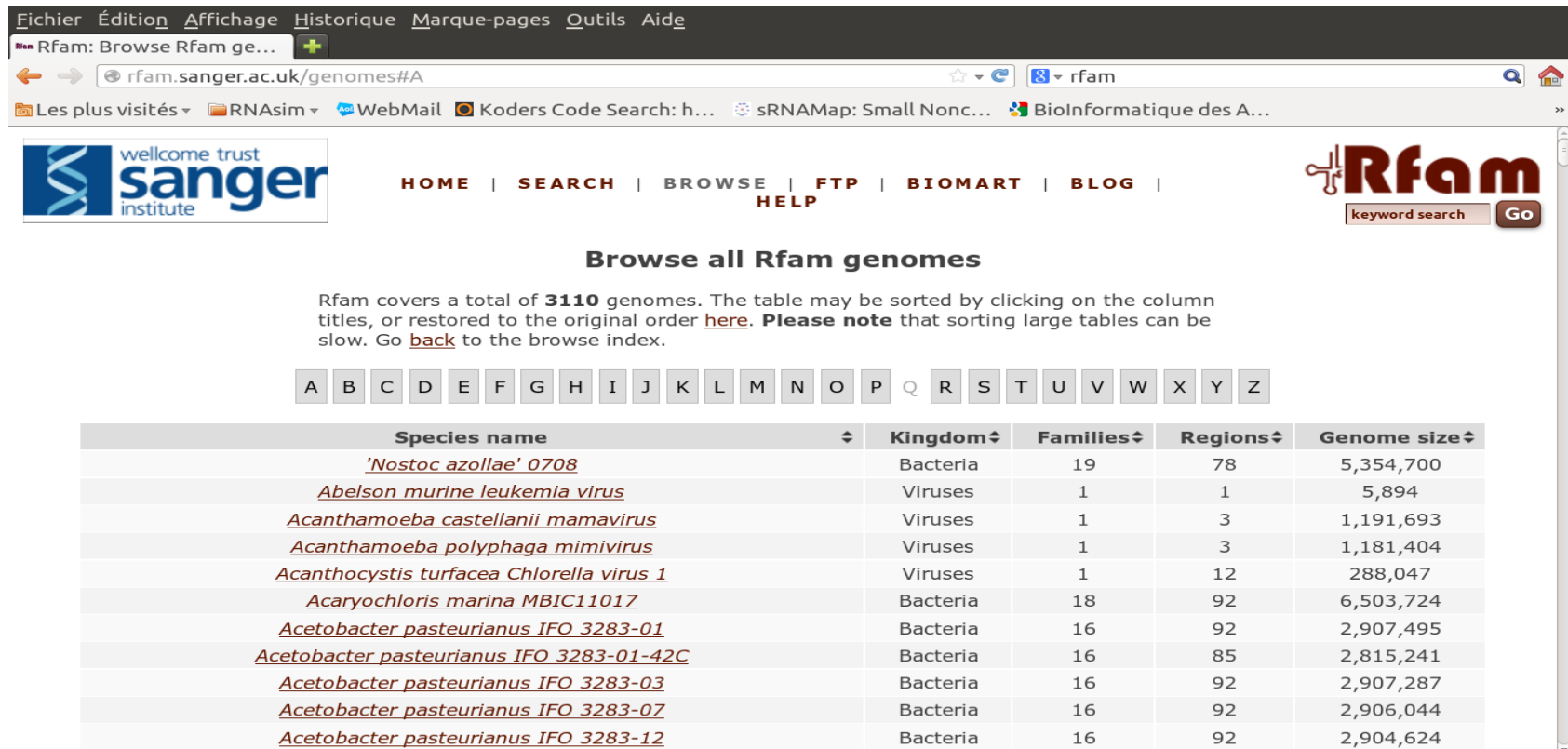
0 - 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

ID ↕	Accession↕	Type ↕	Seed↕	Full↕	Average length ↕	Sequence identity (%) ↕	Description ↕
23S-methyl	RF01065	Cis-reg;	19	590	100.70	59.00	23S methyl RNA motif
5S_rRNA	RF00001	Gene; rRNA;	712	229,497	116.60	60.00	5S ribosomal RNA
5_8S_rRNA	RF00002	Gene; rRNA;	61	375,612	152.20	69.00	5.8S ribosomal RNA
6C	RF01066	Cis-reg;	20	150	75.50	73.00	6C RNA
6S	RF00013	Gene;	153	3,521	180.30	45.00	6S / SsrS RNA
6S-Flavo	RF01685	Gene; sRNA;	89	131	108.40	68.00	6S-Flavo RNA
7SK	RF00100	Gene;	45	21,885	322.20	83.00	7SK RNA
ACA59	RF01293	Gene; snRNA; snoRNA; HACA-box;	3	47	154.30	73.00	Small nucleolar RNA ACA59
ACA64	RF01225	Gene; snRNA; snoRNA; HACA-box;	30	334	127.60	76.00	Small nucleolar RNA ACA64

The non coding protein RNA world

Generalist databases

- RFAM



Fichier Édition Affichage Historique Marque-pages Outils Aide

Non Rfam: Browse Rfam ge... +

rfam.sanger.ac.uk/genomes#A

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAmap: Small Nonc... BioInformatique des A...

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | BIOMART | BLOG | HELP

Rfam

keyword search Go

Browse all Rfam genomes

Rfam covers a total of **3110** genomes. The table may be sorted by clicking on the column titles, or restored to the original order [here](#). **Please note** that sorting large tables can be slow. Go [back](#) to the browse index.



A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Species name	Kingdom	Families	Regions	Genome size
'Nostoc azollae' 0708	Bacteria	19	78	5,354,700
Abelson murine leukemia virus	Viruses	1	1	5,894
Acanthamoeba castellanii mamavirus	Viruses	1	3	1,191,693
Acanthamoeba polyphaga mimivirus	Viruses	1	3	1,181,404
Acanthocystis turfacea Chlorella virus 1	Viruses	1	12	288,047
Acaryochloris marina MBIC11017	Bacteria	18	92	6,503,724
Acetobacter pasteurianus IFO 3283-01	Bacteria	16	92	2,907,495
Acetobacter pasteurianus IFO 3283-01-42C	Bacteria	16	85	2,815,241
Acetobacter pasteurianus IFO 3283-03	Bacteria	16	92	2,907,287
Acetobacter pasteurianus IFO 3283-07	Bacteria	16	92	2,906,044
Acetobacter pasteurianus IFO 3283-12	Bacteria	16	92	2,904,624

The non coding protein RNA world Generalist databases

- RFAM

[Fichier](#) [Édition](#) [Affichage](#) [Historique](#) [Marque-pages](#) [Outils](#) [Aide](#)
 Rfam: Genome: Triticu...
 rfam.sanger.ac.uk/genome/4565#tabview=tab1
 Les plus visités | RNAsim | WebMail | Koders Code Search: h... | sRNAMap: Small Nonc... | BioInformatique des A...


[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [BIOMART](#) | [BLOG](#) | [HELP](#)

[keyword search](#) [Go](#)

Genome: *Triticum aestivum* (bread wheat)
 NCBI taxonomy ID: 4565

1 sequence 1 species 0 structures

Summary
Chromosomes
Jump to...
 enter ID/acc [Go](#)

Chromosomes
 This section shows a chromosome-by-chromosome breakdown of the genome. This genome has [EMBL accessions](#). Each row shows the details of a particular chromosome. You can see information about the families found on that chromosome by clicking the "Show" button at the end of the row.

EMBL accession	Description	Length	Family count	Download GFF file	Show all regions
FN645450.1	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0011b	1,266,078	30	Download	Show

The non coding protein RNA world Generalist databases

- RFAM

Fichier Édition Affichage Historique Marque-pages Outils Aide

Rfam: Genome: Triticu...

rfam.sanger.ac.uk/genome/4565#tabview=tab1

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc... BioInformatique des A...

keyword search **Go**

Genome: *Triticum aestivum* (bread wheat)
NCBI taxonomy ID: 4565

1 sequence 1 species 0 structures

Summary

Chromosomes

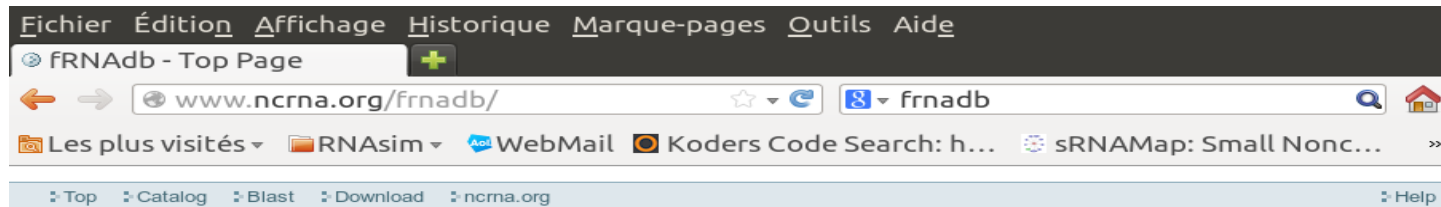
Jump to...
enter ID/acc **Go**

This section shows a chromosome-by-chromosome breakdown of the genome. This genome has [EMBL accessions](#). Each row shows the details of a particular chromosome. You can see information about the families found on that chromosome by clicking the "Show" button at the end of the row.

EMBL accession	Description	Length	Family count	Download GFF file	Show all regions
FN645450.1	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0011b	1,266,078	30	Download	Hide
Family	Start	End	Bits score		
MIR1122	15,164	15,256	81.00		
SS_rRNA	581,112	581,186	25.44		
SS_rRNA	581,221	581,312	28.57		
tRNA	657,418	657,344	58.30		
tRNA	658,128	658,025	25.55		
tRNA	661,702	661,629	59.90		
tRNA	674,444	674,530	48.09		
Intron_gpII	676,664	676,788	45.26		
Intron_gpII	677,993	678,114	32.34		
Intron_gpII	679,446	679,578	41.06		
Intron_gpII	681,388	681,521	46.77		
Intron_gpII	683,873	683,946	52.03		
Intron_gpII	686,204	686,282	54.78		
tRNA	694,524	694,595	58.85		
tRNA	695,300	695,227	63.61		
tRNA	696,004	696,077	62.32		
tRNA	742,744	742,673	46.43		
tRNA	804,767	804,696	46.65		
Plant_U3	898,830	898,694	98.41		
tRNA	907,724	907,653	39.28		
tRNA	950,362	950,433	52.30		
MIR1122	985,184	985,308	80.97		
MIR1122	1,043,280	1,043,166	91.01		
tRNA	1,096,147	1,096,076	52.30		
tRNA	1,101,627	1,101,556	52.30		
SS_rRNA	1,105,304	1,105,390	26.79		
SS_rRNA	1,105,430	1,105,521	28.57		
MIR1122	1,196,181	1,196,315	85.81		
MIR1122	1,220,853	1,220,719	83.62		
IRES_L-myc	1,256,838	1,256,680	39.66		

The non coding protein RNA world Generalist databases

- fRNAdb



A comprehensive non-coding RNA sequence database ver. 3.4

fRNAdb is [Web Service \(SOAP, REST\)](#) Ready.

Total: 510,055 entries



Catalog



Blast



Download



Help

Please input some query keywords for retrieving RNAs via simple text search e.g. "miRNA" or "snoRNA".

Try add disease/tissue name to make your search more specific e.g. "miRNA oncogene".

Or can be more specific e.g. "miRNA oncogene human".

The icons used in this page are a part of Tango Desktop Project [created by Tango Desktop Project](#) which are available under CC BY-SA License [\[2\]](#).

This site uses Yahoo! User Interface Library (YUI) [which is available under BSD License](#) [\[3\]](#).



fRNAdb is licensed under a Creative Commons [表示-非営利-改変禁止 2.1 日本 License](#).
Based on a work at [www.ncrna.org](#).

The non coding protein RNA world Generalist databases

- fRNAdb

Fichier Édition Affichage Historique Marque-pages Outils Aide

FRNAdb - Catalog (Data... +

www.ncrna.org/frnadb/catalog_datasou ☆ f n r n a d b

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc...

Top Catalog Blast Download ncrna.org Help

fRNAdb::Catalog::Datasource

- Length
- Datasource
- Sequence Ontology
- Taxonomy
- MicroArray
- Transcription Factor Binding Sites
- QMIM
- Disease
- Anatomy
- Biological Process
- Molecular Function
- Cellular Component
- RNA Name

Datasource ?

Datasource	fRNAdb Ids
RNAdb	289,317
Rfam	108,043
Gene Expression Omnibus	102,130
miRBase	5,912
NONCODE	5,511
H-invitational	2,149
snoRNA-LBME-db	359
European ribosomal RNA	44

The icons used in this page are a part of [Tango Desktop Project](#) created by Tango Desktop Project which are available under [CC BY-SA License](#).
This site uses [Yahoo! User Interface Library \(YUI\)](#) which is available under [BSD License](#).

Fichier Édition Affichage Historique Marque-pages Outils Aide

FRNAdb - Catalog (Leng... +

www.ncrna.org/frnadb/catalog_length ☆ f n r n a d b

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc...

Top Catalog Blast Download ncrna.org Help

fRNAdb::Catalog::Length

- Length
- Datasource
- Sequence Ontology
- Taxonomy
- MicroArray
- Transcription Factor Binding Sites
- QMIM
- Disease
- Anatomy
- Biological Process
- Molecular Function
- Cellular Component
- RNA Name

Length ?

Length	Range	fRNAdb Ids
Long	10000- nt	22
	1000-9999 nt	28,509
Medium	500-999 nt	6,615
	200-499 nt	32,715
Small	100-199 nt	69,502
	24-99 nt	264,228
Micro	-23 nt	108,464

The icons used in this page are a part of [Tango Desktop Project](#) created by Tango Desktop Project which are available under [CC BY-SA License](#).
This site uses [Yahoo! User Interface Library \(YUI\)](#) which is available under [BSD License](#).

The non coding protein RNA world

Specific databases

- miRBase

The screenshot shows the miRBase website with a dark blue header and a light blue sidebar. The main content area is white. The header includes a navigation menu with links: Fichier, Édition, Affichage, Historique, Marque-pages, Outils, Aide. Below the header is a search bar with the text 'miRBase' and a plus sign. The browser address bar shows 'www.mirbase.org'. The main content area features a large 'miRBase' logo and a search bar. The sidebar contains a 'Latest miRBase blog posts' section with two entries: 'Website at risk, Tues 19th March 8am-9am GMT' and 'miRBase web site down time, Oct 22nd-23rd'. The main content area also has a 'miRNA count: 21264 entries' section, a 'Search by miRNA name or keyword' section, and a 'Download published miRNA data' section. The footer includes a 'References' section.

miRBase: the microRNA database

miRBase provides the following services:

- The [miRBase database](#) is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature 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](#).

References

The non coding protein RNA world Specific databases

- miRBase

Fichier Édition Affichage Historique Marque-pages Outils Aide

Triticum aestivum miR... +

www.mirbase.org/cgi-bin/mirna_summa miRBase

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc...

miRBase MANCHESTER 1824

Home Search Browse Help Download Blog Submit

Triticum aestivum miRNAs (42 sequences)

ID	Accession	Chromosome	Start	End	Strand	Fetch
tae-MIR156	MI0016450					<input type="checkbox"/>
tae-MIR159a	MI0006170	CA731881	258	434	+	<input type="checkbox"/>
tae-MIR159b	MI0006171	CA484819	231	485	+	<input type="checkbox"/>
tae-MIR160	MI0006172	CJ641547	388	535	+	<input type="checkbox"/>
tae-MIR164	MI0006173	CA704421	8	163	+	<input type="checkbox"/>
tae-MIR167a	MI0006174	CK209908	358	465	+	<input type="checkbox"/>
tae-MIR167b	MI0016456	CK209889	362	451	+	<input type="checkbox"/>
tae-MIR171a	MI0006175	CD910903	77	206	+	<input type="checkbox"/>
tae-MIR171b	MI0016468	BJ275219	329	466	-	<input type="checkbox"/>
tae-MIR319	MI0016453	CA483944	88	298	-	<input type="checkbox"/>
tae-MIR395a	MI0016463	CV763592	90	166	+	<input type="checkbox"/>
tae-MIR395b	MI0016464					<input type="checkbox"/>
tae-MIR398	MI0016466	TA109388_4565	60	179	+	<input type="checkbox"/>
tae-MIR399	MI0006176	TA93688_4565	77	202	+	<input type="checkbox"/>
tae-MIR408	MI0006177	BE419354	48	234	+	<input type="checkbox"/>
tae-MIR444a	MI0006178	TA93849_4565	61	486	+	<input type="checkbox"/>

The non coding protein RNA world Specific databases

- **Silva**

The screenshot shows the SILVA website interface. At the top is a navigation bar with links: [Fichier](#), [Édition](#), [Affichage](#), [Historique](#), [Marque-pages](#), [Outils](#), [Aide](#). Below this is a search bar with the text "Silva" and a plus icon. The address bar shows "www.arb-silva.de" and the search bar contains "silva database". Below the address bar are several icons and links: "Les plus visités", "RNAsim", "WebMail", "Koders Code Search: h...", and "sRNAMap: Small Nonc...".

The main content area is divided into two columns. The left column contains the SILVA logo and a navigation menu: [Home](#), [Browser](#), [Search](#), [Aligner](#), [Download](#), [Documentation](#), [Projects](#), [FISH & Probes](#), [Shop](#), [Contact](#).

The right column contains a "News" section with several updates:

- 03.03.2013**
Meet ARB & SILVA at VAAM 2013
Talk to the ARB and SILVA developers at VAAM 2013 (10.03-13.03) in Bremen, Germany. Follow the link to see the sessions where you will find us.
- 18.02.2013**
LTP 111 released
Version 111 of the "All Species Living Tree" (LTP) has been released. Check the project website (link above) for more information ...
- 09.02.2013**
ARB & SILVA at Biosystematics 2013
Join us at the Software Bazar at Biosystematics in Vienna, February 18-22, 2013.
- 09.02.2013**
SILVA Stickers have arrived!
If you like one - contact us. We are happy to ship them to you.
[go to Archive ->](#)

Below the news section is a table titled "SILVA 111 - full release" showing statistics for different releases:

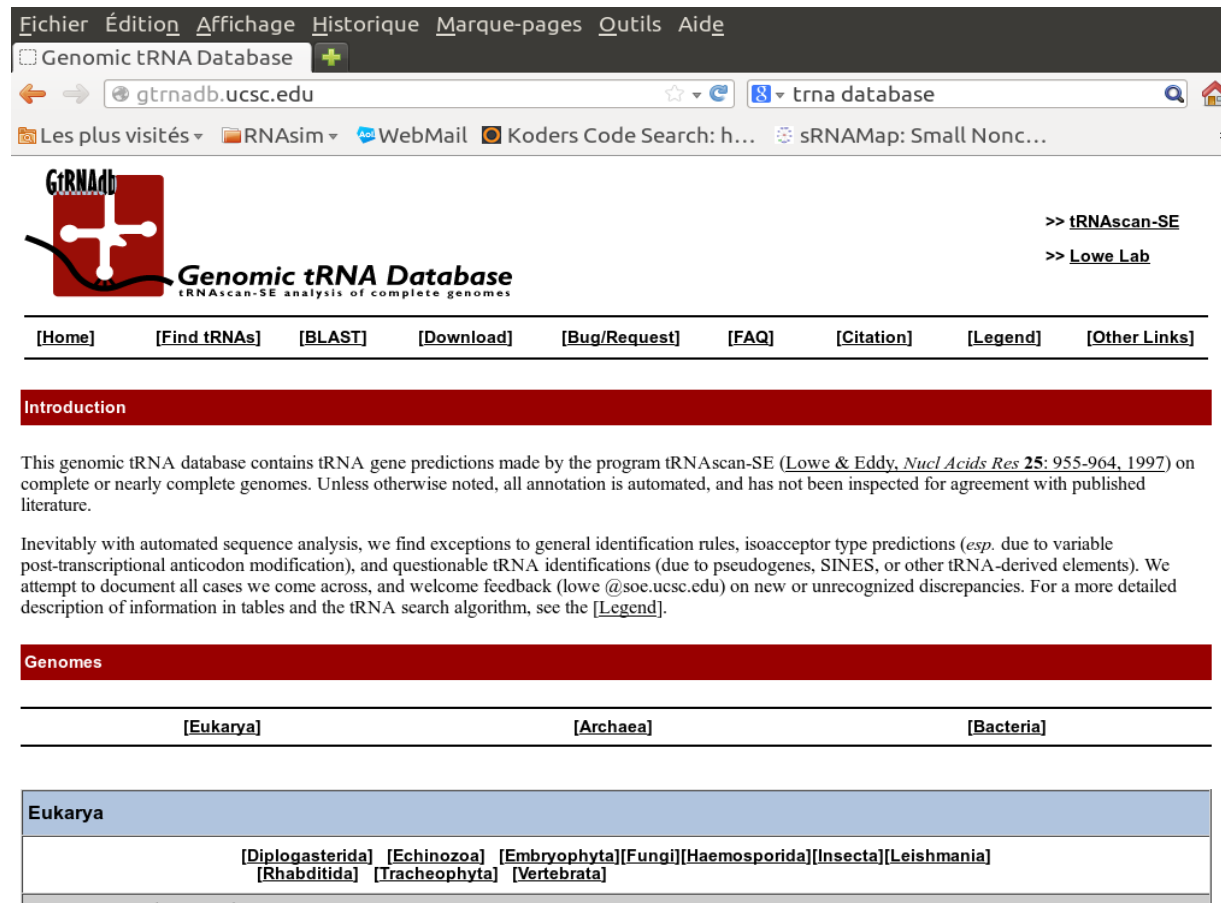
	SSU Parc	SSU Ref	SSU Ref NR	LSU Parc	LSU Ref
Minimal length	300	1200/900	1200/900	300	1900
Quality filtering	basic	strong	strong	basic	strong
Guide Tree	no	yes	yes	no	yes
Release version	111	111	111	111	111
Release date	27.07.12	27.07.12	27.07.12	27.07.12	27.07.12
Aligned rRNA sequences	3,194,778	739,633	286,858	288,717	29,306

Below the table is a section titled "SILVA 114 - web release (Ref datasets and ARB files not updated)".

The left column also contains a "Welcome to the SILVA rRNA database project" section, an "ARB" section, and a "The MEGX.net data portal" section.

The non coding protein RNA world Specific databases

- tRNA



The screenshot shows the Genomic tRNA Database website. The browser address bar displays 'gtrnadb.ucsc.edu'. The website features a navigation menu with links: [Home], [Find tRNAs], [BLAST], [Download], [Bug/Request], [FAQ], [Citation], [Legend], and [Other Links]. Below the navigation menu is an 'Introduction' section, followed by a 'Genomes' section. The 'Genomes' section is organized into a table with columns for [Eukarya], [Archaea], and [Bacteria]. Under the [Eukarya] column, there is a list of taxonomic groups: [Diplogasterida], [Echinozoa], [Embryophyta], [Fungi], [Haemosporida], [Insecta], [Leishmania], [Rhabditida], [Tracheophyta], and [Vertebrata].

Fichier Édition Affichage Historique Marque-pages Outils Aide

Genomic tRNA Database

gtrnadb.ucsc.edu

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc...

Genomic tRNA Database
tRNAscan-SE analysis of complete genomes

>> tRNAscan-SE
>> Lowe Lab

[Home] [Find tRNAs] [BLAST] [Download] [Bug/Request] [FAQ] [Citation] [Legend] [Other Links]

Introduction

This genomic tRNA database contains tRNA gene predictions made by the program tRNAscan-SE (Lowe & Eddy, *Nucl Acids Res* **25**: 955-964, 1997) on complete or nearly complete genomes. Unless otherwise noted, all annotation is automated, and has not been inspected for agreement with published literature.

Inevitably with automated sequence analysis, we find exceptions to general identification rules, isoacceptor type predictions (*esp.* due to variable post-transcriptional anticodon modification), and questionable tRNA identifications (due to pseudogenes, SINES, or other tRNA-derived elements). We attempt to document all cases we come across, and welcome feedback (lowe @soe.ucsc.edu) on new or unrecognized discrepancies. For a more detailed description of information in tables and the tRNA search algorithm, see the [Legend].

Genomes

[Eukarya]	[Archaea]	[Bacteria]
Eukarya [Diplogasterida] [Echinozoa] [Embryophyta] [Fungi] [Haemosporida] [Insecta] [Leishmania] [Rhabditida] [Tracheophyta] [Vertebrata]		

The non coding protein RNA world

Specific databases

- CRISPRdb database

Fichier Édition Affichage Historique Marque-pages Outils Aide

CRISPRs Database

crispr.u-psud.fr/crispr/

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc...

UNIVERSITÉ PARIS-SUD 11

Navigation

- Home page
- CRISPRs database
 - Browse CRISPRs
 - BLAST CRISPRs
 - FlankAlign
 - CRISPRs utilities
 - My CRISPRs DB
- CRISPRs finder
- CRISPRs comparison

Other Tools

- CRISPRcompar
- CRISPRtionary
- CRISPRfinder

GPMS Links

- GPMS Team
- Tandem Repeats DB
- MLVA Web Service

IGM

CRISPRdb

Thursday April 25th 2013

Home About CRISPRs News FAQs Help Contact Us Examples IGM

- View the strains taxonomy browser
- View the strains alphabetical browser
- View the strains in database processing order
- Page manual

	Genomes analysed	CRISPRs found	Strains with convincing CRISPR(s)	Strains devoid of detectable CRISPRs
Archaea	145	551	123	22
Bacteria	2151	3053	1025	653
Total	2296	3604	1148	675

Click on strain name to get more information.

Bacteria

- 'Nostoc azollae' 0708 (1 CRISPR, 3 questionable structures)
- Acaryochloris marina MBIC11017 (1 CRISPR, 2 questionable structures)
- Acetobacter pasteurianus IFO 3283-01 (1 CRISPR)
- Acetobacter pasteurianus IFO 3283-01-42C (1 CRISPR)
- Acetobacter pasteurianus IFO 3283-03 (1 CRISPR)
- Acetobacter pasteurianus IFO 3283-07 (1 CRISPR)

Related Works

CRISPR evolution (*Yersinia pestis*)

Data Summary

	Genomes analysed	CRISPRs found (*)
Archaea	145	551(123)
Bacteria	2151	3053(1025)
Total	2296	3604(1148)

*number of convincing CRISPR structures (number of genomes with such CRISPR)

Database status:

Last update : 2013-03-23

Contact: Christine POURCEL

The non coding protein RNA world

Specific databases

- snoRNA database

The screenshot shows a web browser window displaying the snoRNABase website. The browser's address bar shows the URL <https://www.snorna.biotoul.fr/>. The website has a navigation menu with links: Home, Information, Search, Find guide Rna, Browse, Human yeast snoRNAs, Links, and Contact. The main content area is titled "snoRNABase, a comprehensive database of human H/ACA and C/D box snoRNAs. Version 3". It provides GenBank accession numbers for rRNA sequences: 28S rRNA: U13369 nts 7935-12969, 18S rRNA: X03205, and 5.8S rRNA: U13369 nts 6623-6779. It also includes a citation for Lestrade, L., and Weber, M. J. (2006). A "What's new" section dated July 2007 mentions updates to the human-yeast snoRNA section and the "Find guide RNA" section.

Fichier Édition Affichage Historique Marque-pages Outils Aide

snoRNA-LBME-db, a co...

https://www.snorna.biotoul.fr/ snornadb

Les plus visités RNAsim WebMail Koders Code Search: h...

UUUUGGACCA-----AUAGGAGCUU GCUCCGUCCA-----CUCCACGCA
AGA GGUAACGGG UGGGGUCCGC GCAGUCCGCC:450
AUGGUGA ACUAUGCCUG GGOAGGGCGA AGC

snoRNABase **LBME**

Home Information Search Find guide Rna Browse Human yeast snoRNAs Links Contact

**snoRNABase, a comprehensive database of
human H/ACA and C/D box snoRNAs.
Version 3**

The GenBank accession numbers for the rRNA sequences used are:
28S rRNA: U13369 nts 7935-12969
18S rRNA: X03205
5.8S rRNA: U13369 nts 6623-6779

These sequences, and the numbering of nucleotides, differs from those used in the previous version of the database, and from those used by certain authors.

If you make use of the data presented here, please cite the following article in addition to the primary data sources:
- Lestrade, L., and Weber, M. J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res 34, D158-162.

What's new

July 2007:
The human-yeast snoRNA section has been modified to use the same yeast rRNA sequences (U53879) as the [Yeast snoRNA database at UMass-Amherst](#).
Section [Find guide RNA](#) contains two new tables containing a list of modified bases in human 18S and 28S

The non coding protein RNA world

Specific databases

- Plant databases

Fichier Édition Affichage Historique Marque-pages Outils Aide

Arabidopsis Homepage

chualab.rockefeller.edu/gbrowse2/homepage.html

plant ncna database

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc... BioInformatique des A...

PLncDB: Plant Long noncoding RNA Database
released August 24th, 2012

Home GBrowse Download Tutorial Publication Citation

Our lab has released 6,480 lincRNA a

News

[verify 1,340 intergenic regions](#) | Mar 19, 2012
[profile 4,650 lincRNAs](#) | Sep 19, 2011
[find 6,480 lincRNAs](#) | July 19, 2011
[find 13,230 intergenic regions](#) | June, 12, 2011

Visitors

US 299 GB 13
 CN 186 CA 12
 MX 42 PL 11
 DE 18 IT 9
 FR 15 KR 9
 IN 15 JP 7

FLAG counter

Overview

We have identified a large number of Arabidopsis long noncoding RNAs by analysis of 200 tiling array datasets and RNA-seq data derived from 4 libraries. More than 13,000 RNAs were found transcribed from intergenic regions of the *Arabidopsis thaliana* genome. These intergenic transcription units (TUs) could be further classified into following groups:

	Re-analysis of previously reported data				New data in this study	
	Other RNAs	EST[unigene]	Tiling array analysis (Seedlings)	Tiling array analysis (Seeds)	RepTAS	RNA-seq
LincRNAs	36	36	32	61	6,480	278
GATUs	52	69	369	434	—	370
TUCPs	5	0	0	0	22	7
ECTUs	55	83	172	237	6,728	678
OITUs	1	8	1	6	—	7
Total intergenic TUs	149	196	574	738	13,230	1,340

Long noncoding RNAs are expressed in a temporal and/or spatial specific manner. Many genomic regions encoding long noncoding RNAs were found to be associated with DNA methylation and histone modification, such as H3K4me3, H3K27me3, H3K36me3 and H3K9me3, etc. To provide comprehensive information for the plant research community, we collected a variety of RNA-seq, tiling array, CHIP-chip, CHIP-seq and small RNA datasets, and integrated them into the genome browser. These datasets are shown in following table:

The non coding protein RNA world Specific databases

• Bacteria small regulatory database

Fichier Édition Affichage Historique Marque-pages Outils Aide

Bacterial Small Regulat...

← → bac-srna.org/BSRD/index.jsp# ☆ http://kwanlab.bio.cuhk.edu.hk/BSRD

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc... BioInformatique des A... Metagenomics — Bioi...

Menu

- Home
- Search BSRD
- Hierarchical taxonomy
- Regulatory network
- BLAST BSRD
- Download
- sRNADeep
- Submission
- Latest publications

Home

- [About BSRD](#)
- Current release by host
- Current release by sRNAs
- Contact us
- Contributions by others
- FAQ
- Help
- Latest Update

BSRD
Bacterial Small Regulatory RNA Database

In bacteria, small (~30-500 nt) non-coding RNAs (sRNAs) are the most abundant class of post-transcriptional regulators that are involved in diverse processes including quorum sensing, stress response, virulence and carbon metabolism. Based on the target molecules, sRNAs can be divided into two major groups: (i) mRNA-binding antisense sRNAs and (ii) protein-binding sRNAs. The antisense RNAs can further be categorized as cis-encoded antisense sRNAs, which are completely complementary to their targets, and trans-encoded antisense sRNAs, which are only partially complementary to their targets. In any case, the interaction between antisense RNAs and target mRNAs could direct a plethora of biological regulatory circuits. Recent developments in high-throughput techniques, such as genomic tiling arrays and RNA-Seq have provided invaluable insights into the detection and characterization of bacterial sRNAs. However, a comprehensive bacterial sRNA database is not yet available, especially for integrating and analyzing high-throughput sequencing data.

Here, we have designed and constructed BSRD (Bacterial Small regulatory RNA Database) which hosts sRNAs collected from over 783 bacterial species and 957 strains.

The distinctive features of BSRD are:

- (1) BSRD hosts sRNAs retrieved from online databases including Rfam, sRNAMap, GenBank, RegulonDB and EcoCyc, as well as manual curation. In addition, we have also integrated 20,115 regulatory elements in BSRD.
- (2) BSRD collects sRNAs targets predicted by computational algorithms, IntaRNA and RNAplex, as well as experimentally validated sRNAs targets in sRNATarBase and related literatures.
- (3) BSRD includes information on regulatory relationships between transcription factors (TF) and their target genes, which could provide insights into the combinatorial regulations of sRNAs and TF to their common targets.
- (4) BSRD has integrated expression data from NCBI GEO (Gene Expression Omnibus), which provides detailed evidences for sRNA expression profiling and re-annotation.
- (5) BSRD includes multiple new sRNA annotations from manually curated literature mining, including growth phase, Hfq binding, dual function and Rho-independent terminators.
- (6) BSRD harbors a novel RNA-Seq analysis platform, sRNADeep, that allows perform comprehensive sRNA expression profiling and differential expression analysis in large-scale transcriptome sequencing projects. With the aid of sRNADeep, users can (i) filter low-quality

ncRNA annotation

***Methods and tools for ncRNA
prediction and annotation***

The non coding protein RNA world

Methods and tools

- **A variety of existing approaches not always easy to integrate**
 - Known family: Homology prediction
 - Specific family methods
 - Generalist
 - New family: *De novo* prediction
 - Orphan promoter/terminator
 - Comparative analysis of related organisms
 - Bias composition detection between coding and non coding

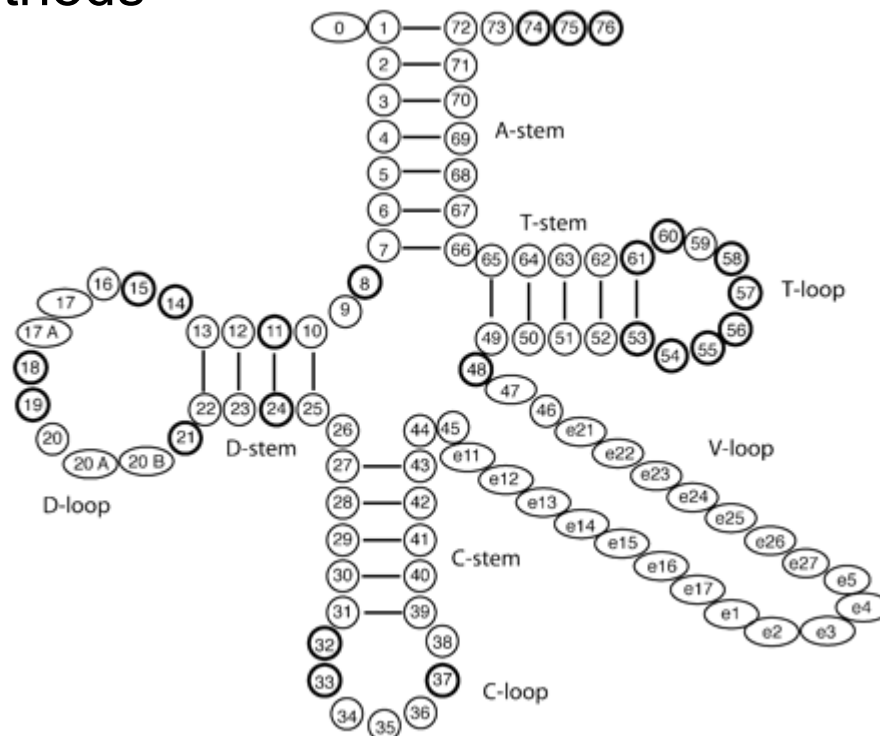
The non coding protein RNA world

Methods and tools

- **Known family: Homology prediction**

- Specific family methods

- tRNA



The non coding protein RNA world

Methods and tools

- **Known family: Homology prediction**
 - Specific family methods
 - tRNA
 - tRNAscan-SE (1997)

© 1997 Oxford University Press

Nucleic Acids Research, 1997, Vol. 25, No. 5 **955–964**

tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence

Todd M. Lowe and Sean R. Eddy*

Department of Genetics, Washington University School of Medicine, 660 South Euclid, Box 8232, St Louis, MO 63110, USA

The non coding protein RNA world Methods and tools

• Known family: Homology prediction

– Specific family methods

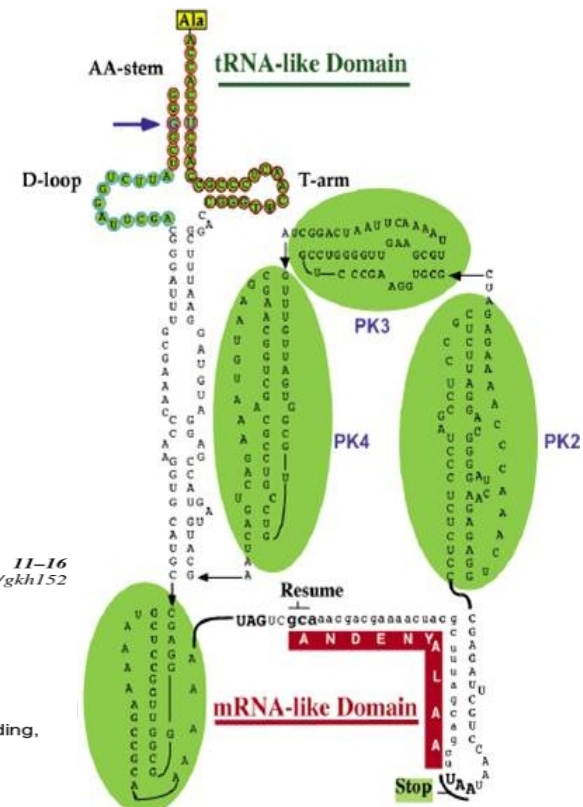
• tRNA

– TRNAscan-SE (1997)

• tRNA+tmRNA (bacteria)

– ARAGORN (2004)

– Single chain tmRNA (ssra)



Nucleic Acids Research, 2004, Vol. 32, No. 1 11-16
DOI: 10.1093/nar/gkh152

ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences

Dean Laslett and Bjorn Canback^{1,*}

Murdoch University, Perth, Western Australia, Australia and ¹Department of Microbial Ecology, Ecology Building, Lund University, S-223 62, Sweden

The non coding protein RNA world

Methods and tools

- **Known family: Homology prediction**
 - Specific family methods
 - rRNA
 - Bacteria and archaea :
 - » 16S rRNA
 - » 23S rRNA
 - » 5S rRNA
 - Eucaryotes
 - » 18S rRNA
 - » 28S rRNA
 - » 5.8S rRNA
 - » 5S rRNA

The non coding protein RNA world

Methods and tools

- **Known family: Homology prediction**
 - Specific family methods
 - rRNA
 - 5S, 16S/18S, 23S/28S
 - Alignment+HMM

3100–3108 *Nucleic Acids Research*, 2007, Vol. 35, No. 9
doi:10.1093/nar/gkm160

Published online 22 April 2007

RNAmmer: consistent and rapid annotation of ribosomal RNA genes

**Karin Lagesen^{1,2,*}, Peter Hallin³, Einar Andreas Rødland^{1,2,4,5}, Hans-Henrik Stærfeldt³,
Torbjørn Rognes^{1,2,4} and David W. Ussery^{1,2,3}**

¹Centre for Molecular Biology and Neuroscience and Institute of Medical Microbiology, University of Oslo, NO-0027 Oslo, Norway, ²Centre for Molecular Biology and Neuroscience and Institute of Medical Microbiology, Rikshospitalet-Radiumhospitalet Medical Centre, NO-0027 Oslo, Norway, ³Center for Biological Sequence Analysis, Biocentrum-DTU, Technical University of Denmark, DK-2800 Lyngby, Denmark, ⁴Department of Informatics, University of Oslo, PO Box 1080 Blindern, NO-0316 Oslo, Norway and ⁵Norwegian Computing Center, PO Box 114 Blindern, NO-0314 Oslo, Norway

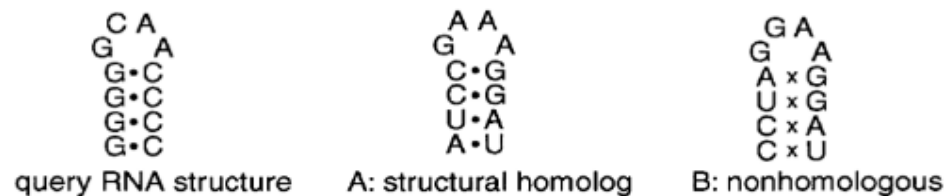
- # STOCKHOLM 1.0
 #=GF AU Infernal 1.0.2

Smr7C GAGCGGCACUCUAUCAAUG.CCGUGAGU..CUGGU.G..AUUGGCGUGUCCCCGCCG.C.CACCAGA..UGAGGACAAAGGCCUCAUCC.CCCUCUCGCGGCCUUUGUCCCGCUUUCAAAGA..G-CCGCGCAGCGGC...AGCCUCCG.CGCCUGGCGGCUUUUU
 Smedr7C GAGCGGCACUCUAUAAAUG.CCGUGUGU..CUGGU.G..AUUGGCGUGUCCCCGCCG.C.CACCAGA..UAGGACAAAGGCCUCAUCC.CCCUCUCGCGGCCUUUGUCCCGCUUUCAAAGA..AACGCUACGCGGC...CGCUCUGC.CGCCUGGACGCUUUUU
 Str7C GAGCGGCACUAACAAUG.CCGUGU-GU..CUGGU.G..AUUGGCGUGUCCCCGCCG.C.CACCAGA..UGUGGACAAAGGCCUCAUCC.CCCUCUCGCGGCCUUUGUCCCGCUUUCACAGA..G-CCGCUACGCGGC...CGCUCUGC.CGCCUGGACGCUUUUU
 Atr7C AGAGCGCGCGCUGAUCAG.CCGUA-GU..CUGAUAG.c-UUGCUGUGUCCCCGCCG-CG.CUGAGA..CC-GGACAAAGGCCUU-UCC.CCCUCUCGCGGCCUUUGUCC----UCCAAUAV..GACCGCAUGCGC..aCCCCUCC.GGCCAUGGCGGCUUUA
 AH13r7C UGGACGGCGCGUGAUCAG.CCGUA-GU..CUGAUAG.G..UAGGCGUGUCCCCGCCG-C.AUCAGA..CC-GGACAAAGGCCUUUCC.CCCUCUCGCGGCCUUUGUCC----UCUUAUAV..GACCGCAUGCGC..AGCCCUCC.GGCCAUGGCGGCUUUUU
 ReCIA7r7C GAUUGC-----UG.CCGAGGUGU..CUGAU.C..ACCGGCUUGUCCCCGCCG.C..AUCAGG..CUUGGACAAAGGCCUUUCC.CCCUCUCGCGGCCUUUGUCCCAUUUU---AA..GGUCGCAUGGCC-...AACCUCCA.GGCCAUGGCGGCUUUU
 Arr7CI GAUUG-----AUUAC.CCAGGGU..CUGAC.A..ACCGCGUGUCCCCGCCA-U.CUGAGA..CUUGGACAAAGGCCUA-UCC.CCCUCUCGCGGCCUUUGUCCCUUUUU---CUA..AGCGCGCAUGGC-...CCCUCCA.GGCCAUGGCGGCUUUU
 Rlt2304r7C GAUUGC-----G.CCGCGCGGU..CUGAU.U..ACCGGCUUGUCCCCGCCG.C..CUGAGG..CUUGGACAAAGGCCUUUCC.CCCUCUCGCGGCCUUUGUCCCCAUUU---UAA..GGUCGCAUGGCC-...AACCUCCA.GGCCAUGGCGGCUUUU
 Avr7CI CAGCGGCGCAUGAUAGCA.CCGCA-GU..CUGAA.A..GUUGGCGUGUCCCCGCCG-C..UUCAGA.cUUGGACAAAGGCCUUUCC.CCCUCUCGCGGCCUUUGUCCCGCUUAGUAGAA..AGUGCUACGCGGC...CGCUCUG-a-GCGGACGACGCUUUU
 RlyrC GAUUGC-----G.UCGCGGGU..CUGAU.U..ACCGGCUUGUCCCCGCCG.C..CUGAGG..CUUGGACAAAGGCCUUUCC.CCCUCUCGCGGCCUUUGUCCCAUU---UAA..AGUGCGCAUGGCC-...AACCUCCA.GGCCAUGGCGGCUUUU
 Rlt1325r7C GAUUGC-----G.UCGCGGGU..CUGAU.U..ACCGGCUUGUCCCCGCCG.C..CUGAGG..CUUGGACAAAGGCCUUUCC.CCCUCUCGCGGCCUUUGUCCCAUU---UAA..AGUGCGCAUGGCC-...AACCUCCA.GGCCAUGGCGGCUUUU
 ReCFNr7C GAUUGC-----G.UCGCGGGU..CUGAU.U..ACCGGCUUGUCCCCGCCG.C..CUGAGG..CUUGGACAAAGGCCUUUCC.CCCUCUCGCGGCCUUUGUCCCAUUUU---AA..AGUGCGCAUGGCC-...AACCUCCA.GGCCAUGGCGGCUUUU
 Mlr7C CACGGGCAUUAUUU---G.CCGAA-GU..CCGGA.G..CAAGGCGCGUCCCCGCCUA.U.CCGGA..CAAUUGGCGCAGCUUACC.CCCUCUCGCGGCUUGCGCAUGGCUUGAGUAU..GAGGCGCGGUGU-U.CCCCUCC-a.CCAGGCGGCGCUUUU
 MbsCNr7C UAUUGCGCAUAUAUU-UUUG.CCGAA-G-..UCCG.G..AAAGGCGUGUCCCCGCCGAA.C.CGGA.CA..UAAAGCGGCAUUAUCC.CCCUCUCGCGGCGUGCC----CACAUUAUAUAA..GGCGCGCGGUGU..CCCCUCC-..ACCACGCGGCGCUUU
 Mer7C CACGGGCAUUAUUU---G.CCGA-GU..CCGGA.G..CAAGGCGCGUCCCCGCCUA.U.CCGGA..CACUUGGCGCAGCUUACC.CCCUCUCGCGGCGCGCAUGGCUUGAGUAUAGAAAGGGCGCGGUGU-U.CCCCUCC-..ACCACGCGGCGCUUU
 Mcr7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Bs23445r7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Bml6Mr7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 BaS19r7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Bm2345r7r7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Bs1330r7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Ba1994r7r7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Rmar7r7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Bor7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Bmr7r7CI CAUUGCACAUUUC-----G.UCGGU-CU..CCGGU.U..UCCGACCGGUGCCCCGUGGAA.ACCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..G-CCGUGUGGUA...UC-CCCU.CACACAGCGGCGCAU
 Oar7CI AAUUCUGAGUAUGCAGCAGUUGCGUGGCGCGGU-U..UCCGACCGGUGCCCCGUGGAA-UCCGGA..UGGUAGCGGUAUGCUUACC.CCCUCUCGCGGCGCGGCUUGAGUAUGAGUA..CAUGGCGGUGCGGUAaCCCCUCC.CGUUAGCGGCAUAUCA

#=GC SS cons
 #CG RF

The non coding protein RNA world Methods and tools

- **Known family : homology prediction**
 - Generic methods
 - Sequence alignment versus structural alignment



primary sequence alignment scoring:

<pre> query: GGGGGCAACCCC x x x x x x x x x A: AUCCGAAAGGAU </pre>	<pre> query: GGGGGCAACCCC x x x x x x x x x B: CCUAGAAAGGAU </pre>
-6	-6

structure + sequence alignment scoring:

<pre> query: GGGGGCAACCCC A: AUCCGAAAGGAU </pre>	<pre> query: GGGGGCAACCCC B: CCUAGAAAGGAU </pre>
+11	-6

The non coding protein RNA world Methods and tools

- **Known family : homology prediction**
 - Generic methods
 - Sequence alignment & structure alignment
 - At the core of Rfam database (<http://rfam.sanger.ac.uk/>)
 - A database of covariance models (probabilistic models of sequence/structure alignments)
 - Sequence alignment tool : Blastn
 - Structure alignment tool : Infernal (cmsearch)

*D136–D140 Nucleic Acids Research, 2009, Vol. 37, Database issue
doi:10.1093/nar/gkn766*

Published online 25 October 2008

Rfam: updates to the RNA families database

**Paul P. Gardner^{1*}, Jennifer Daub¹, John G. Tate¹, Eric P. Nawrocki²,
Diana L. Kolbe², Stinus Lindgreen³, Adam C. Wilkinson¹, Robert D. Finn¹,
Sam Griffiths-Jones⁴, Sean R. Eddy² and Alex Bateman¹**

¹Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK, ²Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, Virginia, USA, ³Center for Bioinformatics, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen N, Denmark and ⁴Faculty of Life Sciences, The University of Manchester, Manchester M13 9PL, UK



The non coding protein RNA world

Methods and tools

- **Known family : homology prediction**
 - Generic methods
 - Descriptor-based methods
 - Subjective and painful descriptor generation
 - Subtle constraints not easily expressed
 - Yes/no answer (no scoring)

© 1996 Oxford University Press

Nucleic Acids Research, 1996, Vol. 24, No. 8 1395-1

Constraints (2008) 13:91-109
DOI 10.1007/s10601-007-9033-9

Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence databases

Bernard Billoud*, Milutin Kontic and Alain Viari

Atelier de Bio-Informatique U

BIOINFORMATICS ORIGINAL PAPER Vol. 2

Genome analysis

Searching RNA motifs and their intermolecular contacts with constraint networks

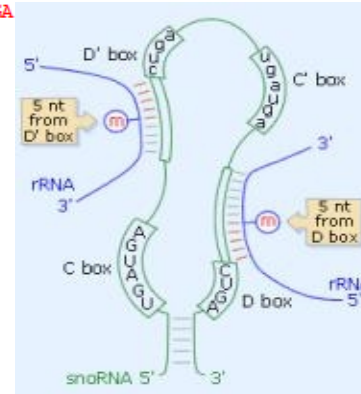
P. Thébault^{1,2}, S. de Givry¹, T. Schiex¹ and C. Gaspin^{1,*}

¹Unité de Biométrie & Intelligence Artificielle INRA, Chemin de Borde Rouge, Auzeville, BP 52627, 31326 Castanet-Tolosan, France and ²Plateforme Bioinformatique, INRA, Chemin de Borde Rouge, Auzeville, BP 52627, 31326 Castanet-Tolosan, France

DARN! A Weighted Constraint Solver for RNA Motif Localization

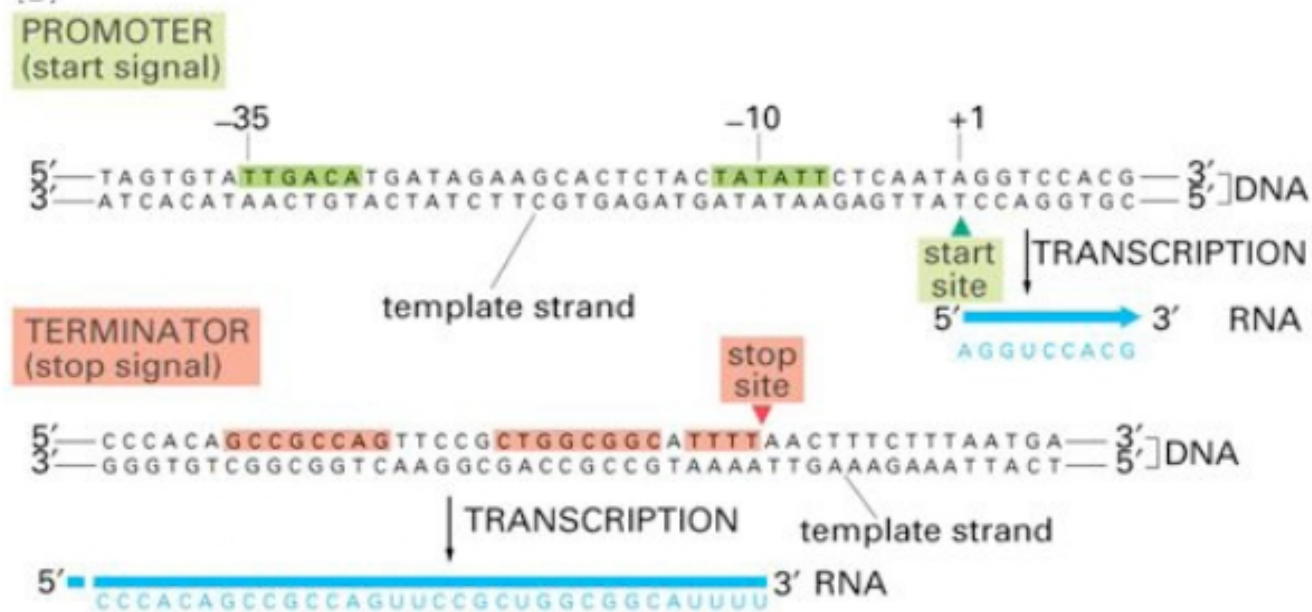
Matthias Zytnicki · Christine Gaspin · Thomas Schiex

GC0CCUUG	UGAUGA	UAGAUUUCAU	CCGA	GCCACC
GC0CCUUG	UGACGA	UGGACGUUUU	CUGA	GCCACC <u>CA</u>
GC0CCUUG	UGACGA	GGCCGUUUCU	CUGA	GUCGCC
GC0CCUUG	UGACGA	GGCCGUUUCU	CCGA	GUCGCC
GC0CCUUG	UGAUGA	GGGUUUUCCA	CCGA	GCCACC
GC0CCUUG	UGACGA	GGCACUUAU	CCGA	GCCACC
GC0CCUUG	UGAUGU	GGUUAUUUAU	CUGA	GCCACC <u>CA</u>
UCCGG	UGAUGA	AGCAGGGGG	CUGA	UGUCCU
UUUUUCCGGA	UGACGA	UAUCAGCACUAU	CUGA	CAAGACUA
GAUUUUUGGA	UGAAGA	CAUCAGCACUAU	CUGA	CACAGC
GAUUUUAGGA	UGACGA	CAUCAGCACUAU	CUGA	CACAGCUA
GAUUUUUGGA	UGAAGA	CAUCAGCACUAU	CUGA	CACAGCUA
GAUUUUUGGA	UGAAGA	CAUCAGCACUAU	CUGA	CACAGCUA
UGUUUACGGA	UGACGA	CAUCAGCACUAU	CUGA	GCGAGC
CGCAAGGUUG	UGAUGA	UGCGGUGU	CUGA	UGUCCU
GCAAAAUA	UGACGA	UAAACUCUAA	CUGA	UGCCGC
GCAAAAAG	UGAUGA	GAACAAUUUC	CUGA	UGCC
GCAAAAAG	UGAUGA	GAACAAUUUC	CUGA	UGCCGC
GCAAAAAG	UGAUGA	GAACAAUUUC	CUGA	UGCCGC
CAUUCG	UGAUGA	AGCAGGGGA	CUGA	TG
UUUUUUGGA	UGAAGA	AAUCGGCACUGU	CUGA	GAGGU
GGA	UGAUGA	UAAGAGGGUAG	CCGA	GGCUU
GGA	UGAUGA	UAAGAGGGUAG	CCGA	GGCUU
AGA	UGAUGA	CAAGAGGGUAG	CCGA	GGCUU
CUGA	UGAUGA	AAAGAGGGUAG	CCGA	GGCCA



The non coding protein RNA world Methods and tools

- **Unknown family**
 - Orphan promoter/terminator

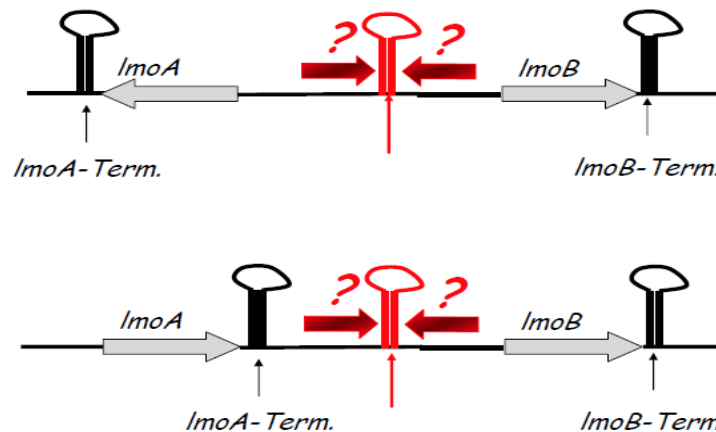


The non coding protein RNA world

Methods and tools

- **Unknown family**
 - Orphan promoter/terminator

Identification by "orphan" terminators prediction



p-independent terminators are identified by standard pattern-searching methods.

The non coding protein RNA world Methods and tools

- **Unknown family**
 - Orphan promoter/terminator
 - Successfully applied in *E. coli*

Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*

Liron Argaman^{*‡}, Ruth Hershberg^{*‡}, Jörg Vogel^{*‡}, Gill Bejerano^{*}, E. Gerhart H. Wagner[†], Hanah Margalit^{*} and Shoshy Altuvia^{*}

Background: Small, untranslated RNA molecules were identified initially in bacteria, but examples can be found in all kingdoms of life. These RNAs carry out diverse functions, and many of them are regulators of gene expression. Genes encoding small, untranslated RNAs are difficult to detect experimentally or to predict by traditional sequence analysis approaches. Thus, in spite of the rising recognition that such RNAs may play key roles in bacterial physiology, many of the small RNAs known to date were discovered fortuitously.

Results: To search the *Escherichia coli* genome sequence for genes encoding small RNAs, we developed a computational strategy employing transcription signals and genomic features of the known small RNA-encoding genes. The search, for which we used rather restrictive criteria, has led to the prediction of 24 putative sRNA-encoding genes, of which 23 were tested experimentally. Here we report on the discovery of 14 genes encoding novel small RNAs in *E. coli* and their expression patterns under a variety of physiological conditions. Most of the newly discovered RNAs are abundant. Interestingly, the expression level of a significant number of these RNAs increases upon entry into stationary phase.

Conclusions: Based on our results, we conclude that small RNAs are much more widespread than previously imagined and that these versatile molecules may play important roles in the fine-tuning of cell responses to changing environments.

Addresses: ^{*}Department of Molecular Genetics and Biotechnology, The Hebrew University-Hadassah Medical School, Jerusalem 91120, Israel. [†]Institute of Cell and Molecular Biology, Biomedical Center, Uppsala University, Box 596, Uppsala 751 24, Sweden.

Correspondence: Shoshy Altuvia, Hanah Margalit, Gerhart Wagner
E-mail: shoshy@cc.huji.ac.il
hanah@md2.huji.ac.il
gerhart.wagner@icm.uu.se

^{*}These authors contributed equally to this work.

Received: **2 May 2001**
Revised: **21 May 2001**
Accepted: **23 May 2001**

Published: **26 June 2001**

Current Biology 2001, 11:941–950

0960-9822/01/\$ – see front matter
© 2001 Elsevier Science Ltd. All rights reserved.

The non coding protein RNA world Methods and tools

- **Unknown family**
 - Orphan promoter/terminator

Research

Open Access

Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake

Carleton L Kingsford, Kunmi Ayanbule and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA.

Correspondence: Carleton L Kingsford. Email: carlk@umiacs.umd.edu

Published: 21 February 2007

Genome **Biology** 2007, **8**:R22 (doi:10.1186/gb-2007-8-2-r22)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2007/8/2/R22>

Received: 14 September 2006

Revised: 1 December 2006

Accepted: 21 February 2007

Published online 7 April 2011

Nucleic Acids Research, 2011, Vol. 39, No. 14 5845–5852
doi:10.1093/nar/gkr168

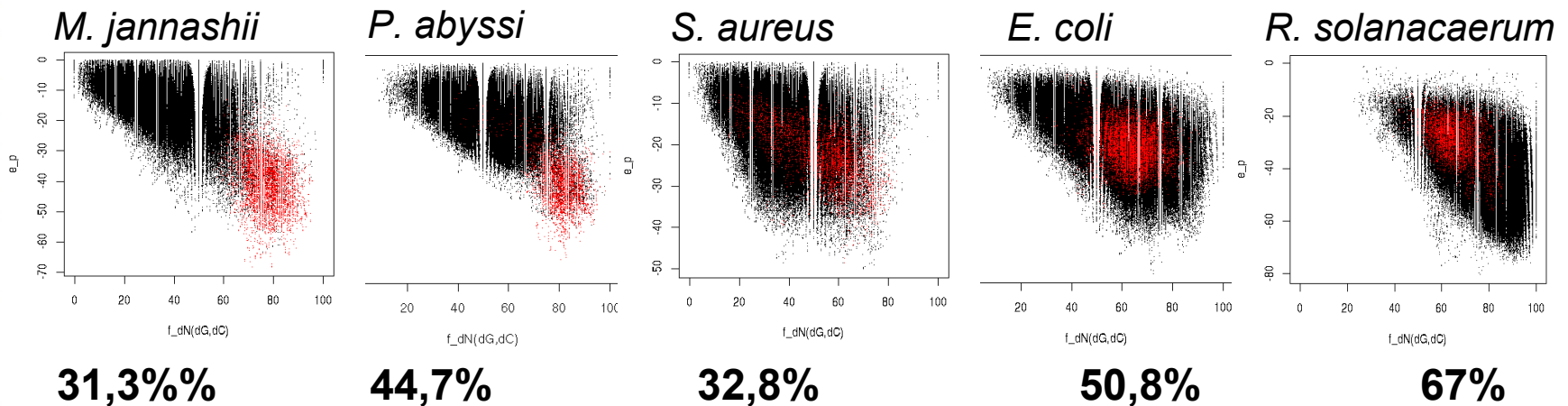
RNIE: genome-wide prediction of bacterial intrinsic terminators

Paul P. Gardner^{1,*}, Lars Barquist¹, Alex Bateman¹, Eric P. Nawrocki² and Zasha Weinberg³

¹Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA0, UK, ²Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147 and ³Howard Hughes Medical Institute, Yale University, Box 208103, New Haven, CT 06520, USA

The non coding protein RNA world Methods and tools

- **Unknown family**
 - Bias composition analysis
 - Schattner, NAR, 2002, Klein & Eddy, PNAS, 2002



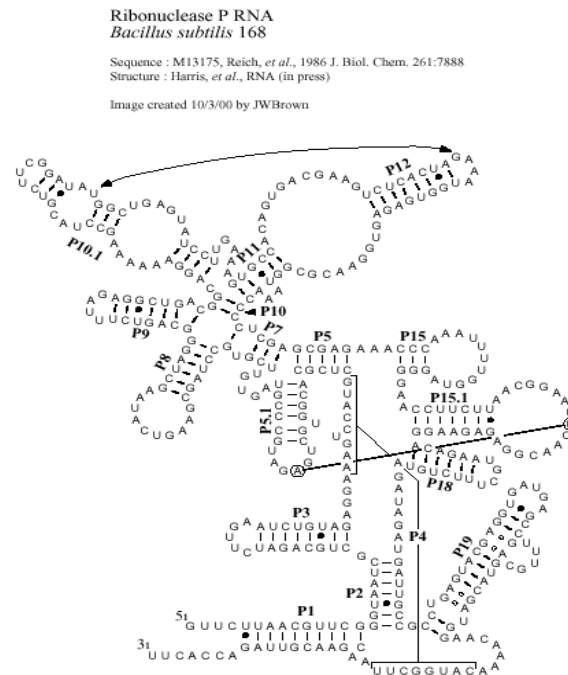
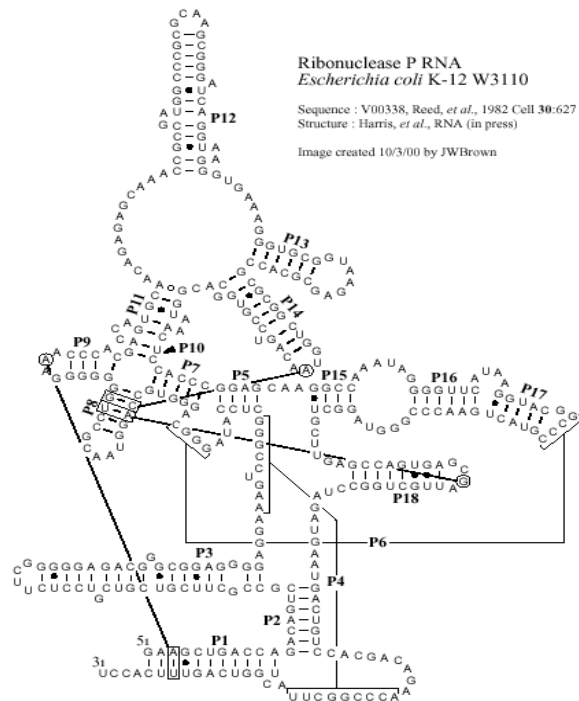
The non coding protein RNA world

Methods and tools

- **Unknown family**
 - Comparative analysis
 - Sequence alignment
 - Structure alignment

The non coding protein RNA world Methods and tools

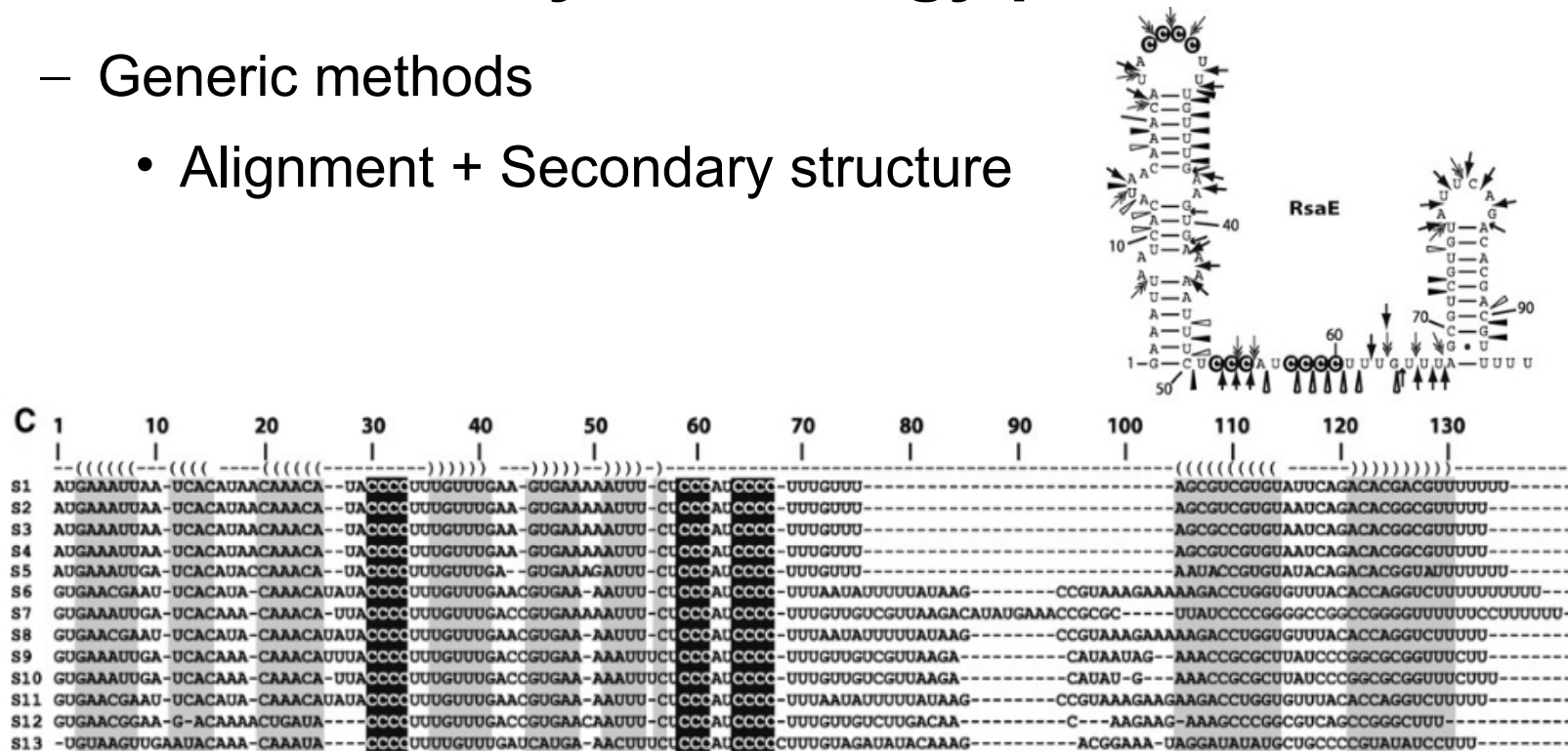
- **Unknown family**
 - Comparative analysis



The non coding protein RNA world

Methods and tools

- **Unknown family: Homology prediction**
 - Generic methods
 - Alignment + Secondary structure



The non coding protein RNA world Methods and tools

- **Unknown family: Homology prediction**
 - RNAz (Washietl et al., 2004)
www.tbi.univie.ac.at/~wash/RNAz
 - Start with an alignment of homologous sequences
 - Compute :
 - Mean free energy of aligned sequences
 - Structure conservation score
 - Mean pairwise identity
 - Number of sequences in the alignment
 - Use a SVM to classify candidates

The non coding protein RNA world

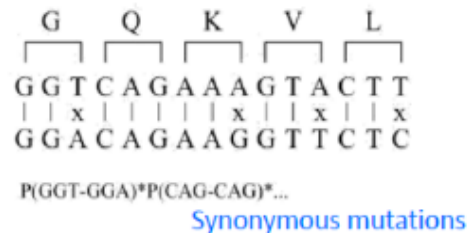
Methods and tools

- **Unknown family: Homology prediction**

- Q-RNA (Rivas & Eddy, 2001)

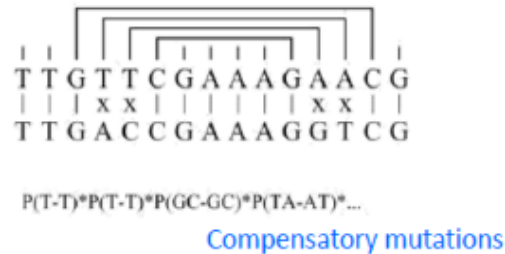
- Start with a blast alignment
- Models to assess coding/non coding

• Model for protein coding gene



• Model for ncRNA

(also include loop probabilities obtained from training set of real ncRNA)



The non coding protein RNA world

Methods and tools

- **miRNA prediction**
 - **De novo prediction**
 - Search for the hairpin structure of the pre-miRNA
 - Hairpin search at the genome scale
 - Exploit conservation between organisms

Example of bacterial sRNA prediction & annotation with RNAspace

RNAspace.org

ncRNA prediction and annotation

RNAspace

The screenshot shows the RNAspace.org website. At the top is a navigation bar with links: [Fichier](#), [Édition](#), [Affichage](#), [Historique](#), [Marque-pages](#), [Outils](#), and [Aide](#). Below this is a browser window showing the address bar with www.rnaspacespace.org and a search bar with the text "rnaspacespace". The main content area has a blue header with the text "RNAspace.org" and "The non-coding RNA annotation platform". To the right of the header is a diagram of an RNA secondary structure. Below the header is a navigation bar with links: [Home](#), [1. Load data](#), [2. Predict](#), [3. Explore](#), and a [HELP](#) button. The main content area is divided into two columns. The left column contains a "Welcome to RNAspace" section with a paragraph about the platform's purpose and a list of links: [Partners](#), [FAQ](#), and [Contact](#). Below this is an "Availability" section with a paragraph about the project's open-source status and its availability on Sourceforge. The right column contains a "News" section with a paragraph about the software version v1.2.1. At the bottom of the page is a "Get Started" button and a footer with the text "Comments and remarks: contact@rnaspacespace.org".

Fichier Édition Affichage Historique Marque-pages Outils Aide

RNAspace - Home

www.rnaspacespace.org

rnaspacespace

Les plus visités RNAsim WebMail Koders Code Search: h... sRNAMap: Small Nonc...

RNAspace.org

The non-coding RNA annotation platform

Home 1. Load data 2. Predict 3. Explore HELP

Welcome to RNAspace RNAspace is a platform which aims at providing an integrated environment for non-coding RNA annotation.

The increasing number of ncRNA discovered since 2000 and the lack of user friendly tools for finding and annotating them, have made necessary to propose to biologists an *in silico* environment allowing structural and functional annotations of these molecules with regard to available protein genes annotation environments.

RNAspace makes available a variety of [ncRNA gene finders](#) and [ncRNA databases](#) as well as user-friendly tools to explore computed results including comparison, visualization and edition of putative RNAs. RNAspace also allows to export putative RNAs in various formats.

[Partners](#) [FAQ](#) [Contact](#)

Availability RNAspace is an open source project. It is developed in Python. It is copyrighted with the GNU General Public License, and is free (in the GNU sense) for all to use, and is in constant development. RNAspace is hosted at [Sourceforge](#). It is also available as a web server at rnaspacespace.org.

News RNAspace software v1.2.1 (July 28, 2011) is running this site.

1 Load your data 2 Select prediction programs 3 Explore your results

Get Started

Comments and remarks: contact@rnaspacespace.org

RNAspace.org

ncRNA prediction and annotation

RNAspace

RNAspace.org
The non-coding RNA annotation platform

Home | 1. Load data | 2. Predict | 3. Explore | HELP

Funding RNAspace has been funded in 2007 by the RNG and in 2010 by ReNaBi (french National Network of Genomic Centers).

Partners

BONSAI team, LIFL - UMR CNRS 8022 - Univ. Lille 1 and INRIA Lille Nord Europe	Unité de Biométrie et Intelligence Artificielle, INRA Toulouse	PF Bioinformatique, INRA Toulouse	PF SIGENAE, INRA Toulouse	IGM, UMR 8621 CNRS - Univ. Paris Sud
Benjamin Grenier-Boley (2008) Antoine de Monte (2008-2010) Laurie Tonon (2010-2011) Hélène Touzet	Marie-Josée Cros Christine Gaspin	Christine Gaspin J-Marc Larré (2008) Jérôme Mariette Guilhem Richard (2010)	Philippe Bardou	Daniel Gautheret

Logos: BIA Toulouse (Biométrie et Intelligence Artificielle), CNRS, geno toul bioinfo, ReNaBi, INRA, Inria, lifl, réseau national genopole®, SIGEN@E, Université Lille 1 Sciences et Technologies, UNIVERSITÉ PARIS-SUD 11

Comments and remarks: contact@rnaspace.org

RNAspace.org

ncRNA prediction and annotation

RNAspace

RNAspace.org

The non-coding RNA annotation platform

[Home](#) | [1. Load data](#) | [2. Predict](#) | [3. Explore](#) | [HELP](#)

Enter the genomic sequences in Fasta or multiFasta format (the number of sequences is limited to 300 per multiFasta). For each sequence, you can describe the name, the domain the species, the strain and the replicon. [\[?\]](#)
Remark: There is a global size limitation of **5.0 Mb** for your data.

Upload sequence(s)

Sequence name:

Domain:

Optional information:

Species:

Strain:

Replicon:

Upload sequence(s) in FASTA format from a file:

Or paste it here:

Sequence name	Size	Domain	Species	Strain	Replicon	Header
sample	100001	bacteria	E.coli	K12	Chromosome	Sample sequence Escherichia coli str. K-12 substr. MG1655 4156417:4256417

Email address(es) [\[?\]](#):

Comments and remarks: contact@rnaspace.org

RNAspace.org

ncRNA prediction and annotation

RNAspace

The non-coding RNA annotation platform

Home 1. Load data 2. Predict 3. Explore HELP

For the sake of clarity, available annotation tools are organized in three sections. Select one or several gene finders to analyse your data. [?]
Remark: Maximum allowed running time for a gene finder is 8 hours.

Homology search

These tools identify regions that are similar to known non-coding RNAs. Similarity is detected at the sequence level and/or at the structure level. [?]

<input checked="" type="checkbox"/> BLAST (sequence homology) [more]	Database: Rfam_10.0_seed parameters
<input checked="" type="checkbox"/> Dam (RNA motif search) [more]	Descriptor: snoRNA-CDbox [A] parameters
<input type="checkbox"/> ERPIN (RNA motif search) [more]	Training set: All [domain] parameters
<input type="checkbox"/> INFERNAL (RNA motif search) [more]	Descriptor: 23S-methyl [RF01065] parameters
<input checked="" type="checkbox"/> RNAmmer (specialized) [more]	parameters
<input checked="" type="checkbox"/> tRNAscan-SE (specialized) [more]	parameters
<input type="checkbox"/> YASS (sequence homology) [more]	Database: Rfam_10.0_seed parameters

Comparative Analysis

You can compare your data to a selection of genome sequences from different species to find out significantly conserved regions that exhibit a consensus secondary structure. Only bacterial and archeal genomes are available by now. [?]

1. Select a set of organisms (at most four)

Select an organism

Pyrococcus_furiosus [1 sequences]	remove
Thermococcus_kodakaraensis_KOD1 [1 sequences]	remove

2. Define your comparative analysis method

Sequence alignment	Sequence aggregation	Structure inference
<input checked="" type="radio"/> BLAST [more] parameters <input type="radio"/> YASS [more] parameters	<input checked="" type="radio"/> CG-seq [more] parameters	<input checked="" type="radio"/> caRNAc [more] parameters <input type="radio"/> RNAz [more] parameters

Ab initio prediction

The last kind of prediction tools uses intrinsic statistical feature of the data. Beware that this approach has been successful only in case of hyperthermophile AT rich genomes. [?]

<input checked="" type="checkbox"/> atypicalGC [more] parameters
--

☒ Combine results [?] Run

ncRNA prediction and annotation

P. abyssi – 2 contigs of *Triticum aestivum*

P. abyssi

Known families

Homology search

- Blast → RFAM
- Darn ! → C/D box sRNA
- RNAmmer
- tRNAscan-SE

New families

Comparative analysis

- *P. furiosus*
- *T. kodakarensis*

Bias composition

T. aestivum

Known families

Homology search

- Blast → RFAM
- RNAmmer
- tRNAscan-SE

ncRNA prediction and annotation

P. abyssi – 2 contigs of *Triticum aestivum*

P. abyssi

Known families

Homology search

- Blast → RFAM : **161**
- Darn ! → C/D box sRNA : **123**
- RNAmmer : **4**
- tRNAscan-SE : **46**

New families

Comparative analysis : 20

- *P. furiosus*
- *T. kodakarensis*

Bias composition : 101

T. aestivum

Known families

Homology search : 118

- Blast → RFAM : **118**
 - * miR1122 : → Infernal : **57**
 - * 5S rRNA : → Infernal : **1**
 - * tRNA : **1** → Infernal : **2**
 - * U4 : **1** → Infernal : **1**
 - * Intron gr II

- RNAmmer : **0**
- tRNAscan-SE : **0**

ncRNA prediction and annotation *P. abyssi*

[Home](#)
[1.Load data](#)
[2.Predict](#)
[3.Explore](#)
[HELP](#)

Current results for the 0057383924590e6 project: **417 putatives RNAs predicted.**

Software tools used and user actions are summarized in the left re-sizable table and query sequence(s) in the right re-sizable table. See the project [history](#) for more details.

Run or user identifier	Description	Number of RNAs	Query sequence(s)					
r02	atypicalGC	101	SeqArchae	1765118 nt	archaea	PAbys	unknown	chromosome
r01	Combine	-74+36						
	BLAST/Rfam_10.0_seed	161						
	RNAmer	4						
	tRNAscan-SE	46						
	Dam	123						
	BLAST/CG-seq/caRNAC	20						
	atypicalGC	0						

Field: Operator: Value (wildcards allowed): Result: Opposite, you can apply successive filters on the list of displayed putative RNAs [?].

Criterion: Comparison: Give value: Add/Update: 417/417 RNAs satisfy filter(s)

[Table view](#) [JBrowse view](#) [CGview view](#)

The table of results may be sorted by clicking on the column titles. You can select predictions by ticking the check boxes in the left column and perform actions on them using the down-drop lists below the table [?].

Predictions 1 - 20 of 417

Display: Terse set Show: 20 Page: 1 of 21 > >>

All	ID	Seq name	Family	Start	End	Size	Strand	Species	Domain	Replicon	Software	Align.	Run
<input type="checkbox"/>	000357	SeqArchae	tRNA-Pro	4930	5007	78	+	PAbys	archaea	chromosome	[combine:BLAST/Rfam_1...]	2	r01
<input type="checkbox"/>	000001	SeqArchae	snRNA-CDbox	8633	8668	36	-	PAbys	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000002	SeqArchae	snRNA-CDbox	9855	9906	52	+	PAbys	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000392	SeqArchae	unknown	15220	15275	56	.	PAbys	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000393	SeqArchae	unknown	26561	26624	64	.	PAbys	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000394	SeqArchae	unknown	30198	30302	105	.	PAbys	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000395	SeqArchae	unknown	56891	57217	327	.	PAbys	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000335	SeqArchae	unknown	56903	57229	327	.	PAbys	archaea	chromosome	BLAST/CG-seq/caRNAC	1	r01
<input type="checkbox"/>	000356	SeqArchae	SRP_euk_arch	56923	57215	293	-	PAbys	archaea	chromosome	[combine:BLAST/Rfam_1...]	3	r01
<input type="checkbox"/>	000209	SeqArchae	SRP_bact	57002	57046	45	-	PAbys	archaea	chromosome	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000188	SeqArchae	sR46	57376	57436	61	+	PAbys	archaea	chromosome	BLAST/Rfam_10.0_seed	4	r01
<input type="checkbox"/>	000396	SeqArchae	unknown	58859	58971	113	.	PAbys	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000240	SeqArchae	sR45	64245	64299	55	+	PAbys	archaea	chromosome	BLAST/Rfam_10.0_seed	4	r01
<input type="checkbox"/>	000003	SeqArchae	snRNA-CDbox	64249	64297	49	+	PAbys	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000218	SeqArchae	sR14	65224	65278	55	-	PAbys	archaea	chromosome	BLAST/Rfam_10.0_seed	3	r01
<input type="checkbox"/>	000004	SeqArchae	snRNA-CDbox	65226	65274	49	-	PAbys	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000241	SeqArchae	sR22	65334	65391	58	+	PAbys	archaea	chromosome	BLAST/Rfam_10.0_seed	4	r01
<input type="checkbox"/>	000005	SeqArchae	snRNA-CDbox	65338	65389	52	+	PAbys	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000170	SeqArchae	sR32	67950	68006	57	+	PAbys	archaea	chromosome	BLAST/Rfam_10.0_seed	3	r01
<input type="checkbox"/>	000186	SeqArchae	sR51	73034	73088	55	-	PAbys	archaea	chromosome	BLAST/Rfam_10.0_seed	1	r01

With selected predictions: Edit... Analyse... Export... With all predictions: EXPORT...

ncRNA prediction and annotation *P. abyssi*

<input type="checkbox"/>	000212	SeqArchae	HgcC	163322	163358	37	+	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	1	r01
<input type="checkbox"/>	000400	SeqArchae	unknown	165348	165400	53	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000401	SeqArchae	unknown	174473	174770	298	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000178	SeqArchae	RNaseP_arch	174477	174806	330	+	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000402	SeqArchae	unknown	204807	206988	2182	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000334	SeqArchae	16s_rRNA	205051	206547	1497	+	PAbyssi	archaea	chromosome	RNAmmmer	0	r01
<input type="checkbox"/>	000358	SeqArchae	tRNA-Ala	206606	206683	78	+	PAbyssi	archaea	chromosome	[combine:BLAST/Rfam_1...]	5	r01
<input type="checkbox"/>	000331	SeqArchae	23s_rRNA	206816	209855	3040	+	PAbyssi	archaea	chromosome	RNAmmmer	0	r01
<input type="checkbox"/>	000403	SeqArchae	unknown	207037	207307	271	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000404	SeqArchae	unknown	207309	207985	677	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000405	SeqArchae	unknown	208012	208831	820	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000406	SeqArchae	unknown	208882	209890	1009	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000220	SeqArchae	PK-G12rRNA	209204	209313	110	+	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000190	SeqArchae	snoR9	230449	230575	127	-	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000014	SeqArchae	snoRNA-CDbox	230464	230513	50	-	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000193	SeqArchae	snoPyro_CD	230633	230690	58	+	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	3	r01
<input type="checkbox"/>	000015	SeqArchae	snoRNA-CDbox	230637	230687	51	+	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000230	SeqArchae	sR49	235441	235496	56	+	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	2	r01
<input type="checkbox"/>	000016	SeqArchae	snoRNA-CDbox	235445	235494	50	+	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000208	SeqArchae	sR13	245976	246030	55	+	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	4	r01
<input type="checkbox"/>	000017	SeqArchae	snoRNA-CDbox	245980	246028	49	+	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000233	SeqArchae	sR43	250933	250988	56	-	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	3	r01
<input type="checkbox"/>	000018	SeqArchae	snoRNA-CDbox	250935	250984	50	-	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000194	SeqArchae	snoPyro_CD	258067	258122	56	+	PAbyssi	archaea	chromosome	BLAST/Rfam_10.0_seed	2	r01
<input type="checkbox"/>	000019	SeqArchae	snoRNA-CDbox	258071	258119	49	+	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000337	SeqArchae	unknown	258075	258119	45	.	PAbyssi	archaea	chromosome	BLAST/CG-seq/caRNAC	1	r01
<input type="checkbox"/>	000020	SeqArchae	snoRNA-CDbox	279210	279261	52	+	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000359	SeqArchae	tRNA-His	296677	296753	77	+	PAbyssi	archaea	chromosome	[combine:BLAST/Rfam_1...]	1	r01
<input type="checkbox"/>	000021	SeqArchae	snoRNA-CDbox	301375	301421	47	-	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000022	SeqArchae	snoRNA-CDbox	303264	303317	54	+	PAbyssi	archaea	chromosome	Dam	0	r01
<input type="checkbox"/>	000407	SeqArchae	unknown	303630	303731	102	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02
<input type="checkbox"/>	000408	SeqArchae	unknown	305598	305716	119	.	PAbyssi	archaea	chromosome	atypicalGC	0	r02

RNASpace.org

P. abyssi

User sequence(s)

ID	SeqName	Family	Begin	End	Size	Strand	Software
000348	SeqArchae	unknown	1083324	1083412	89	.	[BLAST/CG-seq/c...]

Database sequence(s)

Sequence Description	database	Begin	End
gi 57639935 ref NC_006624.1 _Thermococcus_kodakarensis_KOD1_complete_genome_2	NC_006624	914605	914704
gi 18976372 ref NC_003413.1 _Pyrococcus_furiosus_DSM_3638_complete_genome_1	NC_003413	1022786	1022875

Alignment produced by Carnac + Gardenia

```
000348      UUCU---G-U--AA--GC---AC-AAAUCGAUAAUUUUUUAUUAACCUUCAUUUAGACAAGUA
%gi|57639935|re UUCUCACGAUAAAUCGCGGCCUA AACCGAUAAUUUUUUAUUAUCUACACACUAGUUGGGUA
%gi|18976372|re UUCU---G-U--AAUAC---AC-AGACGAUAAUUUUUUAUUAACUUUCACCCUAUUUAGUA
          ****  * *  **  *      * *  * ***** **      ***
```

```
000348      CAAAAAGUGUACUACAAAAUCUGUACUUGGUGGU
%gi|57639935|re CAGAAAAUUGUACUACAAAAUCUGUACCCGGUGGU
%gi|18976372|re CAGAAAAUUGUACUACAAAAUUUGUACUAGGUGGG
          **  **** ***** *  *****  ****
```

```
000348      .....(((((
%gi|57639935|re .....(((((
%gi|18976372|re .....(((((
          UUCU  G U  AA  C  C A A GAUAAUUUUUUAUUAUC      UA      GUA
```

```
000348      ((.....)))))....
%gi|57639935|re (((.....)))))....
%gi|18976372|re (((.....)))))....
          CA AAAA UGUACUACAAAA U UGUAC  GGUGG
```

ncRNA prediction and annotation *T. aestivum*

Home 1.Load data 2.Predict 3.Explore

Current results for the 005739160f9a421 project: **179 putatives RNAs predicted.**

Software tools used and user actions are summarized in the left re-sizable table and query sequence(s) in the right re-sizable table. See the project [history](#) for more details.

Run or user identifier	Description	Number of RNAs
r05	INFERNAL	57
r07	INFERNAL	2
r06	INFERNAL	1
r04	INFERNAL	1
User action	Combine	-136+68
r03	BLAST/Rfam_10.0_seed	67
r02	Combine	-0+0
	INFERNAL	1
User action	Combine	-0+0
	Combine	-0+0
r01	Combine	-2+1
	BLAST/Rfam_10.0_seed	119
	RNAmmer	0
	tRNAscan-SE	0

Query sequence(s)					
contig	983767 nt	eukaryote	Wheat	unknown	unknown
contig_914	2522860 nt	eukaryote	unknown	unknown	unknown

Field: Criterion Operator: Comparison Value (wildcards allowed): Give value Result: 179/179 RNAs satisfy filter(s) Add/Update

Opposite, you can apply successive filters on the list of displayed putative RNAs [?].

Table view JBrowse view CGview view

The table of results may be sorted by clicking on the column titles. You can select predictions by ticking the check boxes in the left column and perform actions on them using the down-drop lists below the table [?].

Predictions 1 - 20 of 179 Display Terse set Show 20 Page 1 of 9 >>

	ID	Seq name	Family	Start	End	Size	Strand	Species	Domain	Replicon	Software	Align.	Run
<input type="checkbox"/>	000018	contig	MIR1122	21364	21441	78	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	1	r01
<input type="checkbox"/>	000261	contig	MIR1122	21364	21445	82	+	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000267	contig	MIR1122	21364	21445	82	-	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000002	contig	5S_rRNA	33257	33337	81	+	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	1	r01
<input type="checkbox"/>	000008	contig	5S_rRNA	42578	42643	66	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000003	contig	5S_rRNA	48687	48750	64	+	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	1	r01
<input type="checkbox"/>	000019	contig	MIR1122	49455	49498	44	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000011	contig	MIR1122	49458	49498	41	+	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	3	r01
<input type="checkbox"/>	000009	contig	5S_rRNA	65214	65278	65	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000010	contig	5S_rRNA	126714	126785	72	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000258	contig	5S_rRNA	150056	150190	135	+	Wheat	eukaryote	unknown	INFERNAL	0	r06
<input type="checkbox"/>	000197	contig	5s_rRNA	150056	150190	135	+	Wheat	eukaryote	unknown	[combine:BLAST/Rfam_1...]	5	explore
<input type="checkbox"/>	000268	contig	MIR1122	151269	151342	74	-	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000259	contig	MIR1122	221919	222033	115	+	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000265	contig	MIR1122	221919	222033	115	-	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000020	contig	MIR1122	221922	222032	111	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	2	r01
<input type="checkbox"/>	000012	contig	MIR1122	221923	222031	109	+	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	2	r01
<input type="checkbox"/>	000193	contig	MIR1122	208145	208146	2	-	Wheat	eukaryote	unknown	[combine:BLAST/Rfam_1...]	6	explore

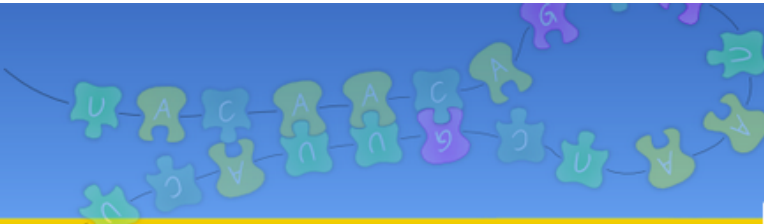
T. aestivum:U4 ?

<input type="checkbox"/>	Accession	Contig	Gene	Start	End	Size	Strand	Species	Category	Method	Score	View	
<input type="checkbox"/>	000259	contig	MIR1122	221919	222033	115	+	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000265	contig	MIR1122	221919	222033	115	-	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000020	contig	MIR1122	221922	222032	111	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	2	r01
<input type="checkbox"/>	000012	contig	MIR1122	221923	222031	109	+	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	2	r01
<input type="checkbox"/>	000193	contig	MIR1122	398065	398146	82	-	Wheat	eukaryote	unknown	[combine:BLAST/Rfam_1...]	6	explore
<input type="checkbox"/>	000189	contig	MIR1122	398066	398146	81	+	Wheat	eukaryote	unknown	[combine:BLAST/Rfam_1...]	6	explore
<input type="checkbox"/>	000257	contig	U4	607456	607607	152	+	Wheat	eukaryote	unknown	INFERNAL	0	r04
<input type="checkbox"/>	000199	contig	U4	607456	607586	131	+	Wheat	eukaryote	unknown	[combine:BLAST/Rfam_1...]	10	explore
<input type="checkbox"/>	000264	contig	MIR1122	629622	629719	98	+	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000263	contig	MIR1122	631502	631599	98	+	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000198	contig	5s_rRNA	759594	759659	66	+	Wheat	eukaryote	unknown	[combine:BLAST/Rfam_1...]	12	explore
<input type="checkbox"/>	000194	contig	MIR1122	795165	795289	125	-	Wheat	eukaryote	unknown	[combine:BLAST/Rfam_1...]	7	explore
<input type="checkbox"/>	000260	contig	MIR1122	795166	795289	124	+	Wheat	eukaryote	unknown	INFERNAL	0	r05

T. aestivum:U4 ?

RNAspace.org

The non-coding RNA annotation platform



RNA features

ID: 000199
Family: U4
Sequence name: contig (*Wheat unknown - eukaryote - unknown*)
Start: 607456
End: 607586
Strand: +
Predicted by: *combine:BLAST/Rfam_10.0_seed* - with a 0.0 score on 25-4-2013

Genome context

GTATAAGAACGGACCATGGCGTATGGAAATGGGC ... [000199] ... TCCTTGGAGAGGGCAAGGGCCTACGAATTAAATAA

Sequence and structure(s)

000199	1	ATTTTTCGCTTGGGGCAATGACGCACCTAGTGAGGTAATACCGAGGCGCGTCAATTGCTGGTTGAAAACATTTCCAAACTCCCTCTTTGGCCCTCACG
	101	GGTCACTGAGAATTTGTGCAAAGGCTCCCTC

Alignment(s)

This prediction is included in 10 [alignment\(s\)](#)

[Edit](#) [Back to explore](#)

Comments and remarks: contact@rnaspace.org

RNAspace.org

T. aestivum:U4 ?

HELP

Page 1 of 1

User sequence(s)

ID	SeqName	Family	Begin	End	Size	Strand	Software
000199	contig	U4	607456	607586	131	+	[combine:BLAST/...]

Database sequence(s)

Sequence Description	database	Begin	End
>U4 J302335.1/69797-69946_JRF00015	Rfam_10.0_seed	1	139
>U4 M479189.1/4511-4661_JRF00015	Rfam_10.0_seed	1	138
>U4 JAA02007064.1/44768-44912_JRF00015	Rfam_10.0_seed	1	130
>U4 J302335.1/69797-69946_JRF00015	Rfam_10.0_seed	1	139
>U4 JAA02007064.1/44768-44912_JRF00015	Rfam_10.0_seed	1	130
>U4 P004858.3/48993-49137_JRF00015	Rfam_10.0_seed	1	130
>U4 ARH01003623.1/42069-42219_JRF00015	Rfam_10.0_seed	1	137
>U4 M479189.1/4511-4661_JRF00015	Rfam_10.0_seed	1	138
>U4 ARH01003623.1/42069-42219_JRF00015	Rfam_10.0_seed	1	137
>U4 P004858.3/48993-49137_JRF00015	Rfam_10.0_seed	1	130

Alignment produced by BLAST/Rfam_10.0_seed

Top sequence: RNA prediction 000199
 Bottom sequence: Rfam_10.0_seed U4|J302335.1/69797-69946_JRF00015
 E-value: 1e-38

```

000199          607456  atttttgcgttggggcaatgacgcacctagtgaggttaata-ccgaggcgcgtcaattgctggtt  607519
|| |||||
%U4|J302335.1/6  1  atctttgcgttggggcaatgacgcagctaatgaggttataaccgaggcgcgtctattgctggtt  65

000199          607520  gaaaactatttccaaactccctctttggc  607548
|||||
%U4|J302335.1/6  66  gaaaactatttccaaacccctcttaggc  94
      
```

Alignment produced by BLAST/Rfam_10.0_seed

Top sequence: RNA prediction 000199
 Bottom sequence: Rfam_10.0_seed U4|M479189.1/4511-4661_JRF00015
 E-value: 1e-31


```

000199          607456  atttttgcgttggggcaatgacgcacctagtgaggttaata-ccgaggcgcgtcaattgctggtt  607519
|| |||||
%U4|M479189.1/4  1  atctttgcgttggggcaatgacgcagctagtgaggttctaaccgaggcgcgtcaattgctggtt  65

000199          607520  gaaaactatttccaaactccctctttggc-----ctcacgggtcactgagaatttggtgcaaa  607577
|||||
%U4|M479189.1/4  66  gaaaactatttccaaacccctctttggcctgggttacgccaggccatcgagaatttctggaag  130

000199          607578  ggctccct  607585
|||||
%U4|M479189.1/4  131  ggctccct  138
      
```

T. aestivum:U4 ?




RNA features

ID: 000199
 Family: U4
 Sequence name: contig (*Wheat unknown - eukaryote - unknown*)
 Start: 607456
 End: 607586
 Strand: +
 Predicted by: *combine:BLAST/Rfam_10.0_seed* - with a 0.0 score on 25-4-2013

Genome context

GTATAAGAACGGACCATGGCGTATGGAAATGGGC ... [000199] ... TCCTTGGAGAGGGCAAGGGCCTACGAATTAAATAA

Sequence and structure(s)

000199	1	ATTTTTCGCTTGGGGCAATGACGCACCTAGTGAGGTAATACCGAGGCGCGTCAATTGCTGGTTGAAAACATTTCCAAACTCCCTCTTTGGCCCTCACG
	101	GGTCACTGAGAATTTGTGCAAAGGCTCCCTC

Alignment(s)

This prediction is included in 10 [alignment\(s\)](#)

[Edit](#) [Back to explore](#)

Comments and remarks: contact@rnaspace.org

T. aestivum:U4 ?

HELP

RNA features

ID:

Family:

Sequence name:

Start:

End:

Strand:

Predicted by: *INFERNAL 1.0.2 with a 2.83e-19 score on 25-4-2013*

Update preview

Genome context

GTATAAGAACGGACCATGGGCGTATGGAAATGGGC ... [000257] ... TACGAATTAAATAATCAAAATTTTAAATTTCTACT

Sequence and structure(s)

000257 1 ATTTTTCGCTTGGGGCAATGACGCACCTAGTGAGGTAAATACCGAGGCGCGTCAATTGCTGGTTGAAACTATTTCCAAACTCCCTCTTTGGCCCTCACG
 101 GGTCACTGAGAATTTGTGCAAAGGCTCCCTCTCCTTGAGAGGGCAAGGGCC

New secondary structure

You can type or paste a secondary structure in bracket-dot format.

Alternatively you can compute the minimal free energy secondary structure with:

Add this structure in the preview page

Reset initial values

Save

Cancel

T. aestivum:U4 ?

RNA features

ID:

000199

Family:

U4

Sequence name:

contig (Wheat unknown - eukaryote - unknown)

Start:

607356

End:

607586

Strand:

+

Predicted by:

combine:BLAST/Rfam_10.0_seed - with a 0.0 score on 25-4-2013

Update preview

Genome context

CGCTTATTACCGCGTTGTTGGAGATGCTCTTAA ... [000199] ... TCCTTGAGAGGGCAAGGGCCTACGAATTAATAA

Sequence and structure(s)

000199

1

TTTGT

TTGTT

GGTAGT

CTGATT

AGTCCC

ACCTCG

GTA

ACTGAG

GCAGG

TGGCA

AGGGG

GAGCT

AGGT

ATA

AGA

ACGG

ACCAT

GGG

CGT

ATG

GAA

ATG

GGC

101

ATTTT

GCGCT

TGGG

CAAT

GACG

CAC

CTAG

TGAG

GTAA

TACCG

AGG

CGCG

CAAT

TGCT

GTT

GAA

AACT

ATTT

CCAA

ACTC

CTCT

TTG

GCCT

CAC

G

201

GGT

CACT

GAG

AATT

TGT

GCA

AAG

GCT

CCCTC

New secondary structure

You can type or paste a secondary structure in bracket-dot format.

Alternatively you can compute the minimal free energy secondary structure with:

Select a software

Add this structure in the preview page

Alignment(s)

This prediction is included in 10 alignment(s)

Reset initial values

Save

Cancel

```

000257  1  TTTGTTTGT  GGTAGTCTGA  TTAGTCCCAC CTCGTAACT  GAGGCAGGTG
      51  GCAAGGGGGA  GCTAGGTATA AGAACGGACC  ATGGGCGTAT  GGAAATGGGC
    101  ATTTTGTGCG  TTGGGGCAAT  GACGCACCTA  GTGAGGTAAT  ACCGAGGCGC
    151  GTCAATTGCT  GGTTGAAAAC  TATTTCCAAA  CTCCTCTTT  GGCCCTCACG
    201  GGTCAGTGA  AATTGTGCA  AAGGCTCCCT  CTCCTTGAG  AGGGCAAGGG
    251  CC
  
```

sRNA prediction and annotation *T. aestivum*

Home 1.Load data 2.Predict 3.Explore ME

Current results for the 005739160f9a421 project: **179 putatives RNAs predicted.**

Software tools used and user actions are summarized in the left re-sizable table and query sequence(s) in the right re-sizable table. See the project [history](#) for more details.

Run or user identifier	Description	Number of RNAs
r05	INFERNAL	57
r07	INFERNAL	2
r06	INFERNAL	1
r04	INFERNAL	1
User action	Combine	-136+68
r03	BLAST/Rfam_10.0_seed	67
r02	Combine	-0+0
	INFERNAL	1
User action	Combine	-0+0
	Combine	-0+0
r01	Combine	-2+1
	BLAST/Rfam_10.0_seed	119
	RNAmer	0
	tRNAscan-SE	0

Query sequence(s)					
contig	983767 nt	eukaryote	Wheat	unknown	unknown
contig_914	2522860 nt	eukaryote	unknown	unknown	unknown

Field: Criterion Operator: Comparison Value (wildcards allowed): Give value Result: 179/179 RNAs satisfy filter(s) Add/Update

Table view JBrowse view CGview view

The table of results may be sorted by clicking on the column titles. You can select predictions by ticking the check boxes in the left column and perform actions on them using the down-drop lists below the table [?].

Predictions 1 - 20 of 179 Display Terse set Show 20 Page 1 of 9 >>

contig	seq name	start	end	score	+	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	1	r01	
<input type="checkbox"/>	000018	contig	MIR1122	21364	21441	78	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	1	r01
<input type="checkbox"/>	000261	contig	MIR1122	21364	21445	82	+	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000267	contig	MIR1122	21364	21445	82	-	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000002	contig	5S_rRNA	33257	33337	81	+	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	1	r01
<input type="checkbox"/>	000008	contig	5S_rRNA	42578	42643	66	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000019	contig	MIR1122	49455	49498	44	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000011	contig	MIR1122	49458	49498	41	+	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	3	r01
<input type="checkbox"/>	000009	contig	5S_rRNA	63214	63276	63	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000010	contig	5S_rRNA	126714	126785	72	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000258	contig	5S_rRNA	150056	150190	135	+	Wheat	eukaryote	unknown	INFERNAL	0	r06
<input type="checkbox"/>	000027	contig	5S_rRNA	150056	150190	135	+	Wheat	eukaryote	unknown	Combine: BLAST/Rfam_10.0_seed	5	r01
<input type="checkbox"/>	000268	contig	MIR1122	151269	151342	74	-	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000259	contig	MIR1122	221919	222033	115	+	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000265	contig	MIR1122	221919	222033	115	-	Wheat	eukaryote	unknown	INFERNAL	0	r05
<input type="checkbox"/>	000020	contig	MIR1122	221922	222032	111	-	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	2	r01
<input type="checkbox"/>	000012	contig	MIR1122	221923	222031	109	+	Wheat	eukaryote	unknown	BLAST/Rfam_10.0_seed	2	r01
<input type="checkbox"/>	000104	contig	MIR1122	308145	308146	82	-	Wheat	eukaryote	unknown	Combine: BLAST/Rfam_10.0_seed	1	r01

ID	SeqName	Family	Begin	End	Size	Strand	Software
000275	contig_914	MIR1122	540570	540684	115	+	INFERNAL
000274	contig_914	MIR1122	2231326	2231440	115	+	INFERNAL
000259	contig	MIR1122	221919	222033	115	+	INFERNAL
000273	contig_914	MIR1122	675545	675659	115	+	INFERNAL
000272	contig_914	MIR1122	648743	648857	115	+	INFERNAL
000279	contig_914	MIR1122	755563	755677	115	+	INFERNAL

P-value: 0.294198

[illegible]

米 米米米米 米米米 米 米 米米米米米米 米米米 米 米米米米 米米米米 米 米 米

000275 UAGAUCAUCCAUUUUUGUGACAGUAUACCGAACGGAGGGAGUACGUC
000274 UAGAUCAUCCAUUUCCGAGCAGUAUACCGAACGGAGGGAGUACUAC
000259 UAGAUCAUCCGUUUGAGCGGACAGUAUUUGGAUCGGAGGGAGUACUA
000273 UAGAUCAUCCAUUUCCAUAGCAGUAUUCCGGAACGGAGGGAGUAAAG
000272 UAGAUCAUCCAUAUGUGCGGACAGUAUACCGAACGGAGGGAGUACUAC
000279 UAGAUAGGUUCAUUUUUUGCAGCAGUAUACCGAACGAGGAGGUAAUA
consensus UAGAUCAUCCAUUUUUGCGGACAGUAUACCGAACGGAGGGAGUAAA

000275
000274
000259
000273
000272
000279
consensus

A UACU CCU G C AUUACU GUC A AUGU UCUA U U U

Sequence logo for the 5' region of the 16S rRNA gene. The y-axis lists positions 000275, 000274, 000259, 000273, 000272, 000279, and a consensus sequence. The x-axis shows nucleotide conservation across 10 positions. The consensus sequence is GGGGGGGGGG.

UAGUAU U C U U GACAAGUA U G CG AGGGAGUA

View consensus secondary structure with **rnaplot**

T. aestivum:miR1122 ?



miRBase

MANCHESTER 1824

[Home](#)
[Search](#)
[Browse](#)
[Help](#)
[Download](#)
[Blog](#)
[Submit](#)
[Search results](#)

Search Results

We found 3 unique results for your query ("mir1122"), in 4 sections of the database.

Section	Description	Number of hits
miRNA name	match the accession or ID of a hairpin precursor entry	3
Previous ID	match the previous ID of a hairpin precursor entry	0
Mature name	match the accession or ID of a mature miRNA sequence	3
Previous Mature ID	match the previous mature ID of a mature entry	0
Dead entry	match the accession or ID of a dead entry	0
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	3
Comments	search miRNA entry comments	1
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	Mature name	Description	Comments
MI0006184	tae-MIR1122	✓	✓	✓	✓
MI0011568	bdi-MIR1122	✓	✓	✓	
MI0016607	far-MIR1122	✓	✓	✓	

ncRNA prediction and annotation

Prediction is different of validation !!!

- **miRNA prediction**

- Homology search
 - Sequence alignment : very good conservation of the mature miRNA : Blastn against miRBase
- Be careful with plants !!!
- Pre-miRNA structure is to verify
- Take care of false positives

[illegible]

```
>[ref|NR_003287.2|] EGM Homo sapiens RNA, 28S ribosomal 1 (RN28S1), ribosomal RNA
Length=5076

GENE ID: 100008589_RN28S1 | RNA, 28S ribosomal 1 [Homo sapiens]
(10 or fewer PubMed links)

Score = 122 bits (66), Expect = 6e-28
Identities = 68/69 (99%), Gaps = 0/69 (0%)
Strand=Plus/Plus

Query 5      AGGTGAAGATCTTGGTGGTAGTAGCAAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGG 64
Sbjct 2341    AGGTGCAGATCTTGGTGGTAGTAGCAAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGG 2400

Query 65      AGAAGGGTT 73
Sbjct 2401    AGAAGGGTT 2409
```