



LABORATORY OF INTEGRATIVE BIOINFORMATICS

## Bioinformática de RNAs

Vinicio Maracaja-Coutinho, PhD



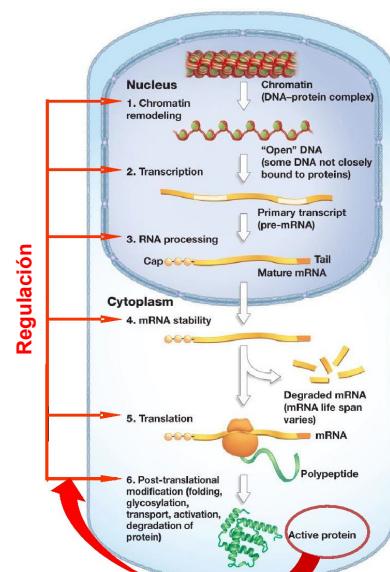
@vin\_maracaja



vinicius.maracaja@uchile.cl

**Los mecanismos genéticos fueron malos interpretados entre las décadas de 1950 y 2000**

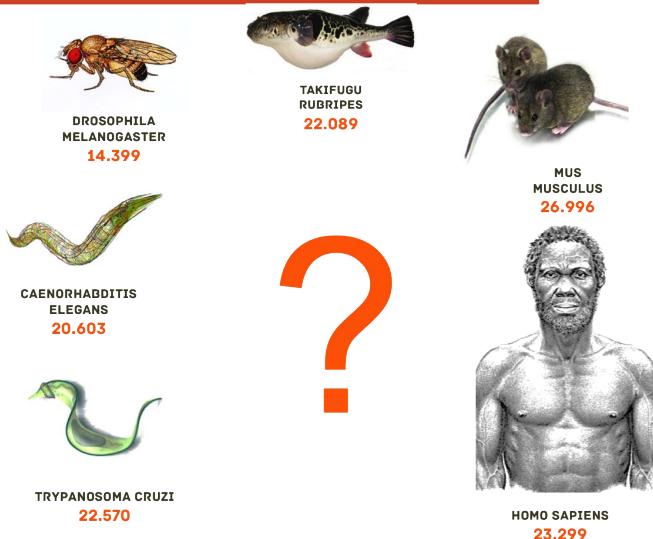
- **Proteínas** como el principal elemento regulador de la expresión génica y de los procesos celulares.



¿No somos tan distintos así?



Número de genes codificadores de proteínas súper similares entre organismos con complejidad tan distintas

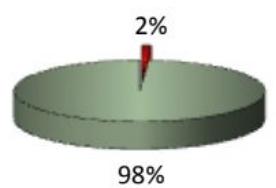


Genes RefSeq codificadores de proteínas no redundantes  
(NCBI, Júlio de 2012)

Solo 2% del genoma codifica proteínas



Donde vienen los elementos responsables por la regulación de los procesos celulares?

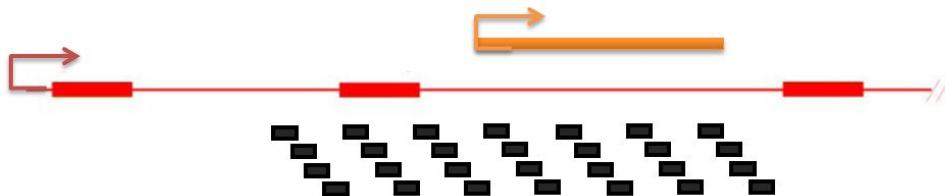


■ Protein-coding region  
■ Non-coding region

Tenemos una transcripción diseminada

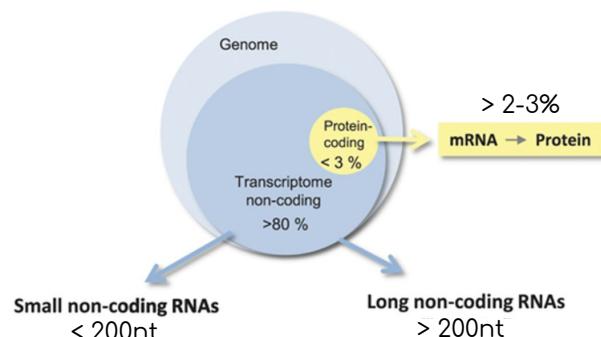
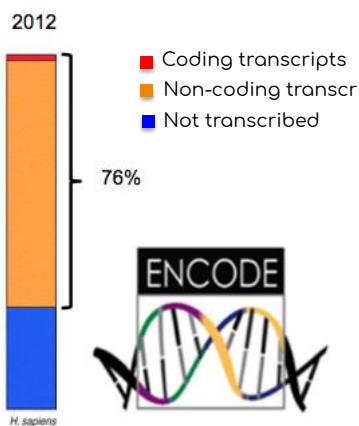


2002 Kapranov y col.: Tenemos 10 veces más regiones transcripcionalmente activas que el número predijo por el mapeo de los genes codificadores conocidos.



Tiling Arrays en todo el cromosoma 21 y 22

Al menos 76% del genoma humano es transcripto en algún tipo de RNAs



Djebali et al, 2012.  
Nature

Uchida and Dimmeler, 2015.  
Circulation Research

Más de 3 mil familias de RNAs



HOME | SEARCH | BROWSE | FTP | BLOG | HELP

Rfam 14.1 (January 2019, 3016 families)

The Rfam database is a collection of RNA families, each represented by **multiple sequence alignments**, **consensus secondary structures** and **covariance models (CMs)**. [More...](#)

Try the **new Rfam search** and [let us know](#) if you have any feedback

Examples: [SAM](#), [Homo sapiens](#), [snoRNA](#), [author: "Weinberg"](#)



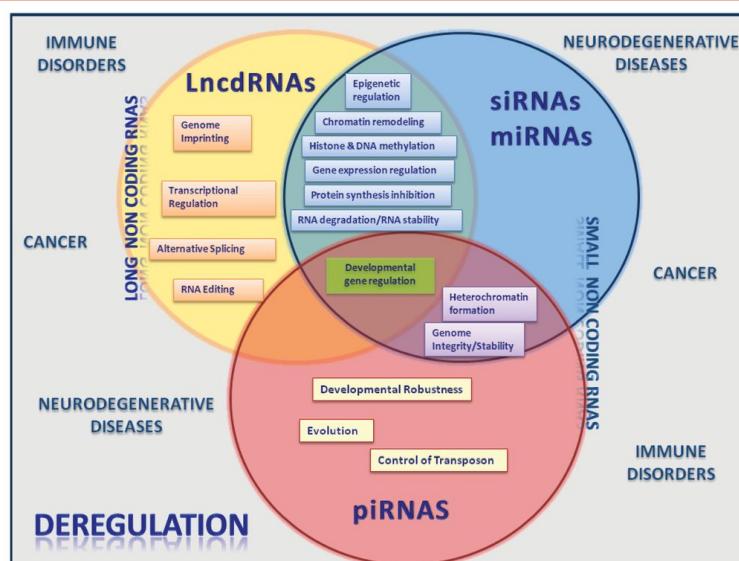
Search Rfam

Go

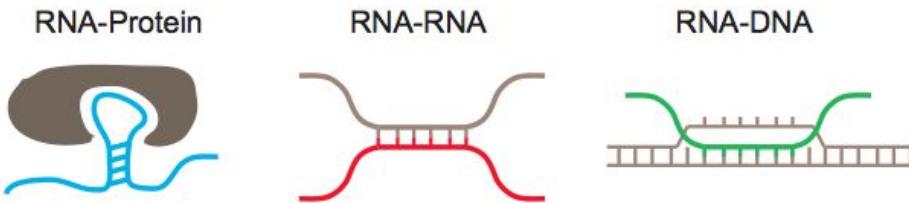
Browse [Families](#), [Clans](#), [Motifs](#), [New Genomes](#), or [Families with 3D structures](#)

Rfam  
database  
Enero 2019

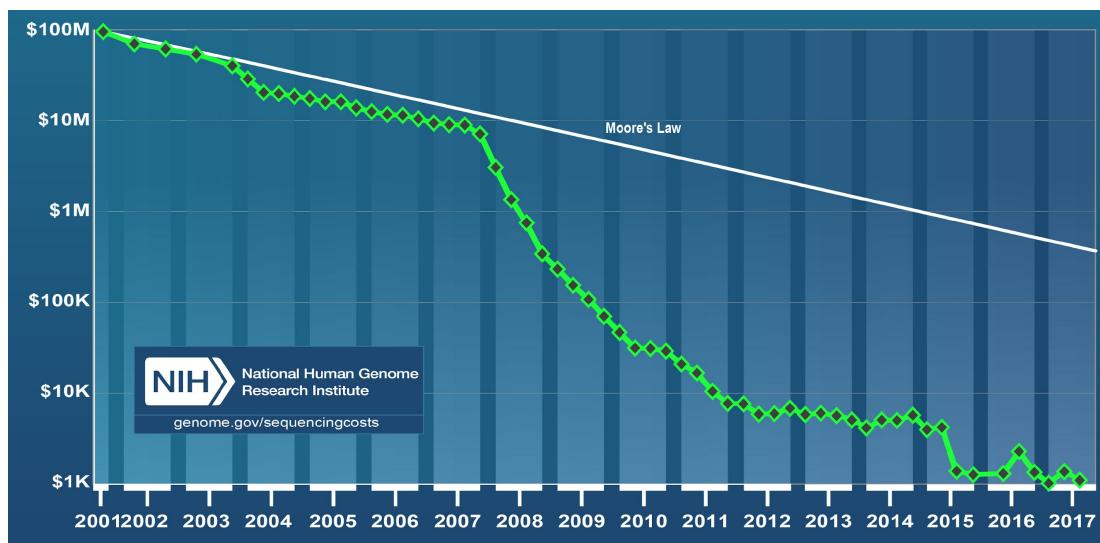
Participan regulando los más diversos procesos celulares y una  
disregulación puede llevar a diferentes enfermedades



Gomes et al., 2013.  
International Journal of Molecular Sciences

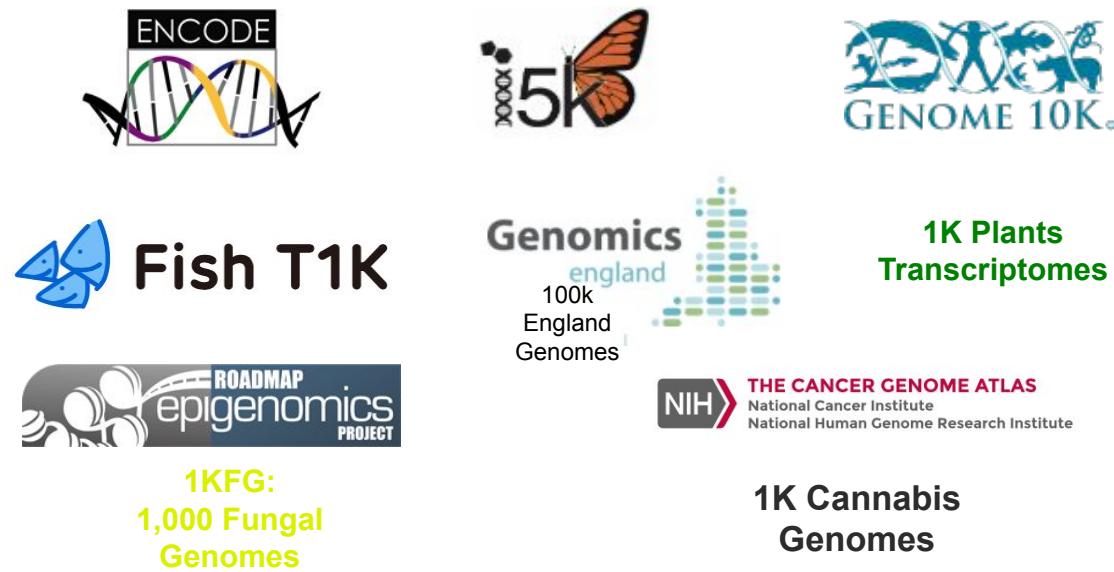


Los costos de la secuenciación han bajado más de 100.00 veces desde la publicación del primer borrador del genoma humano



<https://www.genome.gov/sequencingcostsdata/>

## Revolución “Big Data” en las ciencias de la vida



## Life Sciences “Big Data” Revolution



### Oportunidades!

- Organizar los datos de manera estandarizada;
- Desarrollo de bancos de datos, softwares y algoritmos computacionales para analisar y interpretar estos datos de manera eficiente
- Permitiendo la generación de nuevos *insights* biológicos relevantes

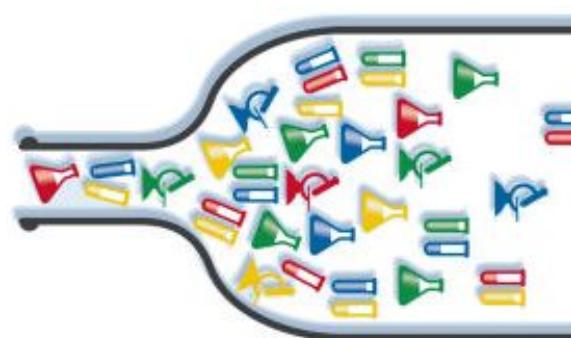
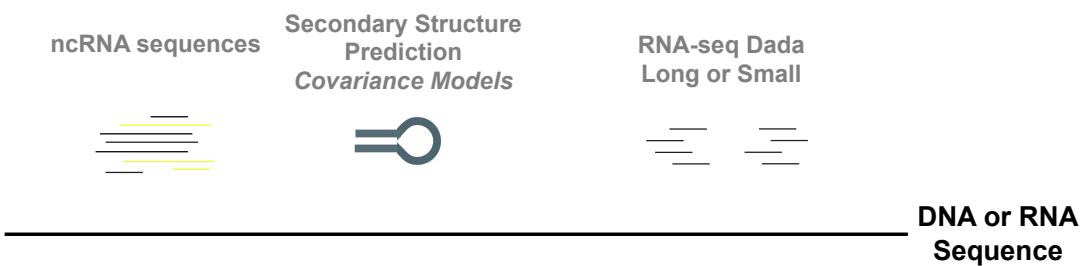


Image:  
<http://thebottleneck.com/>

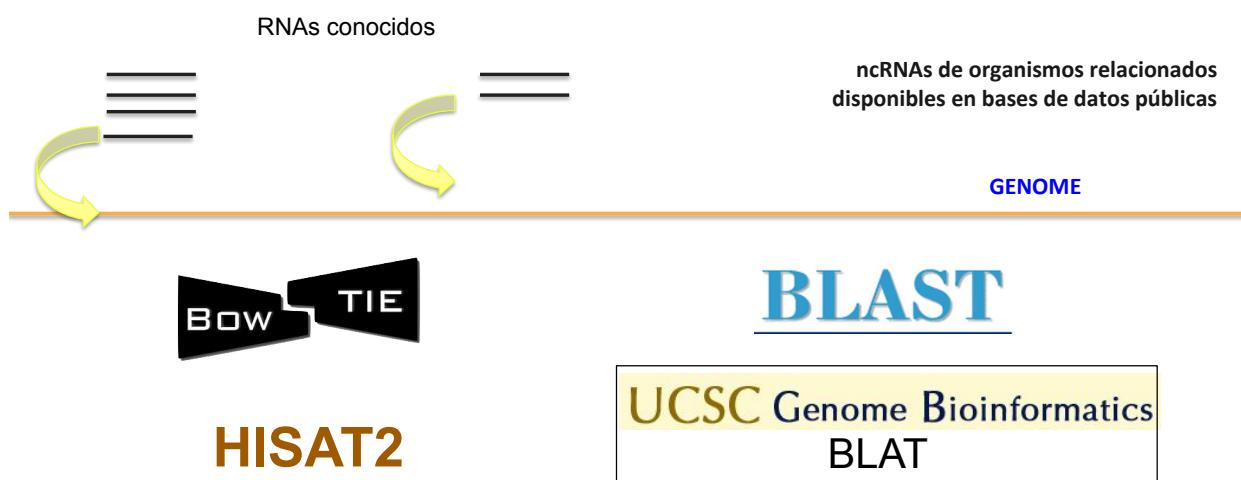
- Identificando familias de RNAs en secuencias de nucleótidos
  - Similitudes de secuencias
  - Modelos de Covariance de RNAs (Estructuras secundarias y/o secuencias)
  - RNA-seq
- Asignando función a RNAs
  - Similitudes de secuencias
  - Predictores específicos
  - Co-expresión de transcritos
- Ejemplos de otras caracterizaciones posibles
  - Región promotora (ej.: regulación por factores de transcripción)
  - Prediciendo interacción con proteínas, RNAs y DNA

Identificando familias de RNAs en  
secuencias de nucleótidos

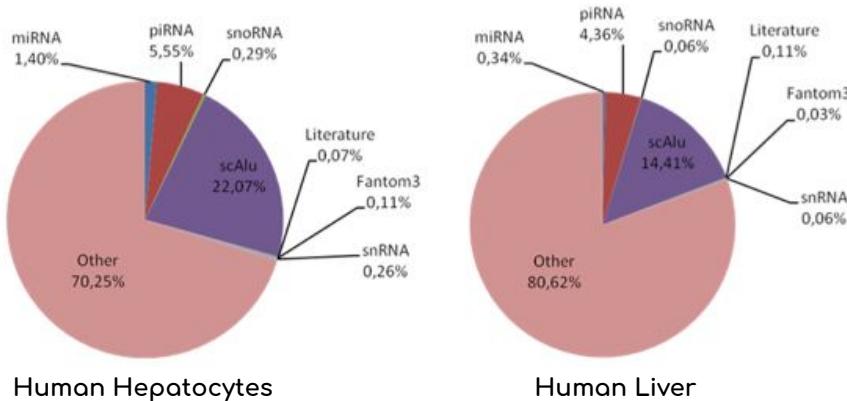
## Predicción de genes de RNAs



## Similitudes de Secuencias



2007 - Non-coding transcriptome was comprised by unclassified ncRNAs



We were not able to functionally annotate the repertoire of ncRNAs using sequence similarity searches or *ab initio* secondary structure predictive tools

## Similitudes de Secuencias





RNA Biology 9(3), 274–282; March 2012; © 2012 Landes Bioscience



## Non-coding transcription characterization and annotation

A guide and web resource for non-coding RNA databases

Alexandre Rossi Paschoal,<sup>1,3,5,†</sup> Vinícius Maracaja-Coutinho,<sup>2,5,†</sup> João Carlos Setubal,<sup>2</sup> Zilá Luz Paulino Simões,<sup>4</sup>  
Sergio Verjovski-Almeida<sup>2</sup> and Alan Mitchell Durham<sup>1,\*</sup>

<sup>1</sup>Departamento de Ciéncia da Computação; Instituto de Matemática e Estatística; Universidade de São Paulo; São Paulo, Brazil; <sup>2</sup>Departamento de Bioquímica; Instituto de Química; Universidade de São Paulo; São Paulo, Brazil; <sup>3</sup>Engenharia da Computação; Universidade Tecnológica Federal do Paraná – Campus Cornélio Procópio; Cornélio Procópio, Brazil; <sup>4</sup>Departamento de Biologia; Faculdade de Filosofia Ciéncias e Letras de Ribeirão Preto; Universidade de São Paulo; São Paulo, Brazil; <sup>5</sup>Programa Interunidades em Bioinformática; Instituto de Matemática e Estatística; Universidade de São Paulo; São Paulo, Brazil

\*Authors contributed equally to this work.

[www.ncrnadatabases.org](http://www.ncrnadatabases.org)



Exhaustive manual survey and curation of literature, introducing four categorizations to classify these databases and to help researchers quickly search and find the information they need:

### 1. RNA families

- *What types of ncRNAs are present in a database?*

### 2. Information source

- *What is the provenance of the ncRNA information? (from experimental evidence; from computational analysis only; from manual curation of information obtained experimentally and/or computationally; and from literature).*

### 3. Information content

- *What are the various types of information stored in a database (e.g., sequence, annotation, expression)?*

### 4. Search mechanisms

- *What search mechanisms are available in a database?*



NRDR portal



Non-coding RNA Databases Resource

Home About NRDR Search Browser Statistics Team NR2 Database Submit

Search in Databases Resource

[www.ncrnadatabases.org](http://www.ncrnadatabases.org)

**Database** i  
Any Database

**Overview (Description)** i  
Any Description

**Organism** i  
Any Organism

**Source** i  
Any Source

**Information Source** i  
 Experimental  In Silico Annotation  Literature  Manual Curation

**RNA Families** i  
 Multiple classes  
 CNE  
 Long RNA  
 miRNA

**Information Content** i  
 3'UTR Regions  
 Acceptor/Donor Site  
 Adenocarcinoma  
 Adult

**Download Type** i  
 BED  
 GFF  
 GTF  
 PSL

**Database has the following Search Method(s)** i  
 Density of ncRNAs  Genomic Location  Keyword  Similarity  Tabular  TAG

**Database Allows** i  
 Multiple Search  Graphic View

**Search**

**229 indexed databases**



Non-coding RNA Databases Resource

NRDR: The Non-coding RNA Databases Resource



Non-coding RNA Databases Resource

Home About NRDR Search Browser Statistics Team NR2 Database Submit

Search / 54 Databases Found

Database Name	Amount of Organisms	Address database website
BioM2MetDisease	Unspecified.	<a href="http://www.bio-bigdata.com/Bio...">http://www.bio-bigdata.com/Bio...</a>
CCGD	Homo sapiens.	<a href="http://crdd.osdd.net/raghava/c...">http://crdd.osdd.net/raghava/c...</a>
dbDEMC	Homo sapiens.	<a href="http://159.226.118.44/dbDEMC/">http://159.226.118.44/dbDEMC/</a>
doRTNA	Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans.	<a href="http://dorina.mdc-berlin.de">http://dorina.mdc-berlin.de</a>
EpimiRBase	Unspecified.	<a href="http://www.epimirbase.eu">http://www.epimirbase.eu</a>
ExprTargetDB	Homo sapiens.	<a href="http://www.scandb.org/apps/mic...">http://www.scandb.org/apps/mic...</a>
HOCTARdb	Homo sapiens.	<a href="http://hoctar.tigem.it/">http://hoctar.tigem.it/</a>
IGDB_NSCLC	Homo sapiens.	<a href="http://igdb.nsclc.ibms.sinica...">http://igdb.nsclc.ibms.sinica...</a>
Isomirs	Homo sapiens, Mus musculus.	<a href="http://hood.systemsbiology.net...">http://hood.systemsbiology.net...</a>
mESAdb	Homo sapiens, Mus musculus, Danio rerio.	<a href="http://konulab.fen.bilkent.edu...">http://konulab.fen.bilkent.edu...</a>



Non-coding RNA Databases Resource

## miRBase

RNA Type: miRNA

**Overview:** miRBase is one of the central repositories of microRNA, with information based on published experimental data and in-house annotation. The database is divided into three parts: (i) miRBase Registry, which provides a nomenclature for functional annotation of miRNAs; (ii) miRBase Targets, which contains predicted miRNA target genes; and (iii) miRBase Sequence, which contains published miRNA sequences and their annotation. The release 16 contains some dataset obtained from next generation sequencing.

### Search Methods:

- Similarity: search sequences using the BLAST or SSEARCH.
- Keyword: search by miRNA accession or name (general search); or by miRNA id, gene name, Ensembl identifier or GO term (target search).
- TAG: search by genome or GO class.
- Tabular: search by species.
- Genomic Location: choosing for an organism and chromosome location.
- Density of ncRNAs: for a cluster inter-miRNA in an organism according to different distances between each miRNA. There is also an option of genomic regions (e.g. CpG island, EST, cDNA, TFBS) surrounding miRNA precursors.

**Source:** Literature, Ensembl, UCSC Genome Browser.

**Information Source:** Experimental, In silico annotation.

**Information Content:** Annotation, Cluster, Evidence, Expression, Next-generation sequencing, Nomenclature of miRNA, Sequence, Structure, Target gene, Tissue.

**Reference:** Griffiths-Jones et al., 2008

**PubMedID:** 14681370, 16381832, 17991681, 20205188, 21037258.

**Year:** 2011;2010;2008;2006;2004

**Multiple search:** Yes

**Download:** GFF, EMBL, FASTA, Other.

**Genomic Overview:** No

**Organism:** *Homo sapiens* (Human), *Rattus norvegicus* (Norway Rat), *Canis familiaris* (Dog), *Mus musculus* (House Mouse), *Danio rerio* (Zebrafish), *Takifugu rubripes* (Fugu), *Gallus gallus* (Chicken), *Ciona intestinalis*, *Drosophila melanogaster* (Common Fruit Fly), *Caenorhabditis elegans*, *Arabidopsis Thaliana* (Thale Cress), *Macaca mulatta* (Rhesus Monkey), *Bos Taurus* (Cattle), *Ovis aries* (Sheep), *Sus scrofa* (Wild Boar) and 142 others.

**Url:** <http://www.mirbase.org/>



## NRDR: The Non-coding RNA Databases Resource



January, 2019

## Chapter 10

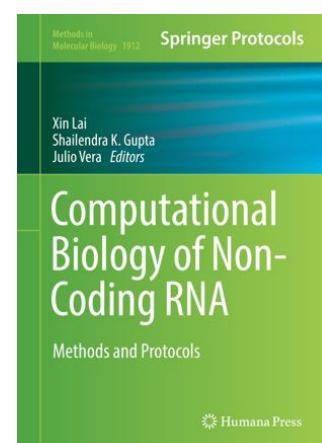
### Noncoding RNAs Databases: Current Status and Trends

Vinicius Maracaja-Coutinho, Alexandre Rossi Paschoal,  
José Carlos Caris-Maldonado, Pedro Vinícius Borges, Almir José Ferreira,  
and Alan Mitchell Durham

#### Abstract

One of the most important resources for researchers of noncoding RNAs is the information available in public databases spread over the internet. However, the effective exploration of this data can represent a daunting task, given the large amount of databases available and the variety of stored data. This chapter describes a classification of databases based on information source, type of RNA, source organisms, data formats, and the mechanisms for information retrieval, detailing the relevance of each of these classifications and its usability by researchers. This classification is used to update a 2012 review, indexing now more than 229 public databases. This review will include an assessment of the new trends for ncRNA research based on the information that is being offered by the databases. Additionally, we will expand the previous analysis focusing on the usability and application of these databases in pathogen and disease research. Finally, this chapter will analyze how currently available database schemas can help the development of new and improved web resources.

**Key words:** Bioinformatics, Databases, Noncoding RNAs, Biomedicine, Biomedical, Disease, Micro-RNA, lncRNA, Circulating RNAs, Review



OMICtools

LIB

RNA databases

SIGN IN or SIGN UP

SOFTWARE 99+ DATABASES 99+ USERS 99+ PIPELINES BETA

1 - 20 of 1244 results

Category

- Gene expression databases 394
- miRNA databases 164
- Genomic databases 132

See more ▾

Taxonomy

- Homo sapiens 340
- Mus musculus 165
- Drosophila melanogaster 53

See more ▾

Disease

- Cancer diseases 148
- Reproductive diseases 112
- Neuronal diseases 105

See more ▾

Data Access

NRDR / Non-coding RNA Databases Resource

Provides a collection of currently available public databases with non-coding RNA information. NRDR permits users to search a list of databases filtered by

FlyBase

Includes several types of information on Drosophila genes and genomes. FlyBase is an online database that curates a variety of data from published biological literature.

GenBank

Provides publicly available nucleotide sequences for formally described species. GenBank is a comprehensive public database of nucleotide sequences.

**BETA VERSION**

OUR PIPELINE EDITOR IS UNDER CONSTRUCTION

We're working very hard to give you the best experience with this one.

RELATED SOFTWARE

- BLASTX
- PSI-BLAST
- GuideScan

See all

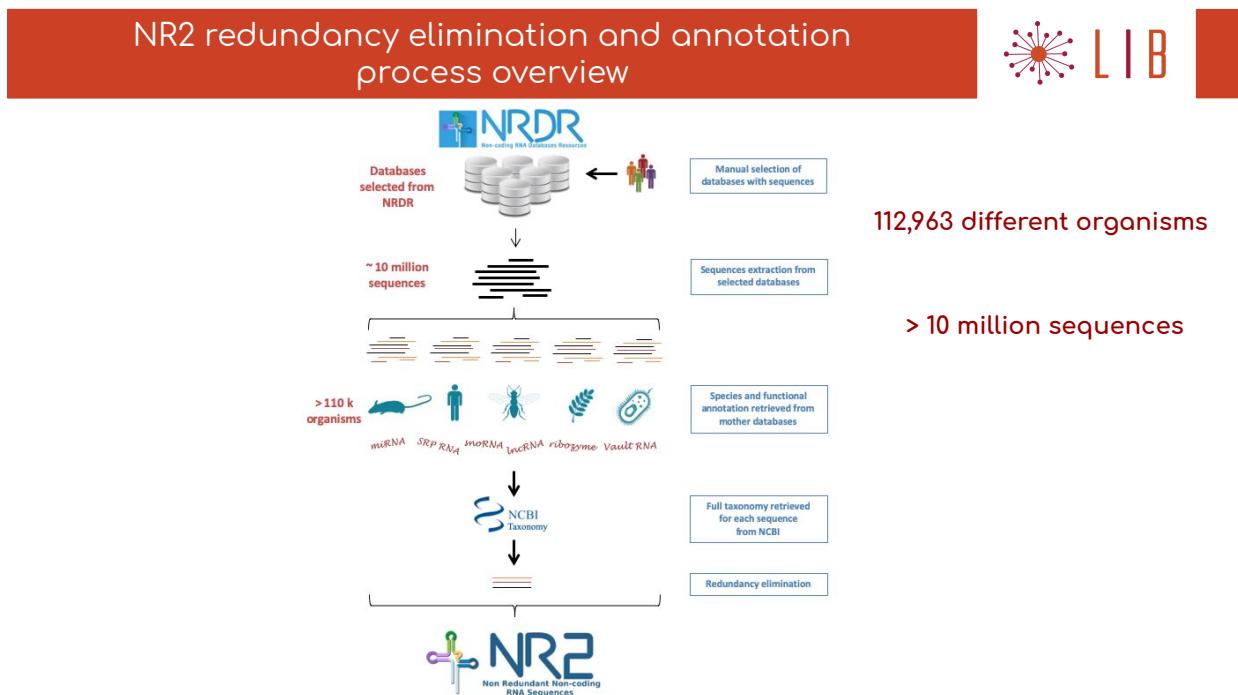
RELATED DATABASES

- NRDR
- FlyBase
- GenBank

See all

RELATED USERS

Chris M. Brown  
University of Otago



## NR2 portal (alfa version available)



NR2

Home About Search Statistics NRDR Database

A Repository for Non-Redundant Non-Coding RNA Sequences

Search a taxon

Search a taxon name

Filter by classes

Select All Clear

Gene tRNA Unclassified Cis-reg Intron piRNA miRNA splicing rRNA ribozyme sRNA lncRNA miRNA riboswitch  
CD-box Intron Ribozyme HACA-box siRNA scaRNA IRES snRNA rRNA RNase P IRES sRNA Riboswitch snoRNA  
antitoxin leader frameshift\_element antisense Telomerase RNA SRP RNA CRISPR thermoregulator snRNA transcript Y RNA  
tmRNA gRNA Ribozyme siRNA/RNAi ta-siRNA scRNA lncRNA rRNA lncRNA scaRNA Vault RNA 7SK leader

## Search for specific taxon using keywords



NR2

Home About Search Statistics NRDR Database

A Repository for Non-Redundant Non-Coding RNA Sequences

Search a taxon

Search a taxon name

Keywords: drosophila

- Drosophila
- Hawaiian Drosophila
- Drosophila grimshawi
- Drosophila melanogaster
- Drosophila simulans
- Drosophila willistoni
- Drosophila eugracilis
- Drosophila kikkawai
- Drosophila ficusphila
- Drosophila elegans

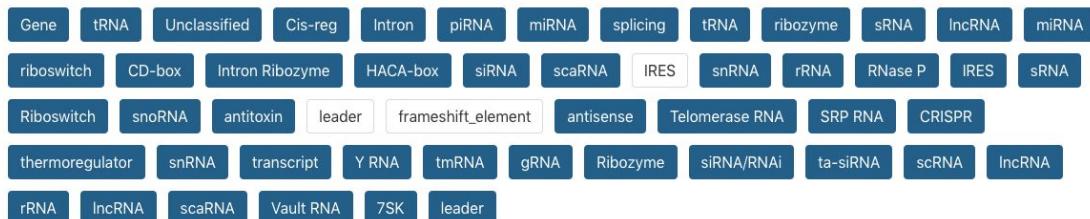
Filter by specific RNA classes of interest



NR2 Home About Search Statistics NRDR Database A Repository for Non-Redundant Non-Coding RNA Sequences

Filter by classes

Select All Clear



Navigate through taxonomy in order to download non-redundant sequences or obtain additional information from the database



NR2 Home About Search Statistics NRDR Database A Repository for Non-Redundant Non-Coding RNA Sequences

Download sequences

Reset

Organisms(2,756,351 sequences)

Find in the list...

		Archaea (9,056)
		Bacteria (242,977)
		Eukaryota (2,285,655)
		other sequences (117)
		unclassified sequences (108,397)
		Viroids (734)
		Viruses (149,857)

Eukaryota(2,285,655 sequences)

Find in the list...

		Alveolata (7,763)
		Amoebozoa (2,171)
		Apusozoa (49)
		Cryptophyta (392)
		environmental samples (8,323)
		Euglenozoa (10,011)
		Fornicata (214)

Euglenozoa(10,011 sequences)

Find in the list...

		Diplonemida (5)
		Euglenida (216)
		Kinetoplastida (9,790)

Navigate through taxonomy in order to download non-redundant sequences or obtain additional information from the database



A screenshot of a web-based taxonomy navigation interface. At the top left, it says "Organisms(2,756,351 sequences)". Below is a search bar with placeholder "Type a text Search". A sidebar on the left lists major taxonomic groups: Archaea (9,056), Bacteria (242,977), Eukaryota (2,285,655), other sequences (117), unclassified sequences (108,397), Viroids (734), and Viruses (149,857). The main content area shows a list titled "Euglenida (216 sequences)" with 14 entries. Each entry includes a sequence ID, type, and organism name. The first few entries are: 107927 CD-Box Euglena\_gracilis, 119365 TRNA Euglena\_stellata, 137710 RRNA Lepocinclis\_spirogyroides, 153456 CD-Box Euglena\_gracilis, 218427 RRNA Euglena\_longa, 221336 RRNA Euglena\_stellata, 259463 CD-Box Euglena\_gracilis, 265133 2 Classes Entosiphon\_sulcatum, 285068 TRNA Euglena\_longa, 285918 TRNA Euglena\_gracilis, and 308454 TRNA Euglena\_gracilis. A "Close" button is at the bottom right of the list.

## NR2 generated information



A screenshot of a web-based interface for NR2 generated information. At the top left, it says "Download sequences". Below is a search bar with placeholder "Type a text Search". A sidebar on the left lists the same taxonomic groups as the previous interface. The main content area shows a list titled "Euglenida (216 sequences)" with a blue header bar labeled "NR2 Database Information". It displays detailed information about a specific sequence: NR2 ID: 869470, Length: 73, Number copies: 1, Number Unclassified Copies: 0, Distinct Classes: [ "Intron Ribozyme" ], Distinct Databases: [ "tRNAb" ], and Distinct Organisms: [ "Euglena\_gracilis" ]. Below this is a "Sequence:" section with a color-coded sequence string: G T A A G T C G T G T G C A T T T G A A A A A T G C C A T G C A C C G A T T T T T T T A G G G A A T C A C A T T A G T G T A T T T T C T A C T A C G G A A. A "Close" button is at the bottom left of the sequence panel.

## Connection with RNA Central API



A screenshot of the LIB software interface. A central modal window displays search results for "Apicomplexa (3,837 Sequences)". The window has a green header bar labeled "RNA Central Search". Below the header, there is a small icon of a multi-colored flower or star-like shape. The main content area contains the following information:

- ID: URS0000664389
- Description: snoRNA from 9 species
- Publications: <http://rnacentral.org/api/v1/rna/URS0000664389/publications>
- Count Distinct Organisms: 9
- Distinct Databases: [ "Rfam" ]

At the bottom of the modal is a "Close" button. The background of the interface shows a dark sidebar on the left with icons for "Organisms", "Search", and "Find", and a vertical list of numbers on the right ranging from 37 down to 1.

## Connection with Rfam API



A screenshot of the LIB software interface. A central modal window displays search results for "Apicomplexa (3,837 Sequences)". The window has a black header bar labeled "Rfam Search". Below the header, there is a small Rfam logo icon. The main content area contains the following information:

- Rfam Request Opened: 2019-04-17 06:22:38
- Rfam Job ID: D093C578-60D0-11E9-B53C-805FD1B96DDE
- Estimated Time: 3s
- Result URL: <http://rfam.org/search/sequence/D093C578-60D0-11E9-B53C-805FD1B96DDE>

At the bottom of the modal, there is a navigation bar with arrows and the text "2580 TRNA From Babesia\_bovis". A "Close" button is at the bottom right. The background of the interface shows a dark sidebar on the left with icons for "Organisms", "Search", and "Find", and a vertical list of numbers on the right ranging from 37 down to 1.

Identify redundancies and improve RNA annotations



Peach Latent Mosaic Viroid (281 sequences)

NR2 Database Information

NR2 ID: 13358  
Length: 53  
Number copies: 17

Identify redundancies in databases

Annotate unclassified RNAs

Distinct Classes: ["Ribozyme", "ribozyme", "Unclassified"]  
Distinct Databases: ["tRNADB", "Rfam", "NONCODE"]  
Distinct Organisms: ["Peach latent mosaic viroid"]

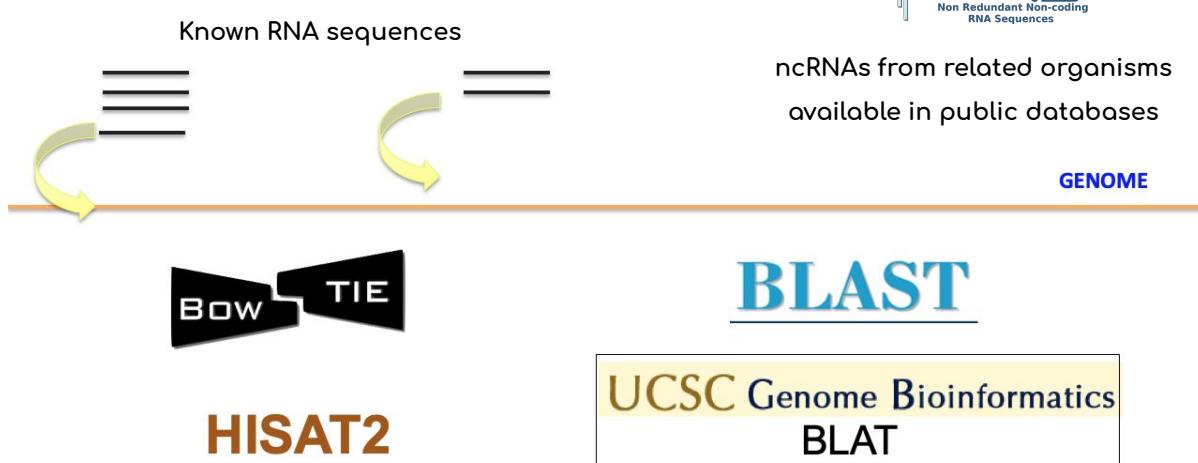
Sequence:

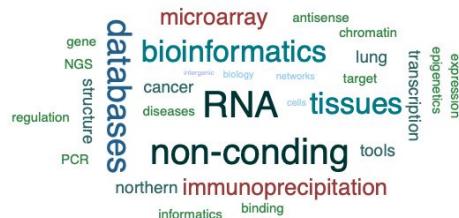
G A A G A G T C G T G C T T A G C A C A C T  
G A T G A G T C T C T G A A A T G A G A C G  
A A A C T C T T T

Close

UNIVERSIDAD DE CHILE

Useful for non-coding RNAs Identification based on sequence similarity searches

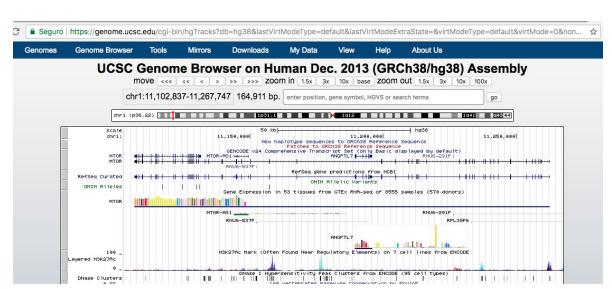
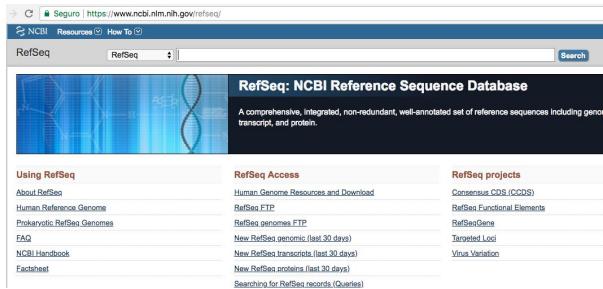
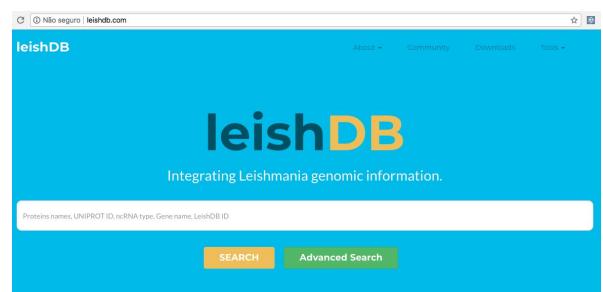


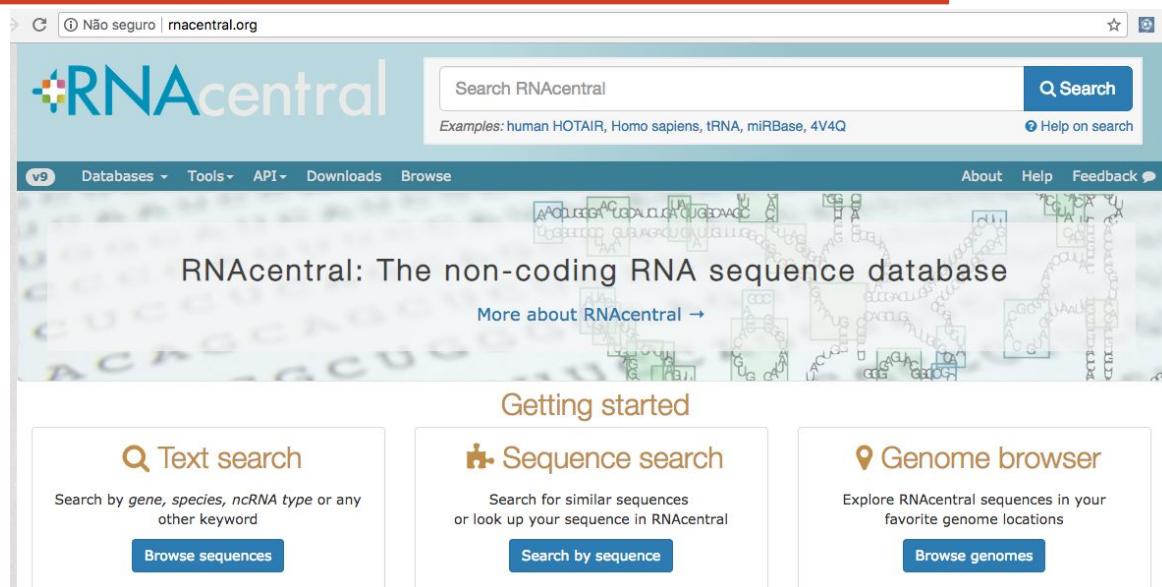


Search and locate information about which ncRNA internet databases contain information pertinent to your research.



Search and download non redundant ncRNA datasets according to its classes and taxonomic classification from our database of more than 2,000,000 non redundant sequences.





Não seguro | rnacentral.org

Search RNACentral

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

Q Search

Help on search

Databases Tools API Downloads Browse About Help Feedback

v9

RNACentral: The non-coding RNA sequence database

More about RNACentral →

Getting started

Text search

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

Browse sequences

Sequence search

Search for similar sequences or look up your sequence in RNACentral

Search by sequence

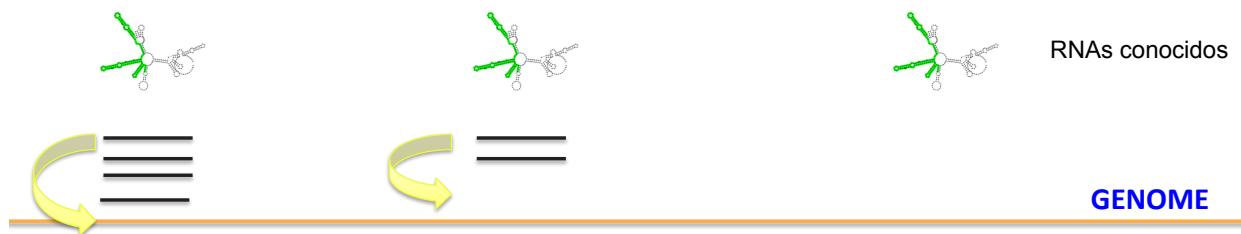
Genome browser

Explore RNACentral sequences in your favorite genome locations

Browse genomes

Identificando RNAs a partir de predicción  
de estructuras secundárias

## Identificando RNAs a partir de predicción de estructuras secundárias



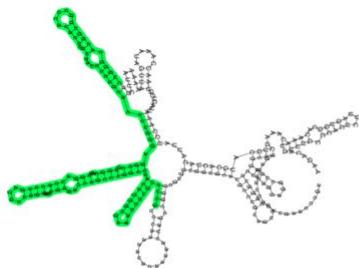
Algunas veces las secuencias primarias no son conservadas,  
sin embargo sus estructuras secundarias si son conservadas

## Identificando RNAs a partir de predicción de estructuras secundárias



Algunas veces las secuencias primarias no son conservadas,  
sin embargo sus estructuras secundarias si son conservadas

## Identificando RNAs a partir de predicción de estructuras secundárias



((((.....)))).....((((.....))))))
ACGC      GCGT
X            X
TGTT      ACCA
X            X
CCTA      TAGG

```
>Ec_OxyS,-1-109
aacgcggaccccgccuuuaacccuugaagucacugccguuucgagaguuccaacaucuaagccaacgugaacuuuugcggaucuccaggauccgcu
...(((((((.....(((((.....))))....)))))).((((.....)))).....(((((((.....))))))). (-29.40)
>Ec_RhyB,-7-66
aaccugaaagcaccacauugcucacaugcuuccagauuaacuuagccagccggugcuggcuuu
...(((.....((.....))))....(((((.....)))))... (-17.40)
>Ec_RprA,-1-106
cgguuauuaaucaacauauugauuuuuuaggcauggaaaucccugagugaaacaacgaaauugcuguguguagucuuugcccaucuccacgauggcuuuuuu
.(.....((.....))))....((.....)).....(((((.....))))....(((((.....))))....)).... (-24.80)
<Ec_mirE -1 82
```

## Identificando RNAs a partir de predicción de estructuras secundárias



### Lowe Lab tRNAscan-SE Search Server

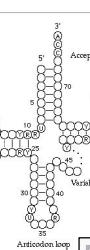
Search for tRNA genes in genomic sequence



#### tRNAscan-SE 1.21

**Plea for letters of support:** Our lab is planning major improvements and upgrades to the capabilities of tRNAscan-SE over the next 3-5 years. If you use this program and/or web server and would like to help us plan a letter of support to lower government officials describing briefly how tRNAscan-SE has helped with your research, and voicing your support for future improvements (if you have specific feature requests, please include in your letter or email). tRNAscan-SE and this web server have been freely available for over ten years, and taking a few minutes to show your appreciation (and help us pursue future funding) is deeply appreciated. -TML

This web server is described in  
Schattner, P., Brooks, A.N., and Lowe, T.M. (2005) Nucleic Acids Res. 33: W686-689.  
The principles underlying the tRNAscan-SE program are described in  
Lowe, T.M. and Eddy, S.R. (1997) Nucleic Acids Res. 25: 955-964.  
If you use this tool in your investigations, please cite one of these references.



#### tRNAs

**CENTER FOR BIOLOGICAL COMPUTATION**

**STAFF**

**CONTACT**

**ABOUT CBS**

**INTERNAL**

**CBS BIOINFORMATICS TOOLS**

**CBS COURSES**

**OTHER BIOINFORMATICS LINKS**

**EVENTS**

**NEWS**

**RESEARCH GROUPS**

**CBS PREDICTION SERVERS**

**CBS DATA SETS**

**PUBLICATIONS**

**EDUCATION**

[CBS >> CBS Prediction Servers >> RNAmmer](#)

**RNAmmer 1.2 Server**

The RNAmmer 1.2 server predicts 58S, 16S/18S, and 23S/28S ribosomal RNA in full genome sequences. This page is the entry point to the Prediction Server for RNAmmer. RNAmmer is available also as a Web Service described by the following [WSDL file](#). Please read the instructions on the [RNAmmer Web Services](#) section.

This page allows academic users to [download RNAmmer](#).

**Download data**

RNAmmer is run daily on the genbank sequences of the [NCBI Entrez Genome Projects](#). MD5 checksums of the raw genome sequence are used to keep track of changes in the genome. From the links below, these data may be downloaded. Please cite [Lagesen et al. 2007](#) when using these results.

#### rRNAs





Arias-Carrasco et al. BMC Bioinformatics (2018) 19:55  
https://doi.org/10.1186/s12859-018-2052-2

BMC Bioinformatics

SOFTWARE

Open Access



# StructRNAfinder: an automated pipeline and web server for RNA families prediction

Raúl Arias-Carrasco<sup>1,2†</sup>, Yessenia Vásquez-Morán<sup>1†</sup>, Helder I. Nakaya<sup>3\*</sup> and Vinicius Maracaja-Coutinho<sup>1,4,5,6\*</sup>

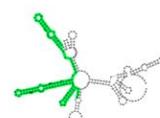
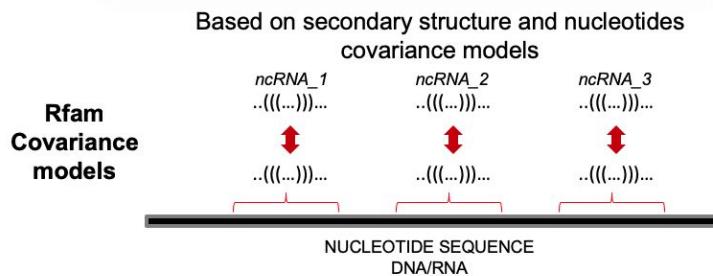


Arias-Carrasco et al, 2018  
BMC Bioinformatics

StructRNAfinder: an automated pipeline and web server for RNA families prediction



It is a pipeline that integrates various tools widely used for the identification of ncRNAs in an automated manner.



<http://structrnafinder.integrativebioinformatics.me>

Arias-Carrasco et al, 2018  
BMC Bioinformatics

Integrates functional annotation of Rfam covariance models to predict and functionally classify ncRNAs in nucleotide sequences



**General description of the RNA families predicted by StructRNAfinder according to Rfam nomenclature. For more details related to each family, please access the Rfam database**

Rfam class	List of available families
Cis-reg	frameshift_element; IRES; leader; riboswitch; thermoregulator It also can be represented by an empty “family,” which are other <i>cis</i> regulatory RNAs that do not have a specific classification
Gene	Antisense; antitoxin; CRISPR; lncRNA; miRNA; ribozyme; rRNA; snRNA; snoRNA; sRNA; tRNA It also can be represented by an empty “family,” which are other RNA genes that do not have a specific classification
Intron	Here, Rfam groups all intronic RNAs responsible for catalyzing their own removal from host transcripts (mRNA, tRNA, and rRNA precursors) in a wide range organisms



StructRNAfinder is available in an stand-alone version



**List of commands that can be used in the stand-alone version of StructRNAfinder**

Command option	Command description
-i or --input	The input file containing the sequences to be performed the search
-d or --database	The file containing the covariance models database. It can be the reference Rfam models automatically downloaded when installing the tool or a set of models generated by the user
-m or --method	Infernal method to search for RNAs families in a sequence dataset, cmscan or cmsearch [default: cmscan]
-n or --otherDB	It is necessary to choose this option when using other covariance models as a database than Rfam. Here, the user needs to indicate the file containing the covariance models to be used in the comparisons
-r or --report	Report all significant annotated RNA hits. One should use this option if interested in performing a genome-wide prediction of RNAs families. By default only the best hit per sequence is shown [default: False]
-s or --score	Filter a minimum infernal score per hit [default: 10]
-e or --e-value	Filter a maximum infernal e-value to each hit [default: 0.01]
-o or --output	Name of the output files generated. By default StructRNAfinder keeps the same name of the files from the INPUT sequences (INPUT_FILE), changing only the extension according to each FILE generated
-c or --cpu	Number of parallel CPUs to be used for a multi-thread analysis

StructRNAfinder -i archivo.fasta -d cm



## StructRNAfinder: Identificando familias de RNAs en secuencias de nucleótidos usando modelos de covariancia



StructRNAfinder -i archivo.fasta -d cm



**-i, --input** archivo en formato fasta  
**-d, --database** input reference database  
**-x, --otherDB** define only if you want use a reference database different to Rfam [default: false]  
**-m, --method** method to search for structural RNAs in a sequence dataset, cmsearch or cmscan [default: cmscan]  
**-r, --report** report total number of annotated RNAs, for default only is showed the best hit per sequence [default: False]  
**-t, --tblout** save parseable table of hits to file [default: nameinfile.tab]  
**-o, --output** direct output to file <f>, not stdout [default: nameinfile.out]  
**-c, --cpu** number of parallel CPU workers to use for multithreads [default: 1 cpu]  
**-p, --posStrand** only search the positive strand [default: false]  
**-n, --negStrand** only search the negative strand [default: false]

Options controlling inclusion (significance) thresholds (only one):

**-s, --score** minimum score to each hit [default: 10]  
**-e, --e-value** maximum e-value to each hit [default: 0.01]

Also available in a user-friendly web server



### JOB LAUNCHER:

Here, you can launch an analysis using StructRNAfinder and access the results in our servers. Before running an analysis, please read the following information:

- Your results will be available until **48 hrs** after its creation, then will be **deleted**.
- All **reports pages** can be **downloaded** in a compressed file (**tar.gz format**) available in the "Files" section of your results html page.
- If you want to analyze a sequence(s) **longer than 10Mbp**, you have to **use the stand-alone verion** available in the [DOWNLOAD](#) section.

You can see an input example [here](#)

Fasta file  Seleccionar Archivo | nenhum arquivo selecionado

Method  cmsearch

Filtering reporting thresholds

E-value  0.01

Search in:  Both strand.

Report only best hit per sequence

RUN! Note: Maximum sequence length 10 Mbp (including all sequences)

StructRNAfinder examples:

*E. coli* genome: ([input](#)|[output](#))  
*E. coli* known noncoding RNAs\*: ([input](#)|[output](#))  
*Leishmania braziliensis* genome: ([input](#)|[output](#))

\*: extracted from Sætrom et. al, 2005.





		Infernal						RNAfold	
Sequence	RNA family	Id	From_seq	To_seq	Score	Evalue	Score	Struct	
Intron									
<a href="#">gi_545778205_gb_U000_71</a>	group-II-D1D4-4	RF02003	4501332	4501193	67.8	9e-14	-42.40		
<a href="#">gi_545778205_gb_U000_72</a>	group-II-D1D4-6	RF02005	4501323	4501067	19.9	0.0017	-84.10		
IRES									
<a href="#">gi_545778205_gb_U000_77</a>	IRES_Bip	RF00223	1472215	1472295	23.6	9.6e-05	-81.20		
<a href="#">gi_545778205_gb_U000_78</a>	IRES_c-myc	RF00216	655274	655419	30.3	2.8e-06	-88.60		
Others-cis									
<a href="#">gi_545778205_gb_U000_140</a>	mini-ykkC	RF01068	4376811	4376857	50.3	2.5e-08	-16.60		
<a href="#">gi_545778205_gb_U000_154</a>	nuoG	RF01748	2400215	2400173	45.5	7.8e-08	-9.30		
<a href="#">gi_545778205_gb_U000_234</a>	SECIS_2	RF01988	1547981	1548035	28.8	0.0025	-14.20		



**Summary RNAs annotated**

**Struct RNA finder**

**General summary**

Number of sequences	1
Number of annotated sequences	1
Total of hits number	488
Average number of hits per sequence	488

**Summary by family**

cis_reg;frameshift	1
cis_reg;IRES	2
cis_reg;leader	32
cis_reg;riboswitch	8
cis_reg;thermoregulator	8
cis_reg;other	37
gene;antisense	10
gene;antitoxin	28
gene;CRISPR	1
gene;lncRNA	2
gene;miRNA	9
gene;ribozyme	4
gene;rRNA	71
gene;snoRNA	7
gene;sRNA	106
gene;tRNA	121
gene;other	39
Intron	2

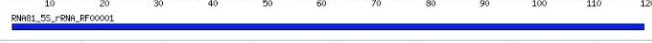
**Percentages of hits per family**

**Loci distribution overview**

**Struct RNA finder**

**Loci distribution overview**

**Sequence**

Sequence	Image
asnT	
gltU	
RyeE	
rrsB	
rrfC	
glyX	
rriA	

**Positive strand**

**Negative strand**

**Struct RNA finder**

Different file formats useful for downstream analyses



Files:	
Infernal	<a href="#">new_e_coli.out</a> <a href="#">new_e_coli.tab</a>
StructRNAfinder Annotation	<a href="#">new_e_coli_filtered.tab</a>
Bed format	<a href="#">new_e_coli_filtered.bed</a>
FASTA Sequence	<a href="#">new_e_coli_mature.fasta</a>
RNAfold	<a href="#">new_e_coli_RNAfold.dbn</a>
All Reports	<a href="#">Download web page</a>



StructRNAfinder: Identificando familias de RNAs en secuencias de nucleótidos usando modelos de covariancia





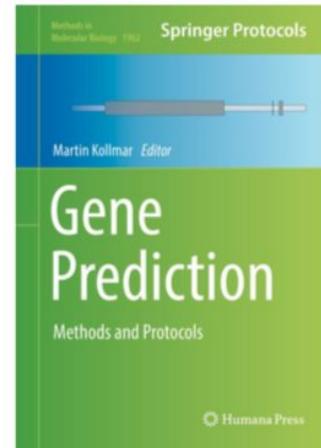
## Chapter 2

### Predicting RNA Families in Nucleotide Sequences Using StructRNAfinder

Vinicius Maracaja-Coutinho, Raúl Arias-Carrasco, Helder I. Nakaya,  
and Victor Aliaga-Tobar

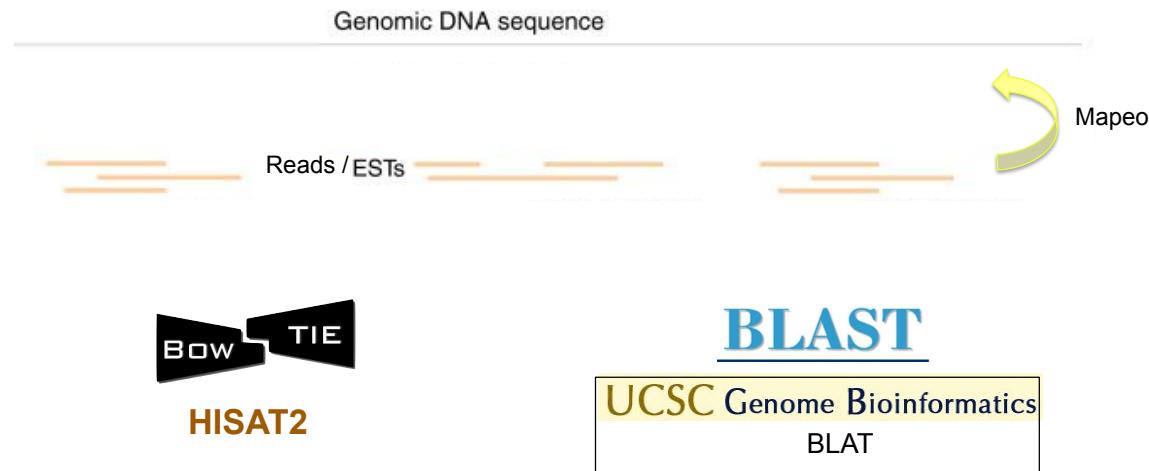
#### Abstract

Noncoding RNA (ncRNA) research is already a routine in every genomics or transcriptomics initiatives. According to their functions, ncRNAs can be grouped into several different RNA families, which can be represented by conserved primary sequences, secondary structures, or covariance models (CMs). CMs are very sensitive in predicting RNA families in nucleotide sequences and have been widely used in characterizing the repertoire of ncRNAs in organisms from all domains of life. However, the large-scale prediction and annotation of ncRNAs require multiple tools along the process, imposing a great obstacle for researchers with lesser computational or bioinformatics background. StructRNAfinder emerged as an automated tool to avoid these bottlenecks, by performing the automatic identification and complete annotation of regulatory RNA families derived directly from nucleotide sequences. In this chapter, we provide a complete tutorial for both stand-alone and web server versions of StructRNAfinder. This will help users to install the tool and to perform predictions of RNA families in any genome or transcriptome sequences dataset.



Identificando familias de RNAs a partir  
de datos de RNA-seq

## Transcriptoma con genoma disponibile



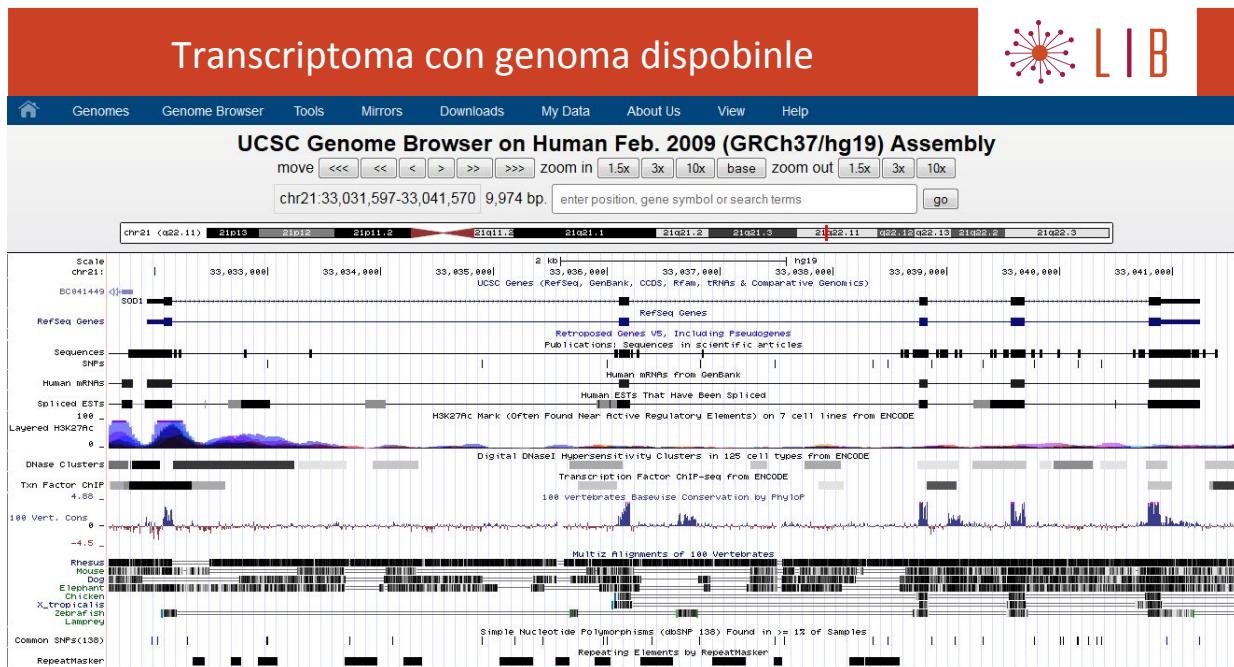
## Transcriptoma con genoma disponible



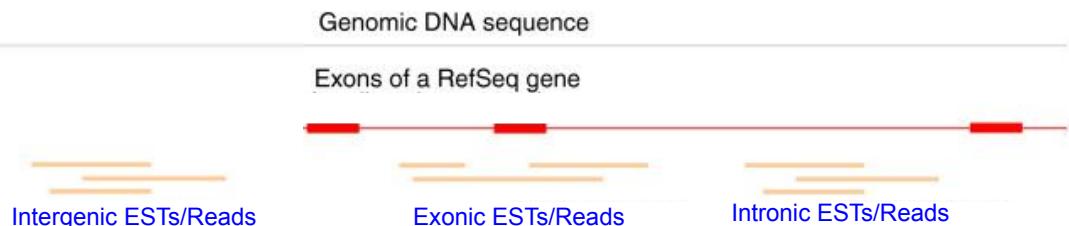
Caracterización usando manipulación de coordenadas genómicas

Coordenadas genómicas son asignadas a genomas

- chr21:33.031.597-33.041.570
- ¿Qué son estas coordenadas?  
UCSC Genome Browser (<http://genome.ucsc.edu>)



Caracterizar su contexto genómico

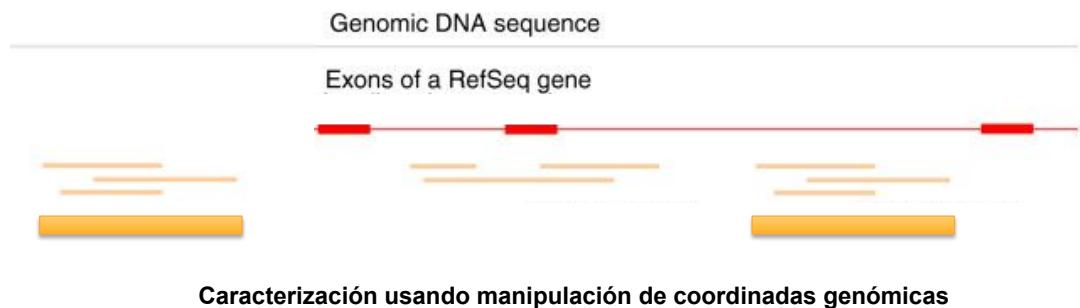


Caracterización usando manipulación de coordenadas genómicas

## Transcriptoma con genoma disponibile



Ensamblar su transcripto por overlapping de coordinadas



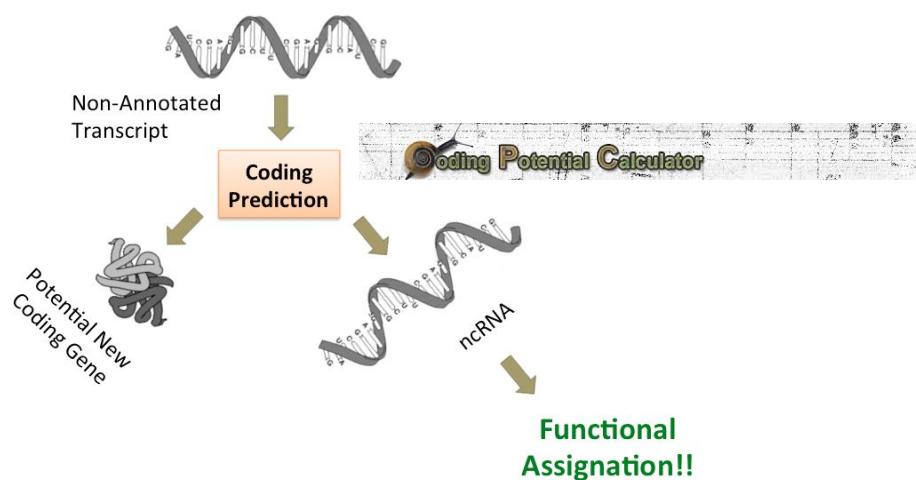
**BEDtools**

**cufflinks**

## Transcriptoma con genoma disponible



Predicción de su potencial codificante



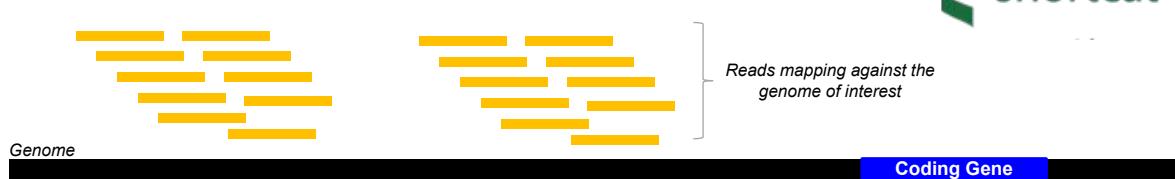


## shortCat: a toolkit for the identification, annotation and expression analysis of small RNAs from RNA-seq data

Victor Aliaga-Tobar, Raúl Arias-Carrasco,  
Vinicius Maracaja-Coutinho

Aliaga-Tobar et al  
To be submitted

Automatically performs the reads trimming and mapping  
against the genome of interest



Reads trimming and mapping against the  
genome sequence

Aliaga-Tobar et al  
To be submitted

Reconstruct small RNA transcripts through genomic coordinates manipulation



Aliaga-Tobar et al  
To be submitted

Automatically annotate ncRNAs according to user-provided references



Annotation based on genomic coordinates from known or predicted functional RNAs from different tools and databases



Aliaga-Tobar et al  
To be submitted

## Expression quantification

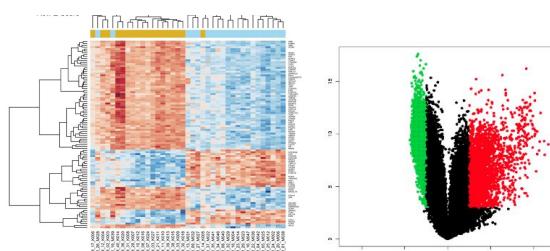


	339	797	339	797	797
SmallRNA_100	339	797	339	797	797
SmallRNA_1058	70	114	70	114	114
SmallRNA_1059	9	66	9	66	66
SmallRNA_1092	6	13	6	13	13
SmallRNA_1094	401	2087	401	2087	2087
SmallRNA_12	714	3082	714	3082	3082
SmallRNA_1238	30	115	30	115	115
SmallRNA_1255	49	99	49	99	99
SmallRNA_1292	3	14	3	14	14
SmallRNA_1295	5	14	5	14	14
SmallRNA_1299	1724	2483	1724	2483	2483
SmallRNA_13	10	42	10	42	42
SmallRNA_1364	40	198	40	198	198
SmallRNA_1366	859	1875	859	1875	1875
SmallRNA_139	16	38	16	38	38
SmallRNA_1405	27	36	27	36	36
SmallRNA_1433	18	45	18	45	45
SmallRNA_1483	5	19	5	19	19
SmallRNA_1503	9	7	9	7	7
SmallRNA_1505	97	171	97	171	171

Expression quantification based on  
reads belonging to assembled  
transcripts

Aliaga-Tobar et al  
To be submitted

## Differential expression analyses



Differential Expression  
Analysis

Aliaga-Tobar et al  
To be submitted



4794–4806 Nucleic Acids Research, 2018, Vol. 46, No. 9  
doi: 10.1093/nar/gky144

Published online 26 February 2018

## A regulatory RNA is involved in RNA duplex formation and biofilm regulation in *Sulfolobus acidocaldarius*

Alvaro Orell<sup>1,2,\*</sup>, Vanessa Tripp<sup>1</sup>, Victor Aliaga-Tobar<sup>3</sup>, Sonja-Verena Albers<sup>4</sup>,  
Vinicius Maracaja-Coutinho<sup>2,5</sup> and Lennart Randau<sup>1,\*</sup>



Neurobiology of Aging  
Volume 64, April 2018, Pages 123-138



Article

### Long Non-Coding RNAs Responsive to Salt and Boron Stress in the Hyper-Arid Lluteño Maize from Atacama Desert

Wilson Huanca-Mamani <sup>1,\*</sup>, Raúl Arias-Carrasco <sup>2</sup>, Steffany Cárdenas-Ninasivincha <sup>1</sup>, Marcelo Rojas-Herrera <sup>2</sup>, Gonzalo Sepúlveda-Hermosilla <sup>2</sup>, José Carlos Caris-Maldonado <sup>1,3</sup>, Elizabeth Bastías <sup>1</sup> and Vinicius Maracaja-Coutinho <sup>2,3,4,5,\*</sup>

Regular article

Genome-wide circulating microRNA expression profiling reveals potential biomarkers for amyotrophic lateral sclerosis

José Manuel Matamala <sup>a, b, c, d</sup>, Raul Arias-Carrasco <sup>d</sup>, Carolina Sanchez <sup>d</sup>, Markus Uhrig <sup>d</sup>, Leslie Bargsted <sup>a, b, f</sup>, Soledad Matus <sup>a, f, g</sup>, Vinicius Maracaja-Coutinho <sup>d</sup>, Sebastian Abarzúa <sup>a</sup>, Brigitte van Zundert <sup>e</sup>, Renato Verdugo <sup>c</sup>, Patricio Manque <sup>d</sup>, Claudio Hetz <sup>a, b, f, h, i, j, k, l</sup>



Database, 2017, 1–11  
doi: 10.1093/database/bax047  
Database tool



Database tool

### LeishDB: a database of coding gene annotation and non-coding RNAs in *Leishmania braziliensis*

Felipe Torres<sup>1,2</sup>, Raúl Arias-Carrasco<sup>3</sup>, José C. Caris-Maldonado<sup>3</sup>, Aldina Barral<sup>1,4,5</sup>, Vinicius Maracaja-Coutinho<sup>3,6,7,\*</sup> and Artur T. L. De Queiroz<sup>1,2,5</sup>

Indian J Microbiol  
<https://doi.org/10.1007/s12088-018-0775-4>

ORIGINAL RESEARCH ARTICLE

### Prediction of MicroRNAs in the Epstein–Barr Virus Reveals Potential Targets for the Viral Self-Regulation

Victor Serrano-Solis<sup>1</sup>, Angelica Cardoso Carlos<sup>1</sup>, Vinicius Maracaja-Coutinho<sup>2,3,4</sup>, Sávio Torres de Farias<sup>1</sup>



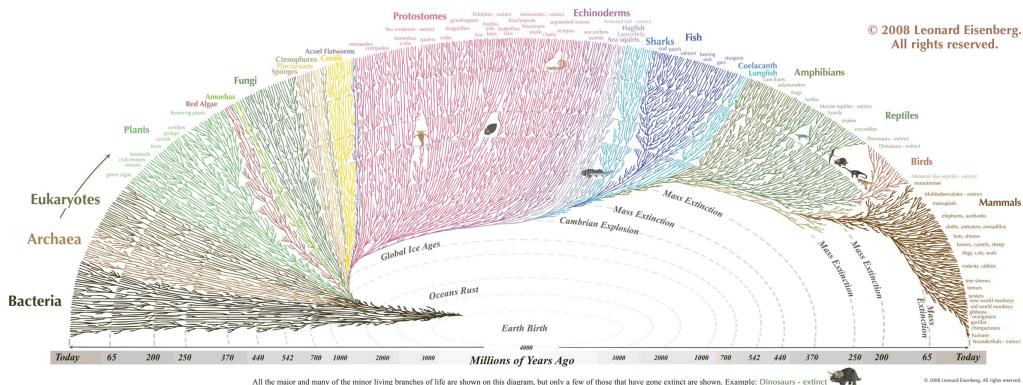
RESEARCH ARTICLE  
Core non-coding RNAs of *Piscirickettsia salmonis*

Christopher Segovia<sup>1,2</sup>, Raul Arias-Carrasco<sup>2,3</sup>, Alejandro J. Yáñez<sup>4</sup>, Vinicius Maracaja-Coutinho<sup>2,5</sup>, Javier Santander<sup>1,\*</sup>

## Non-coding RNAs, complexity, disease and evolution



© 2008 Leonard Eisenberg.  
All rights reserved.



**Virus:** Hepesvirus; Zika Virus

**Archaea:** +250 genomes

**Bacteria:** *Piscirickettsia salmonis*

**Plants:** *Zea mays Lluteño*, *Cannabis sativa*

**Protozoa:** *Leishmania ssp.*

**Platyhelminthes:** *Schistosoma mansoni*

**Insecta:** *Drosophila melanogaster*;

*Apis mellifera*

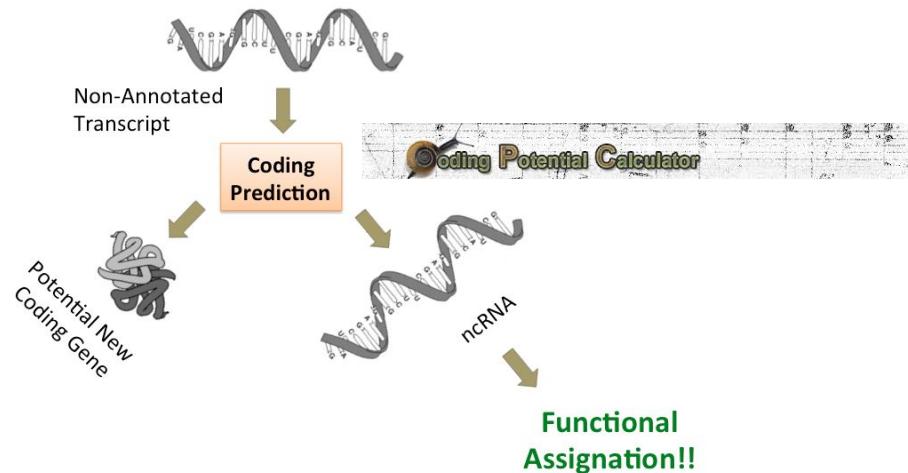
## Transcriptoma sin genoma disponible – De Novo



# Transcriptoma sin genoma dispobinle – De Novo



## Predicción de su potencial codificante



Asignación funcional a ARNs no codificantes por medio de métodos computacionales



Asignación funcional a ARNs no codificantes por medio de  
comparaciones de secuencias (similitudes)



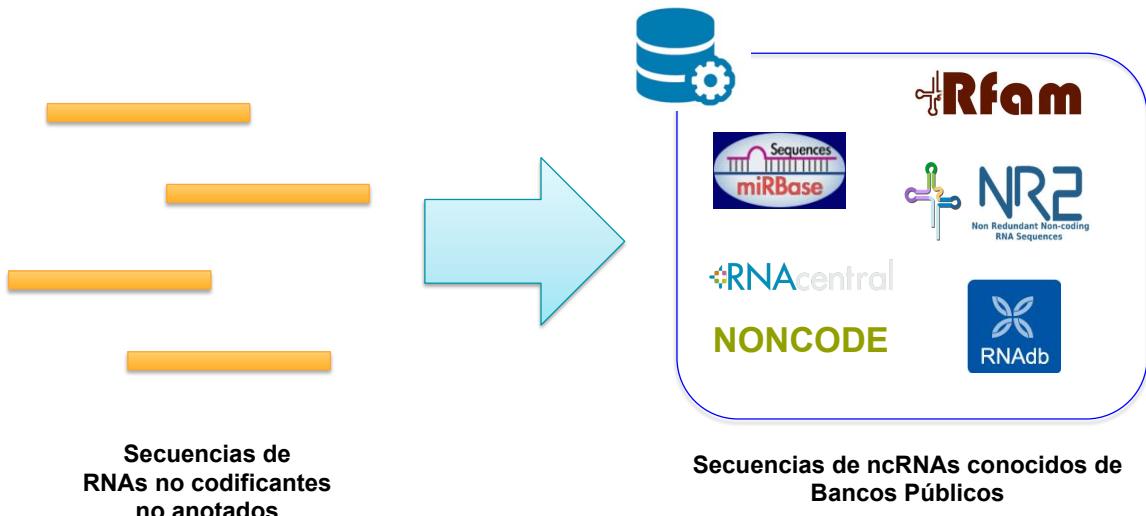
Asignando función a RNAs por medio de  
comparaciones de secuencias

The screenshot shows a web browser window with the URL "bowtie-bio.sourceforge.net/index.shtml". The page header features the "Bowtie" logo (two overlapping shapes) and the text "An ultrafast memory-efficient short read aligner". To the right is the Johns Hopkins University logo. The main content area describes Bowtie as an ultrafast, memory-efficient short read aligner that indexes the genome with a Burrows-Wheeler index. It mentions a rate of over 25 million 35-bp reads per hour and a memory footprint of about 2.2 GB for the human genome (2.9 GB for paired-end). A green "OSI certified" badge is visible in the bottom right corner.

UCSC Genome Bioinformatics  
BLAT

BLASTn

## Asignando función a RNAs por medio de comparaciones de secuencias

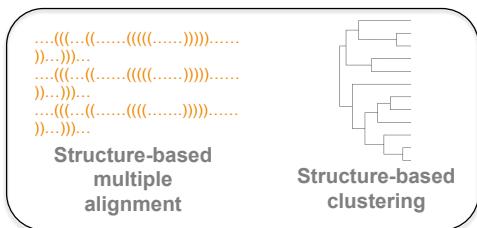


Asignación funcional a ARNs no codificantes por medio de predictores específicos

Asignando función a RNAs por medio de  
otros predictores específicos



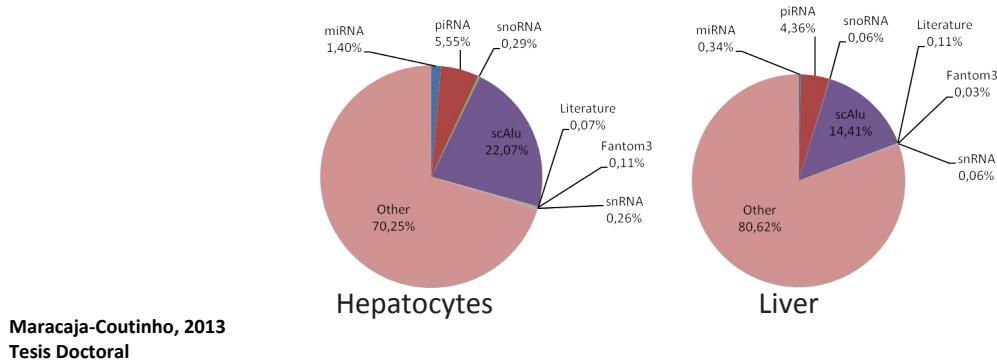
**SnoReport**  
**HHMMIR**  
**MIPRED**  
**Novomir**  
**Triplet**



Asignando función a RNAs por medio de  
co-expresión de transcritos

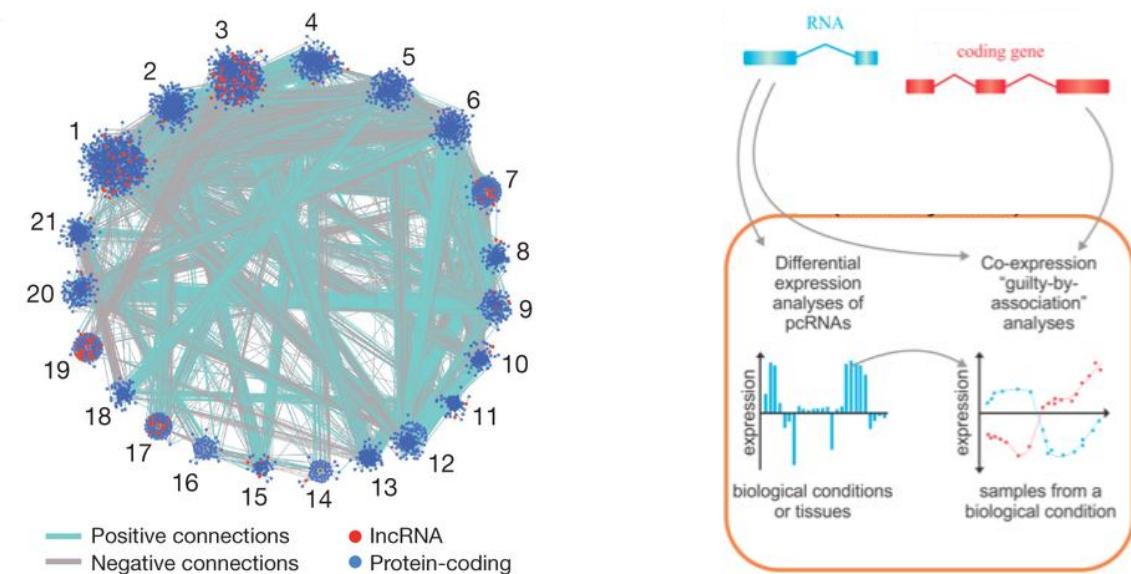


Todavía es muy difícil asignar función a ARNs no codificantes a partir de características en su secuencias

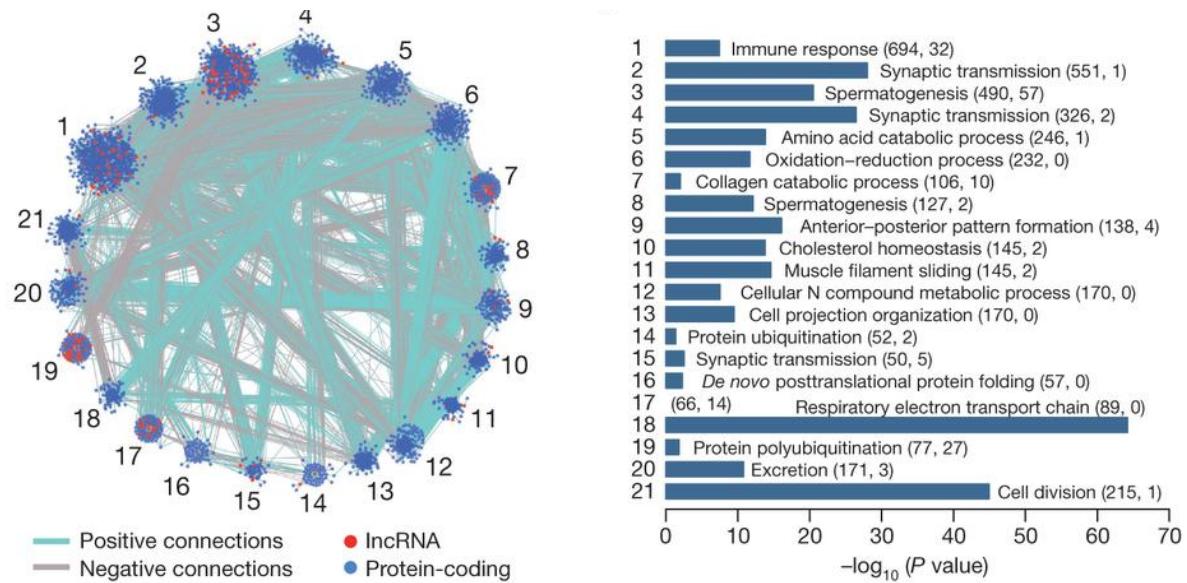


Maracaja-Coutinho, 2013  
Tesis Doctoral

### Asignando función a RNAs por medio de co-expresión de transcritos



## Asignando función a RNAs por medio de co-expresión de transcritos



## Asignando función a RNAs por medio de co-expresión de transcritos



Russo et al. BMC Bioinformatics (2018) 19:56  
<https://doi.org/10.1186/s12859-018-2053-1>

BMC Bioinformatics

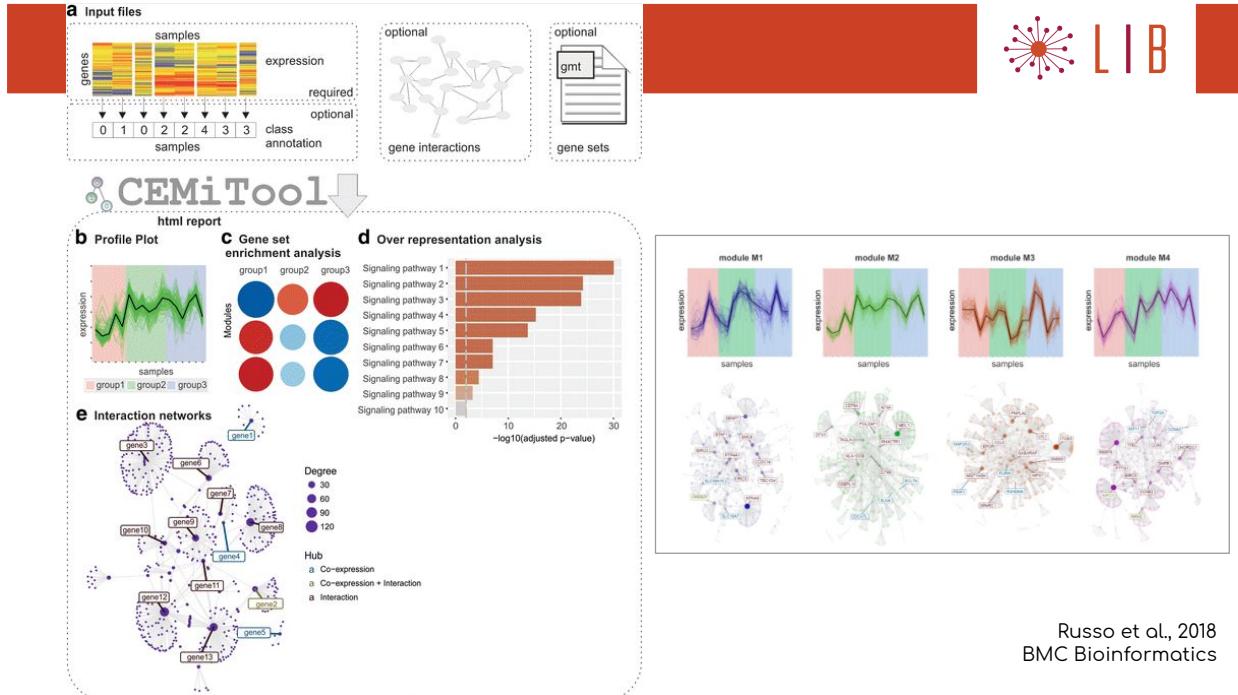
SOFTWARE

Open Access



### CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses

Pedro S. T. Russo<sup>1†</sup>, Gustavo R. Ferreira<sup>1†</sup>, Lucas E. Cardozo<sup>1</sup>, Matheus C. Bürger<sup>1</sup>, Raul Arias-Carrasco<sup>2</sup>, Sandra R. Maruyama<sup>3</sup>, Thiago D. C. Hirata<sup>1</sup>, Diógenes S. Lima<sup>1</sup>, Fernando M. Passos<sup>1</sup>, Kiyoshi F. Fukutani<sup>3</sup>, Melissa Lever<sup>1</sup>, João S. Silva<sup>3</sup>, Vinícius Maracaja-Coutinho<sup>2</sup> and Helder I. Nakaya<sup>1\*</sup>



## Asignando función a RNAs por medio de co-expresión de transcritos

**CEMiTool**

Home Run CSBL

**Submit your data**

Name your analysis. Example: HIV infection

Email to send you the results when ready

p-value for gene-gene correlation. Default=0.05

Number of permutations. Default=1,000

Minimal genes per modules. Default=20

Correlation Method. Default=Spearman

Seleccionar Archivo: *ningun archivo seleccionado*

Submit

File should contain selected genes/probes as rows and samples as columns (tab-delimited).  
Expression table must have less than 4,000 genes (rows) and 400 samples (columns) See example [here](#)

Russo et al. 2019  
Frontiers in Genetics

## LncRNAs associated to vaccine-induced immunity

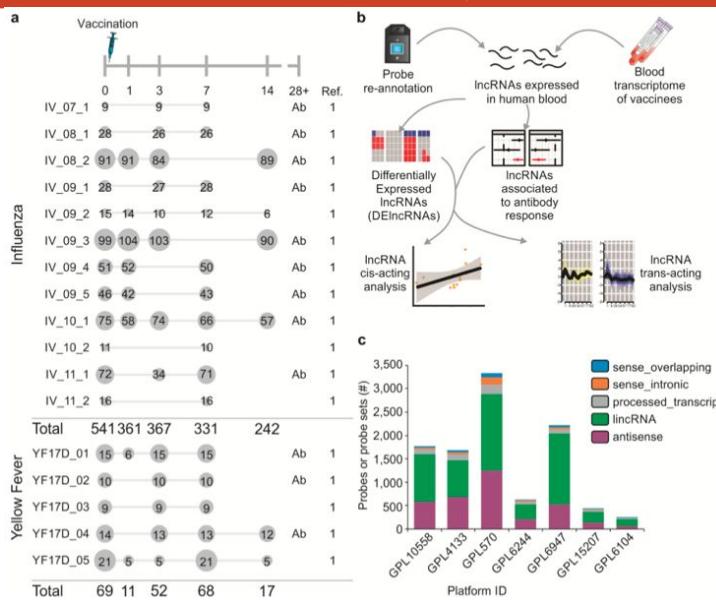
**ACCDIS**  
ADVANCED CENTER FOR CHRONIC DISEASES

Molecular  
Degree  
of  
Perturbation

Meta  
Volcano

CEMiTool

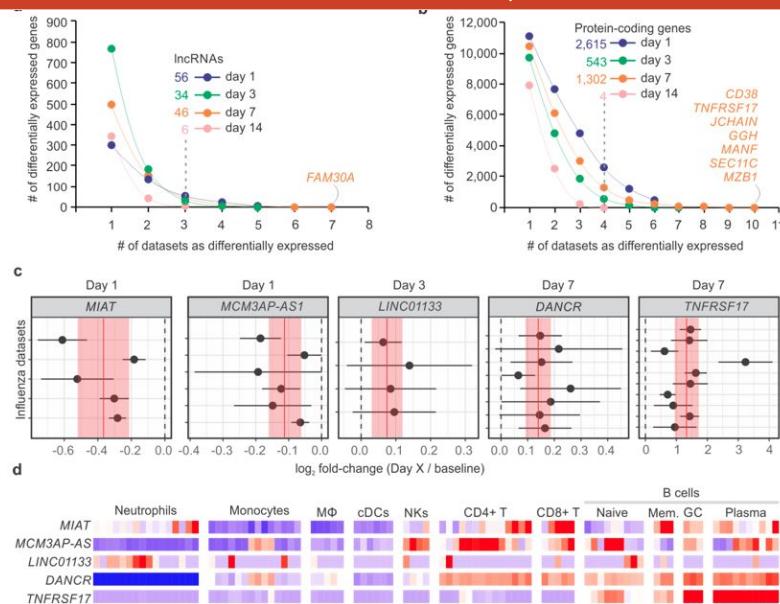
12 cohorts



Lima et al., 2019  
Submitted

## LncRNAs associated to vaccine-induced immunity

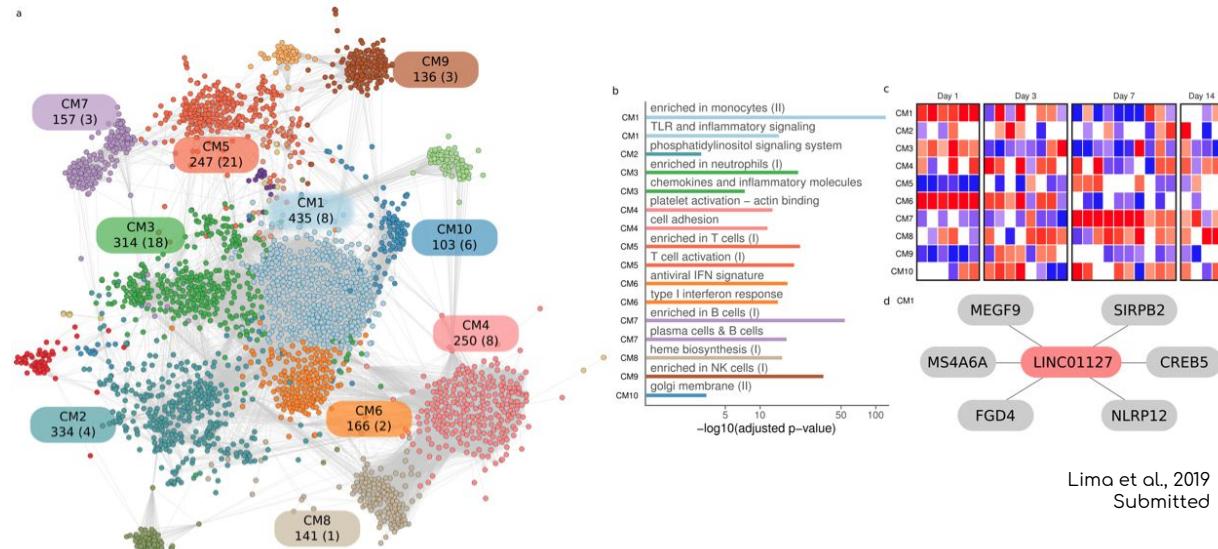
**ACCDIS**  
ADVANCED CENTER FOR CHRONIC DISEASES



Lima et al., 2019  
Submitted

# LncRNAs associated to vaccine-induced immunity

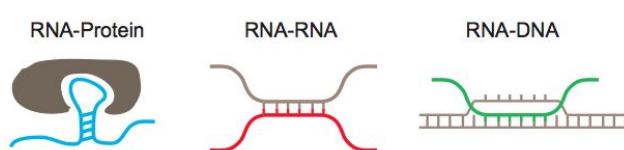
**ACCDIS**  
ADVANCED CENTER FOR CHRONIC DISEASES



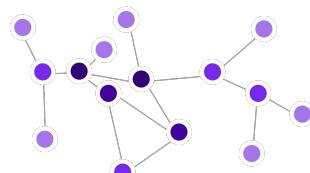
Es una larga jornada....



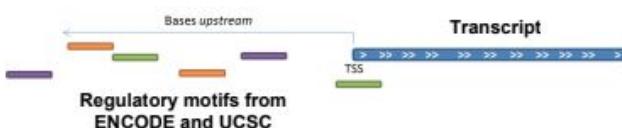
## Interacción con otras moléculas



## Redes de interacción



## Regulación



THANK YOU!



 @vin\_maracaja

 vinicius.maracaja@uchile.cl