LABORATORY OF INTEGRATIVE BIOINFORMATICS

Bioinformática de RNAs

Vinicius Maracaja-Coutinho, PhD

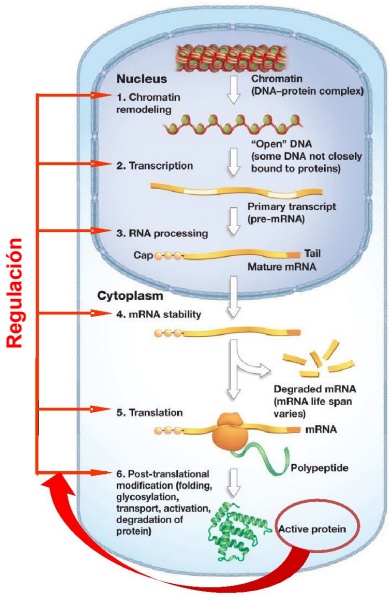
FACULTAD DE CIENCIAS QUÍMICAS Y FARMACÉUTICAS
UNIVERSIDAD DE CHILE

ACCDiS
ADVANCED CENTER FOR CHRONIC DISEASES

@vin_maracaja
vinicius.maracaja@uchile.cl

Los mecanismos genéticos fueran 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

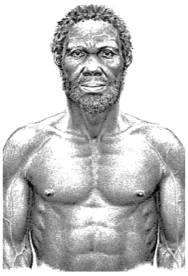
DROSOPHILA MELANOGASTER
14.399

TAKIFUGU RUBRIPES
22.089

MUS MUSCULUS
26.996

CAENORHABDITIS ELEGANS
20.603

?



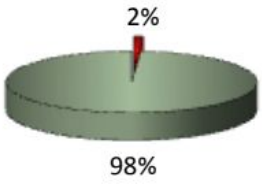
HOMO SAPIENS
23.299

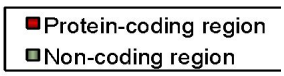
TRYPANOSOMA CRUZI
22.570

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

Solo 2% del del genoma codifica proteínas

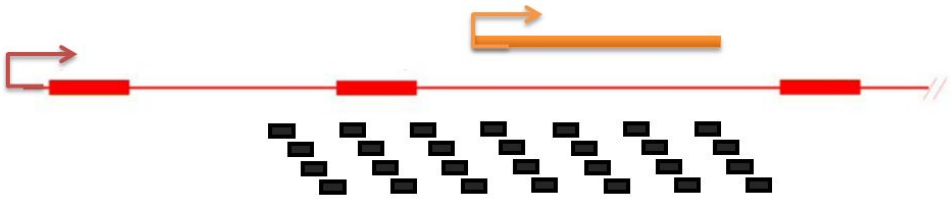
Donde viene los elementos responsables por la regulación de los procesos celulares?





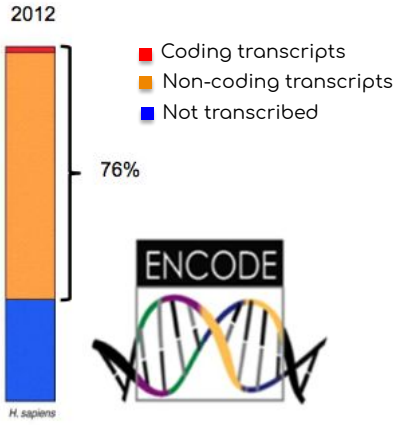
Tenemos una transcripción diseminada

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

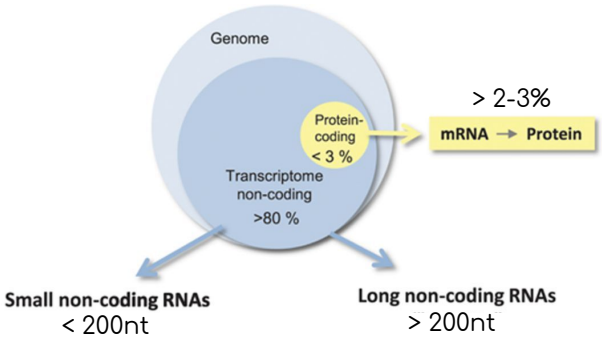


Tiling Arrays en todo el cromosoma 21 y 22

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

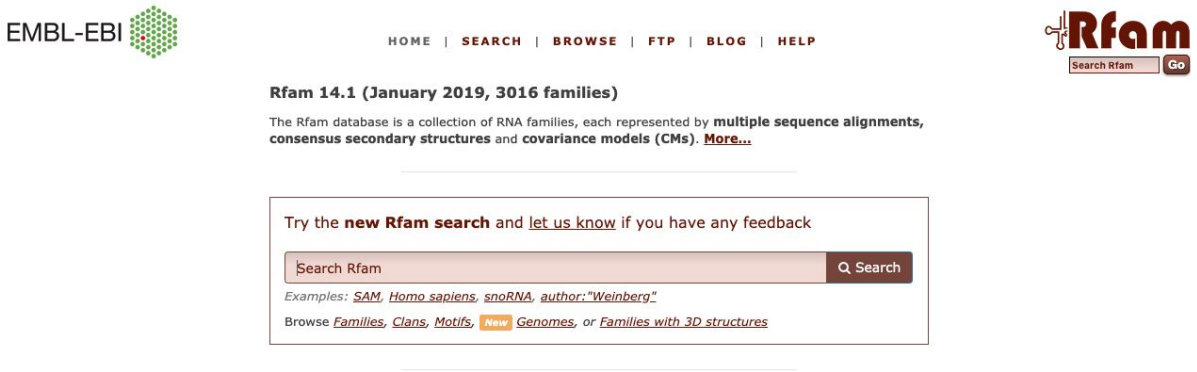


Djebali et al, 2012.
Nature



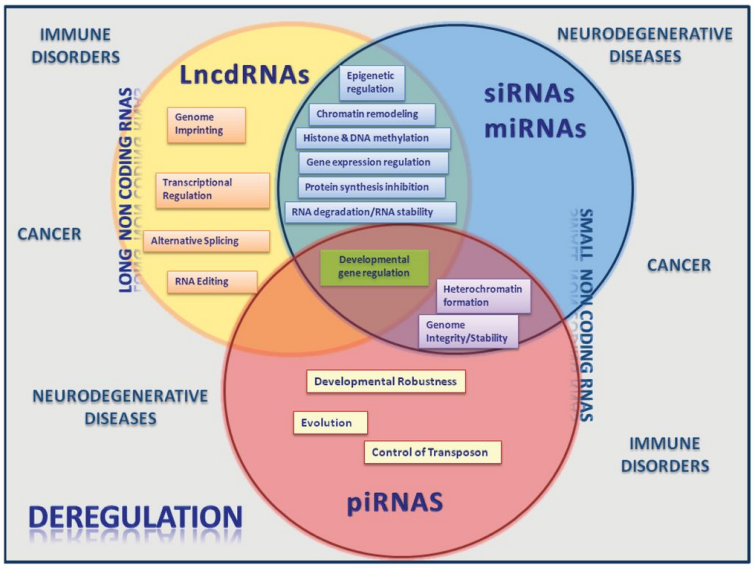
Uchida and Dimmeler, 2015.
Circulation Research

Más de 3 mil familias de RNAs



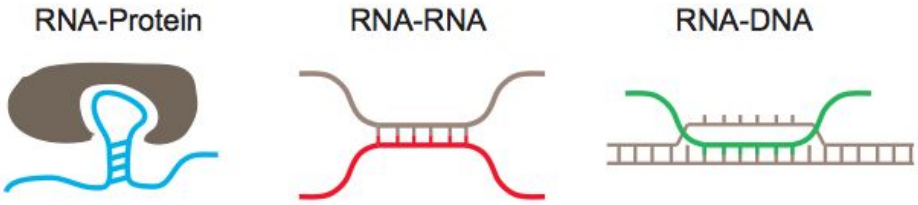
Rfam
database
Enero 2019

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

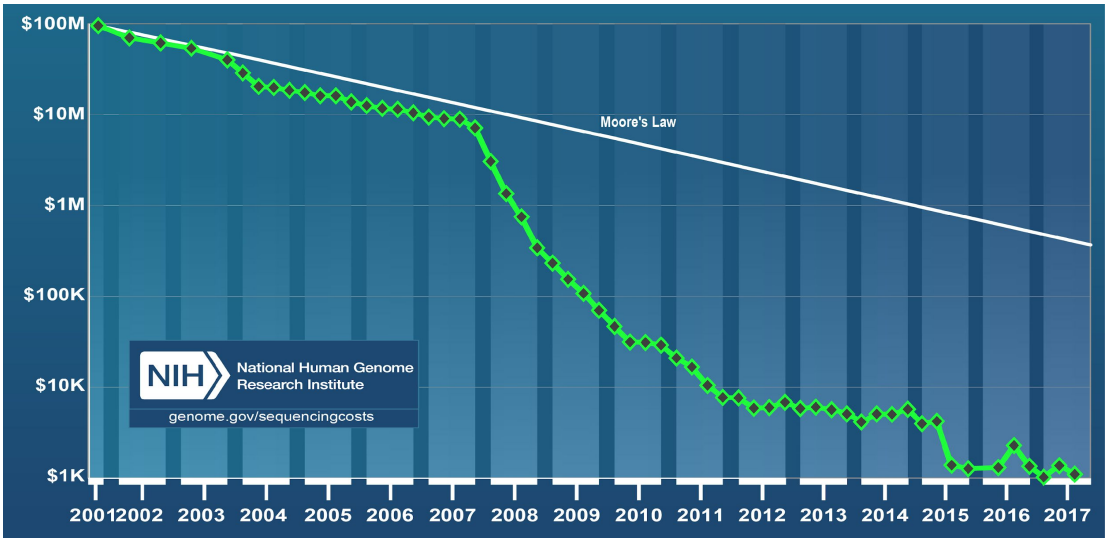


Gomes et al., 2013.
International Journal of Molecular Sciences

RNAs y sus interacciones



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

ENCODE

i5k

GENOME 10K

Fish T1K

Genomics england
100k
England
Genomes

1K Plants
Transcriptomes

ROADMAP epigenomics PROJECT

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

1KFG:
1,000 Fungal
Genomes

1K Cannabis
Genomes

Life Sciences "Big Data" Revolution

Oportunidades!

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

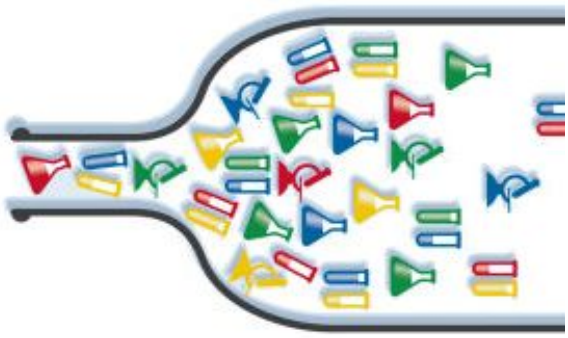


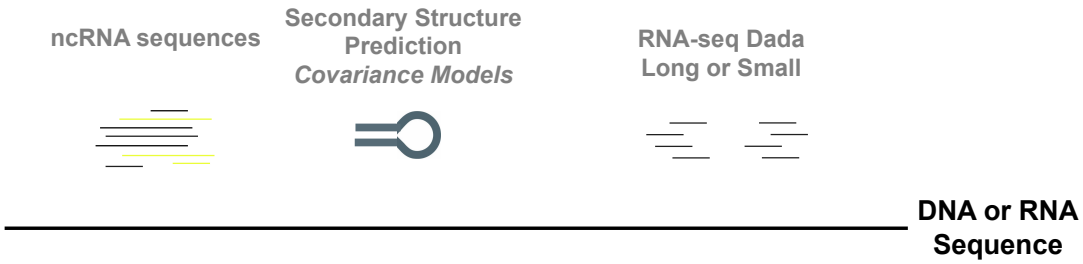
Image:
http://thebottlenecker.com/

- **Identificando familias de RNAs en secuencias de nucleótidos**
  - Similitudes de secuencias
  - Modelos de Covariance de RNAs (Estructuras secundárias 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

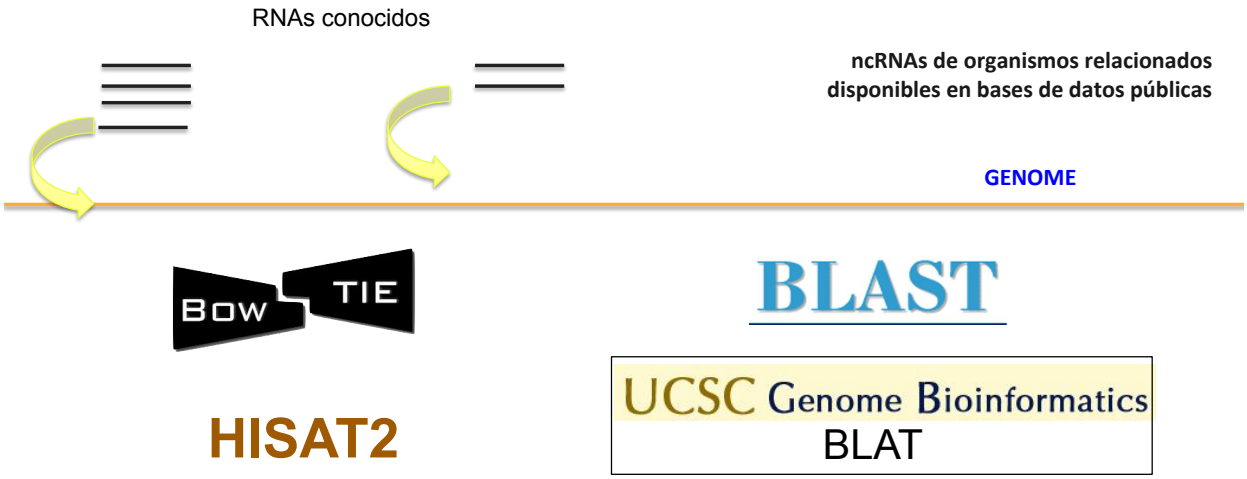
LIB

Identificando familias de RNAs en secuencias de nucleótidos

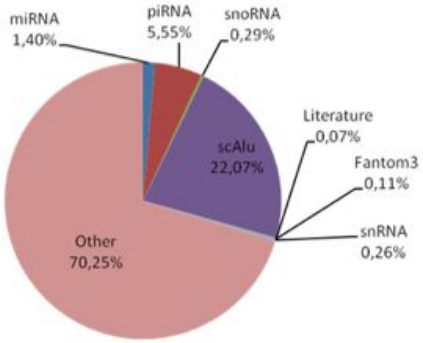
Predicción de genes de RNAs



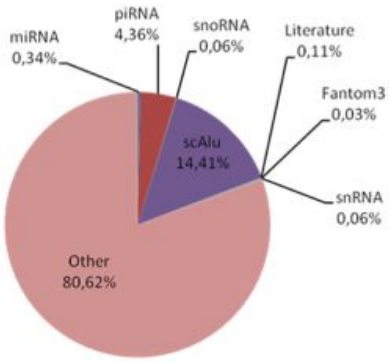
Similitudes de Secuencias



2007 - Non-coding transcriptome was comprised by unclassified ncRNAs



Human Hepatocytes

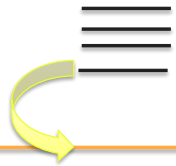


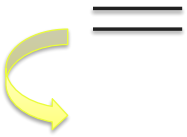
Human Liver

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

RNAs conocidos





ncRNAs de organismos relacionados disponibles en bases de datos públicas

GENOME

Rfam

NONCODE

RNAcentral

RNAdb

Sequences miRBase

NR2 Non Redundant Non-coding RNA Sequences

NRDR: The Non-coding RNA Databases Resource

RNA Biology 9:3, 274-282; March 2012; © 2012 Landes Bioscience

RNAbiology

Non-coding transcription characterization and annotation

A guide and web resource for non-coding RNA databases

Alexandre Rossi Paschoal,[3,5,†] Vinicius Maracaja-Coutinho,[2,5,†] João Carlos Setubal,[2] Zilá Luz Paulino Simões,[4] Sergio Verjovski-Almeida[3] and Alan Mitchell Durham[1,*]

[1]Departamento de Ciência da Computação; Instituto de Matemática e Estatística; Universidade de São Paulo; São Paulo, Brazil; [2]Departamento de Bioquímica; Instituto de Química; Universidade de São Paulo; São Paulo, Brazil; [3]Engenharia da Computação; Universidade Tecnológica Federal do Paraná – Campus Cornélio Procópio; Cornélio Procópio, Brazil; [4]Departamento de Biologia; Faculdade de Filosofia Ciências e Letras de Ribeirão Preto; Universidade de São Paulo; São Paulo, Brazil; [5]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

NRDR Non-coding RNA Databases Resource

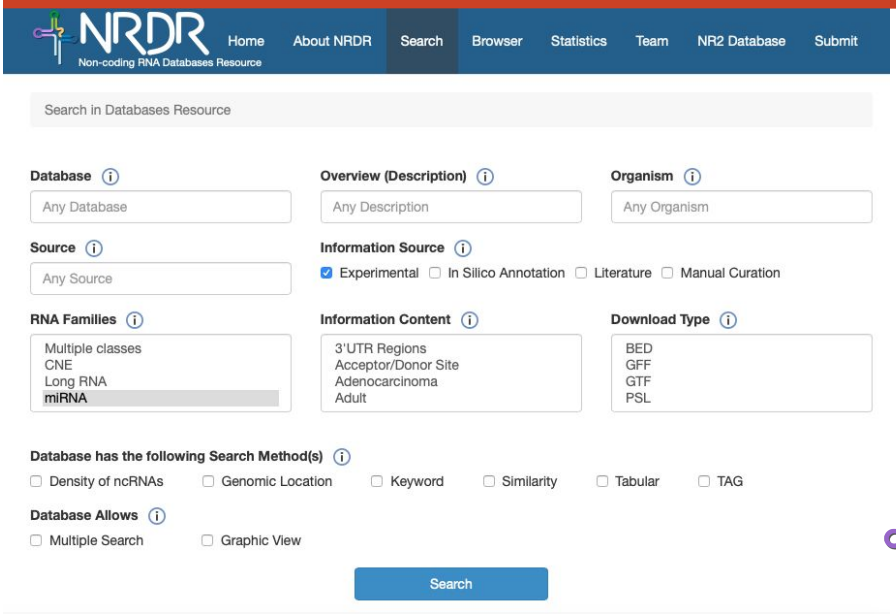
NRDR: The Non-coding RNA Databases Resource

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 Non-coding RNA Databases Resource

NRDR portal

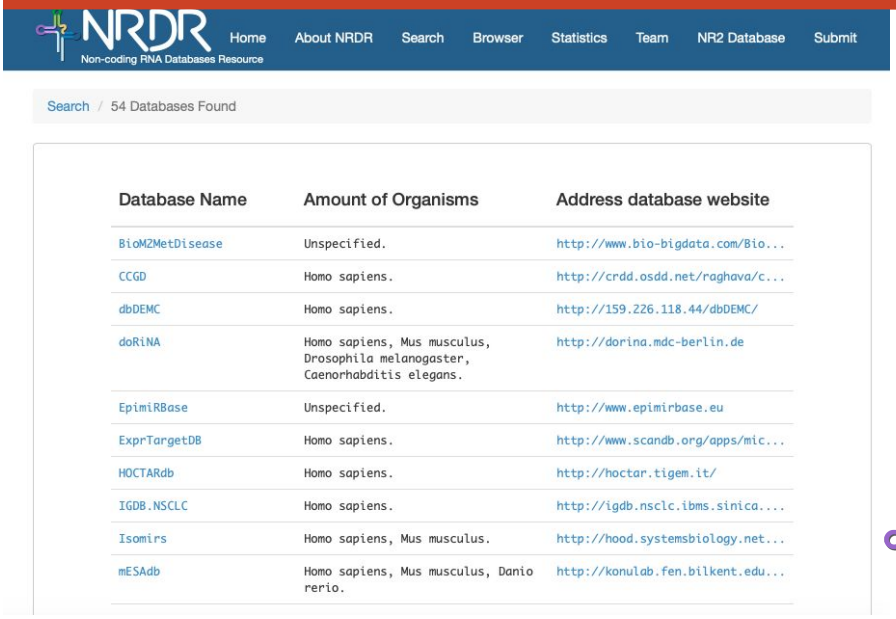


www.ncrnadatabases.org

229 indexed databases

NRDR
Non-coding RNA Databases Resource

NRDR: The Non-coding RNA Databases Resource



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

NRDR
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/

LIB

NRDR Non-coding RNA Databases Resource

NRDR: The Non-coding RNA Databases Resource

LIB

January, 2019

Chapter 10

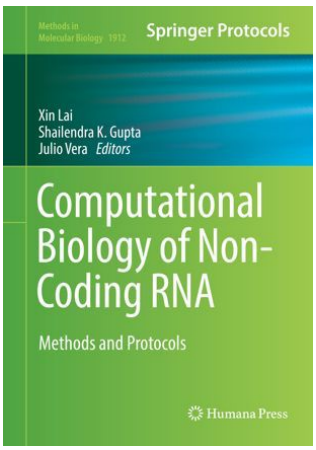
Noncoding RNAs Databases: Current Status and Trends

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

Abstract

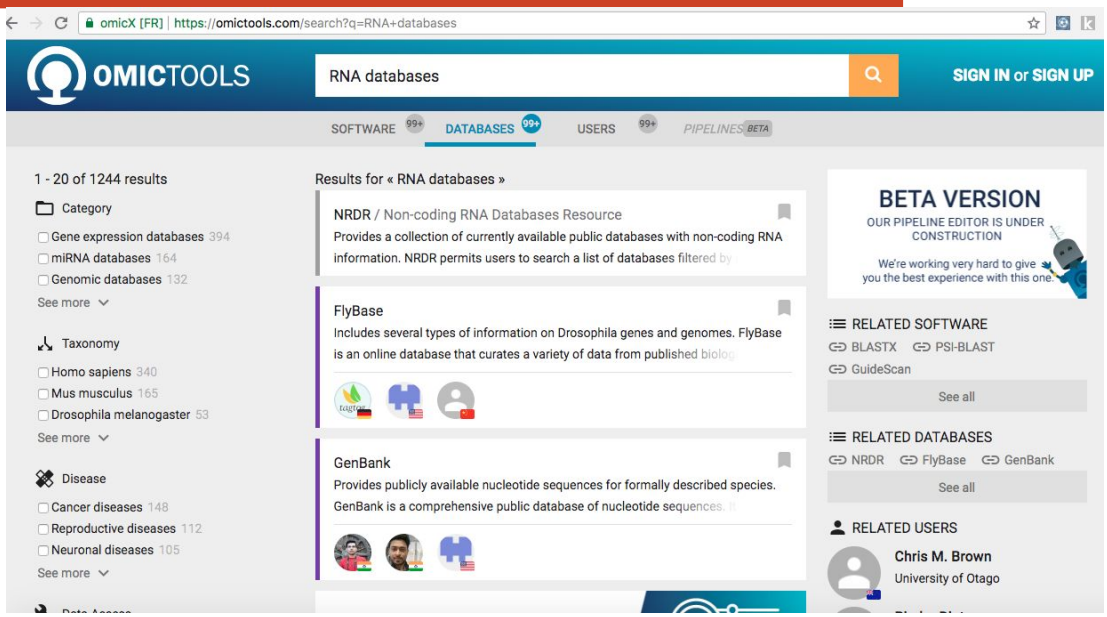
One of the most important resources for researchers of noncoding RNAs is the information available in public databases spread over the internet. However, the effective exploration of this data can represent a daunting task, given the large amount of databases available and the variety of stored data. 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, MicroRNA, lncRNA, Circulating RNAs, Review

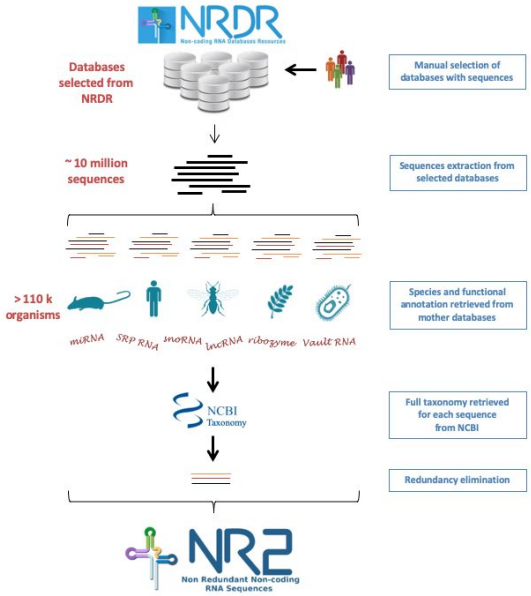


NRDR Non-coding RNA Databases Resource

OMICtools



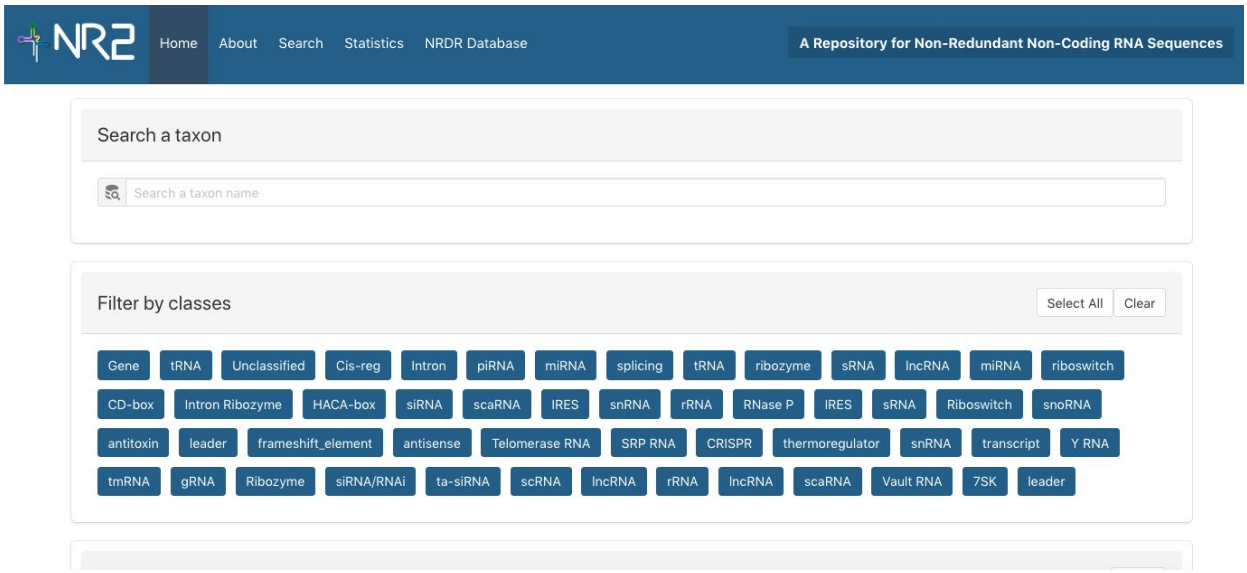
NR2 redundancy elimination and annotation process overview



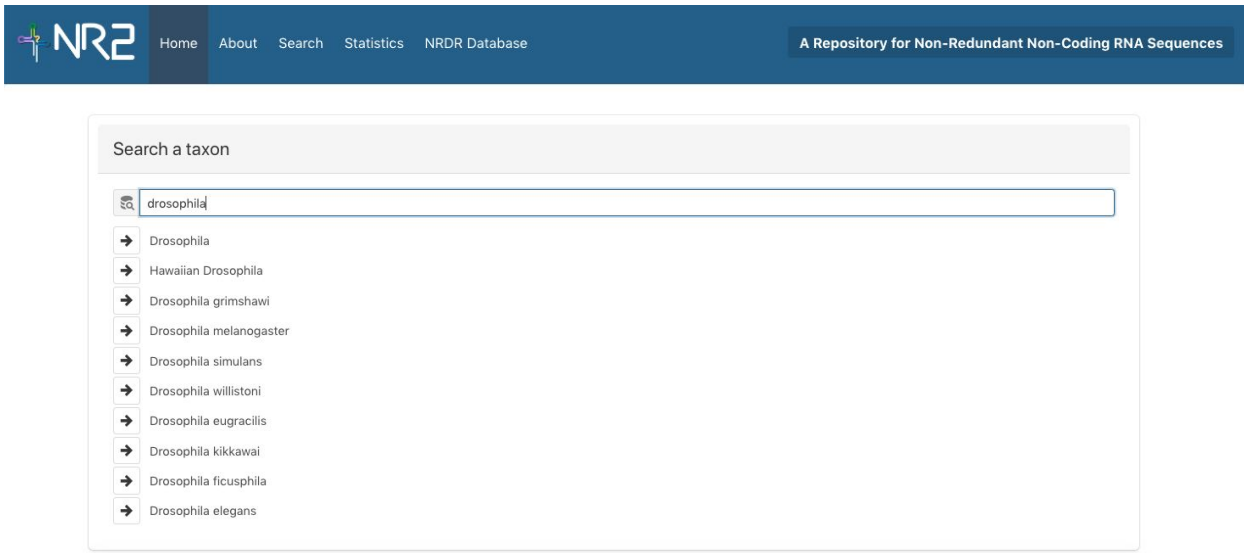
112,963 different organisms

> 10 million sequences

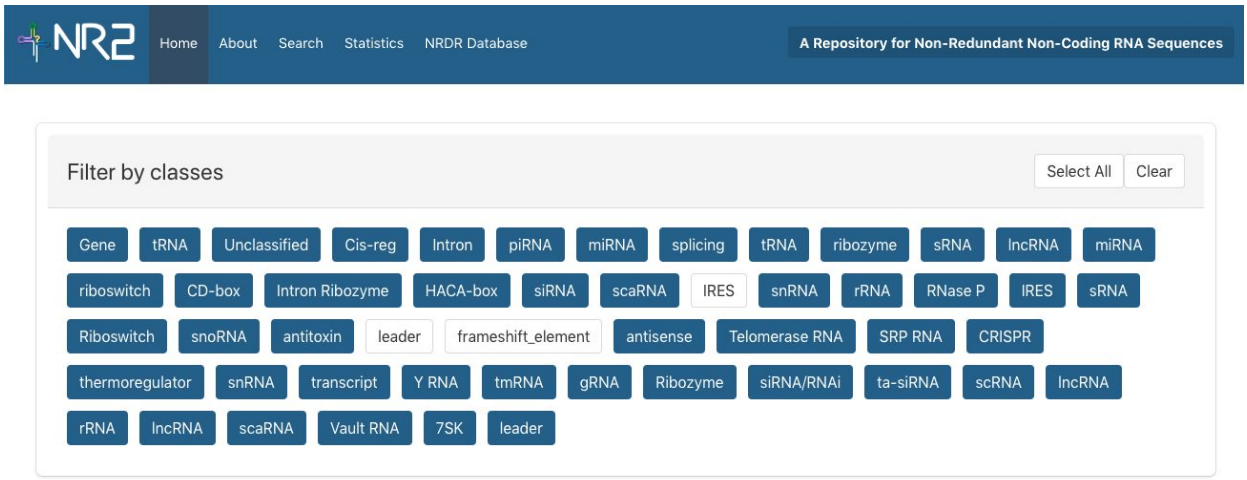
NR2 portal (alfa version available)



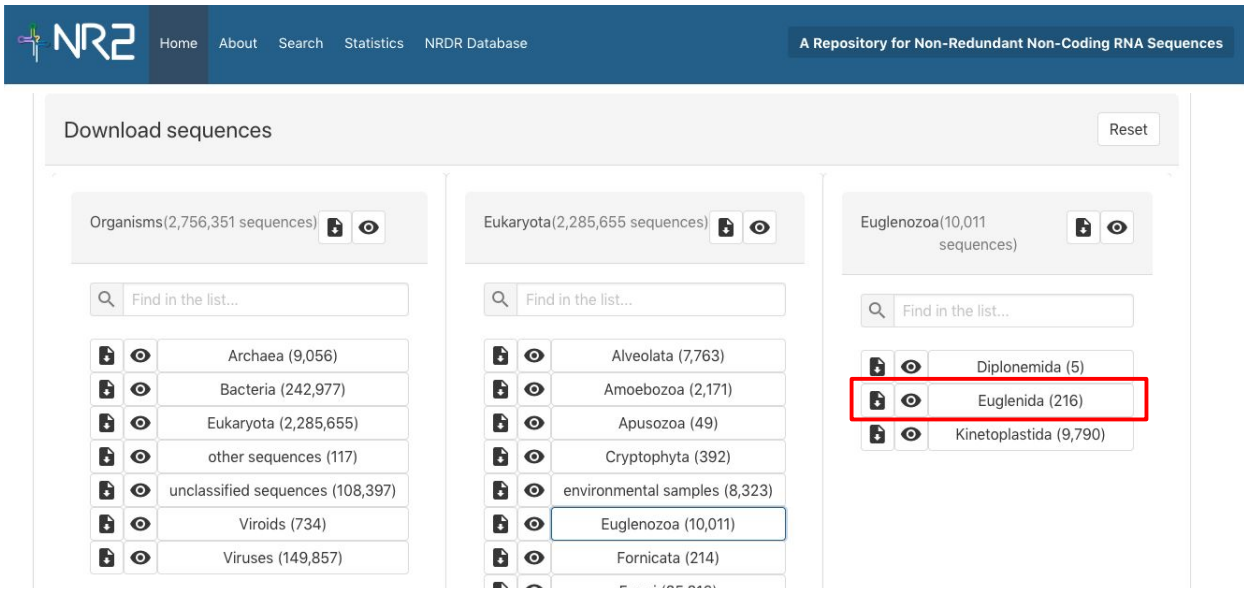
Search for specific taxon using keywords



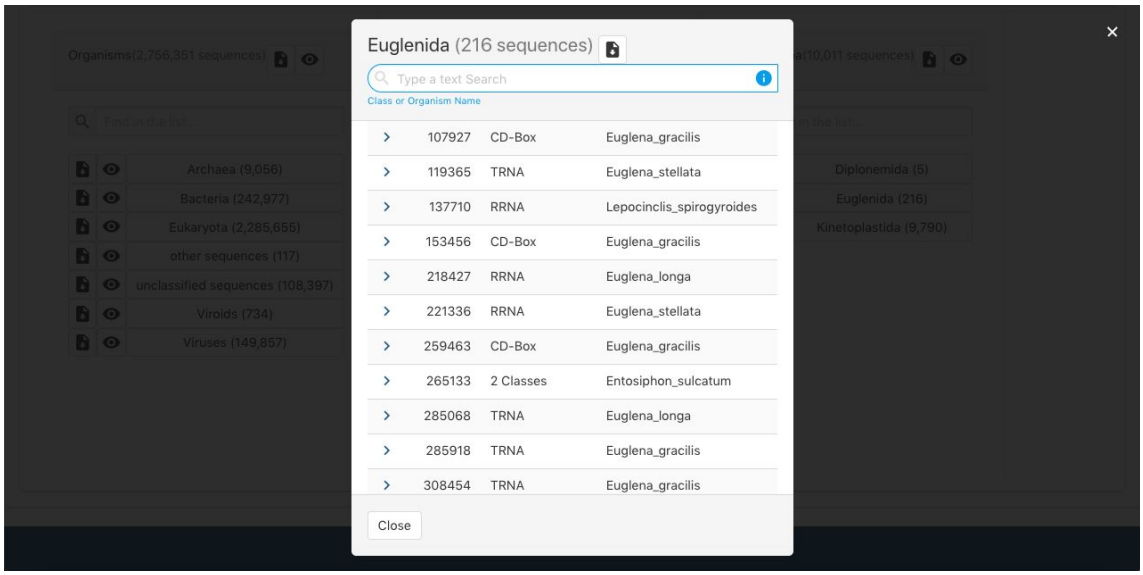
Filter by specific RNA classes of interest



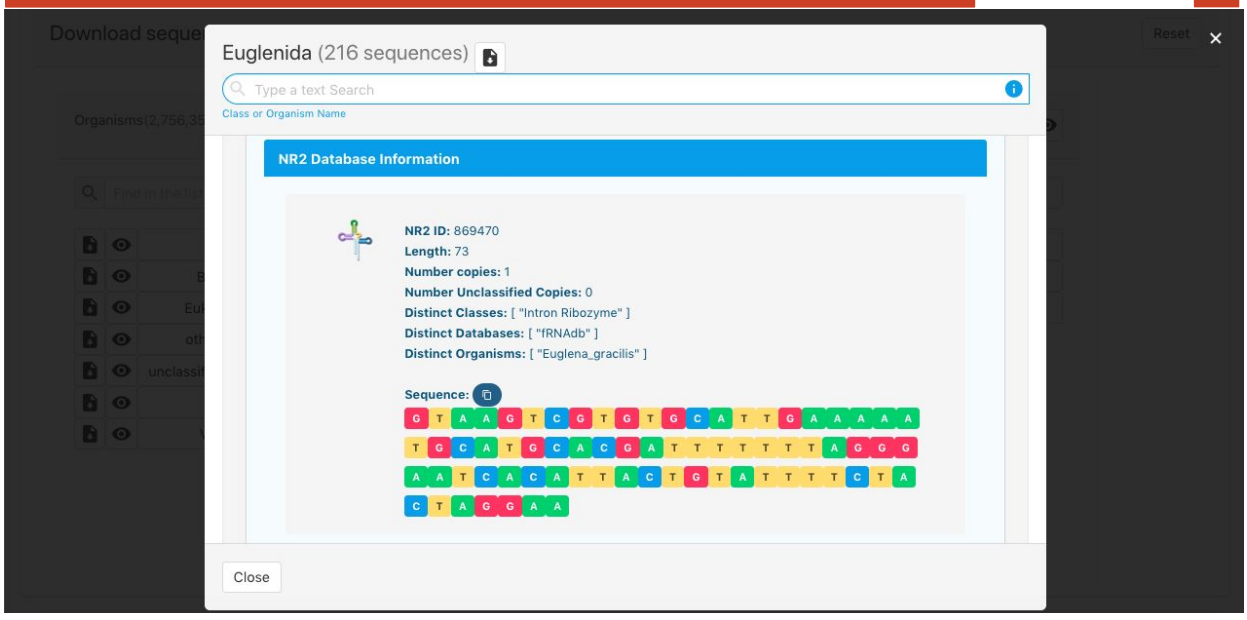
Navigate through taxonomy in order to download non-redundant sequences or obtain additional information from the database



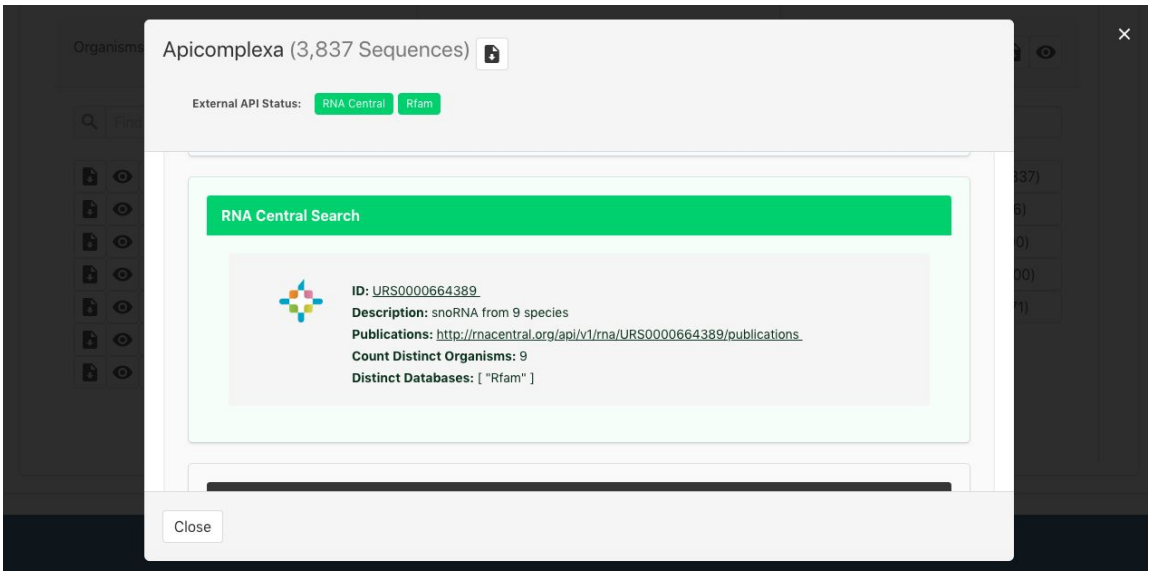
Navigate through taxonomy in order to download non-redundant sequences or obtain additional information from the database



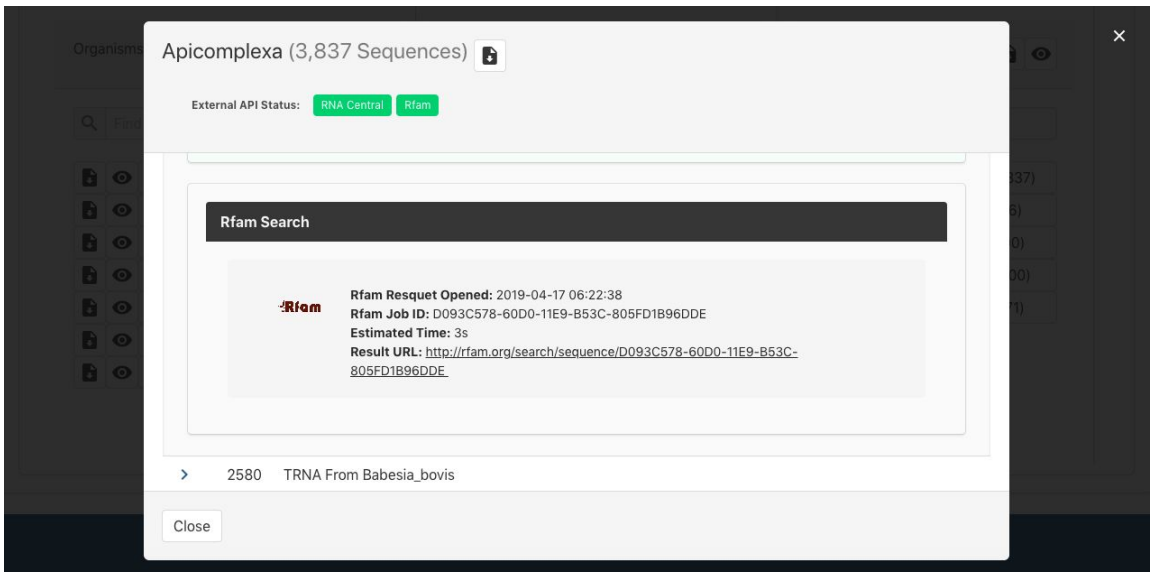
NR2 generated information



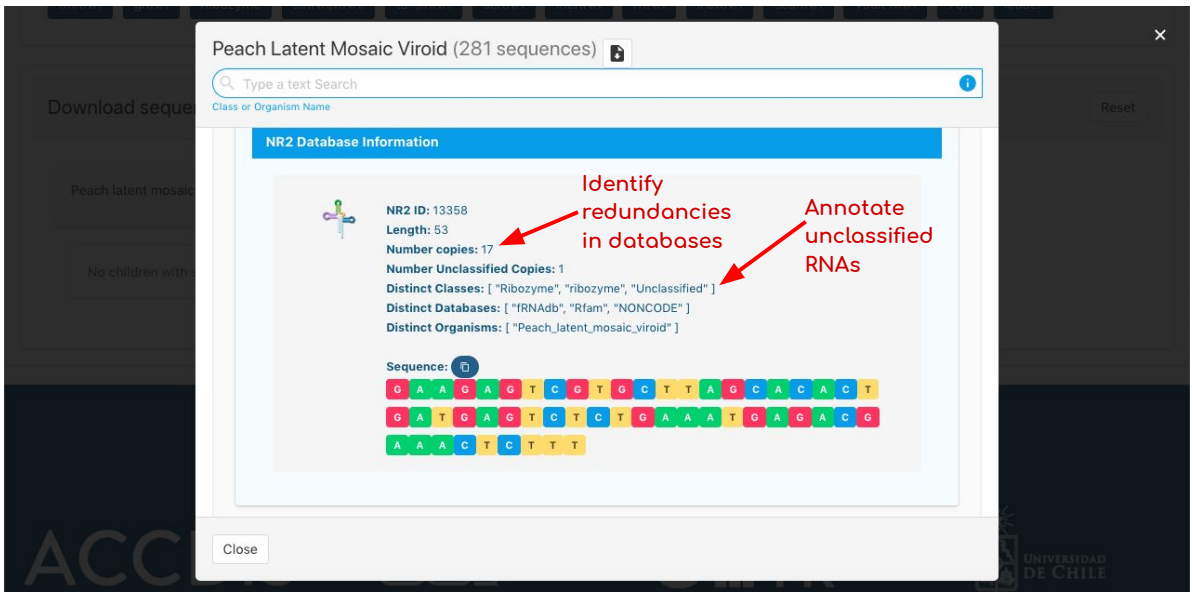
Connection with RNA Central API



Connection with Rfam API



Identify redundancies and improve RNA annotations



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

Known RNA sequences

NR2 Non Redundant Non-coding RNA Sequences

ncRNAs from related organisms available in public databases

GENOME

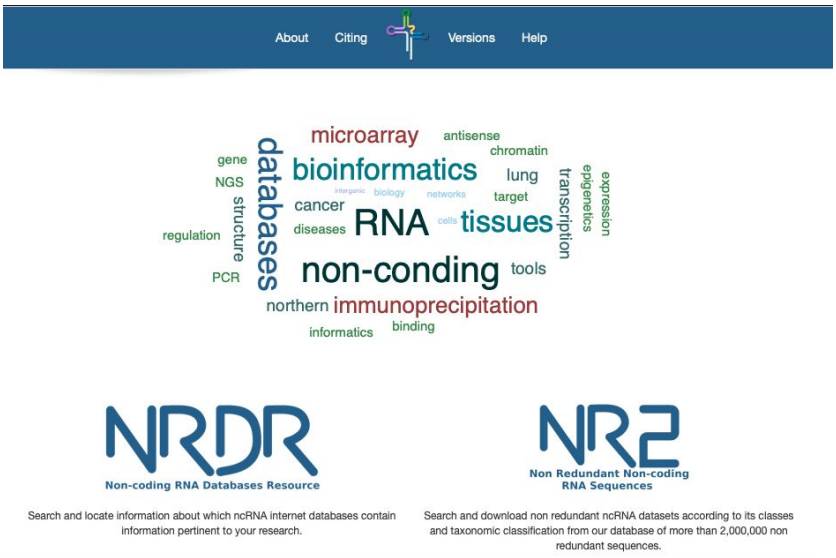
BOW TIE

HISAT2

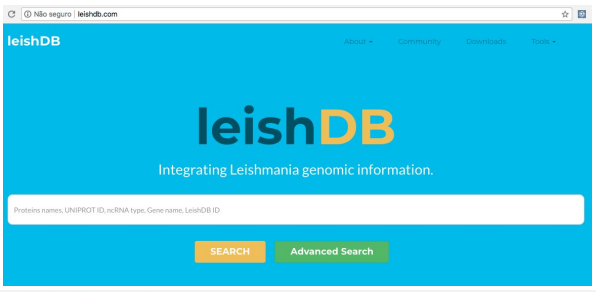
BLAST

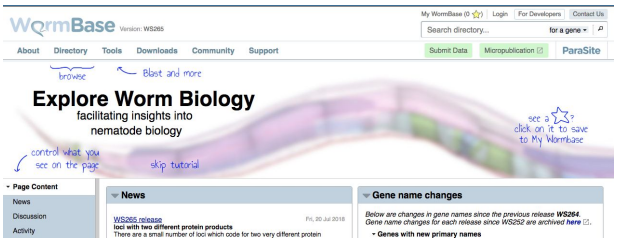
UCSC Genome Bioinformatics
BLAT

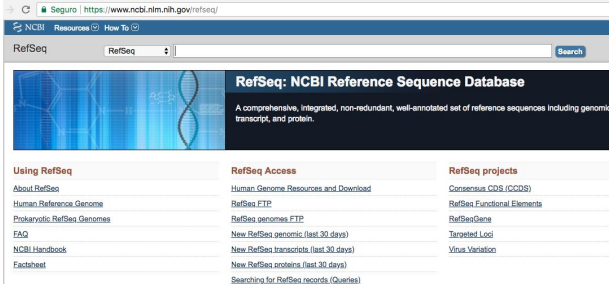
www.ncrnadatabases.org

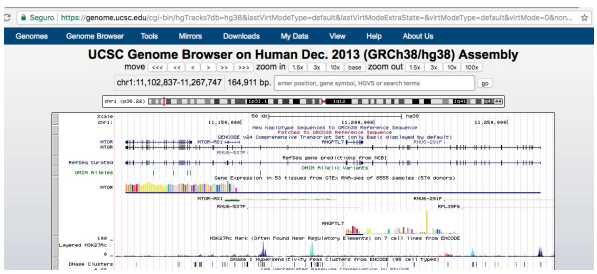


Bancos Específicos y Genéricos

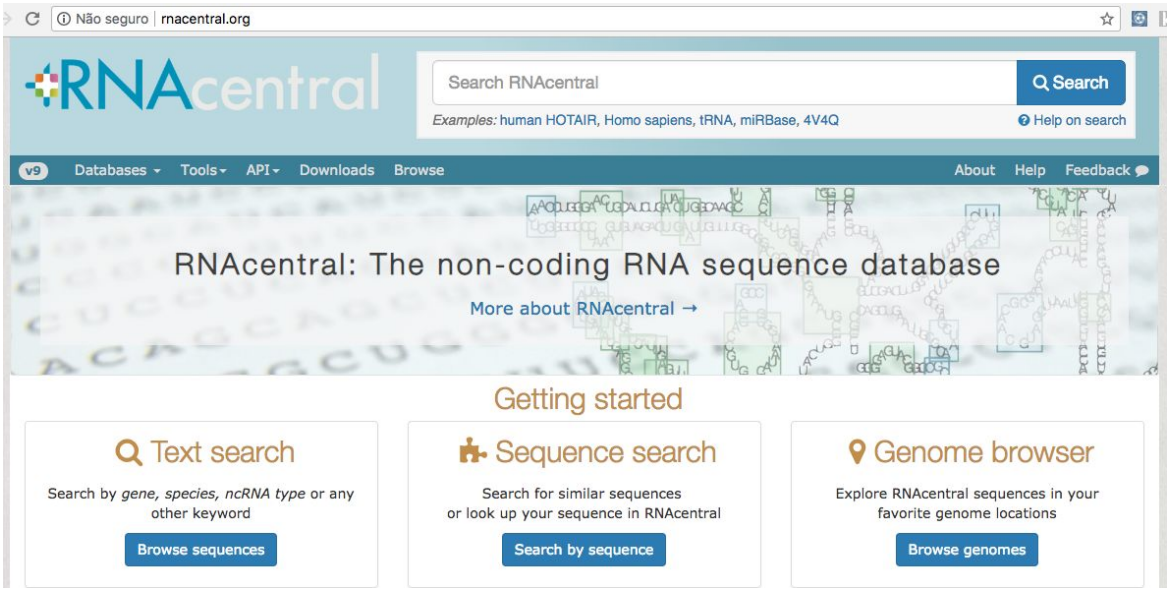








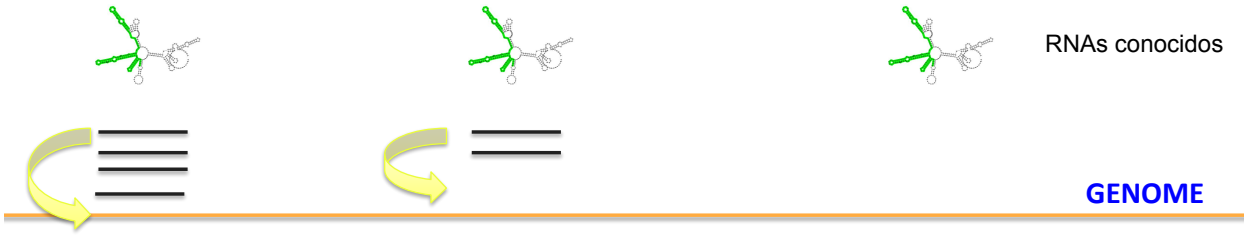
RNAcentral



Identificando RNAs a partir de predicción

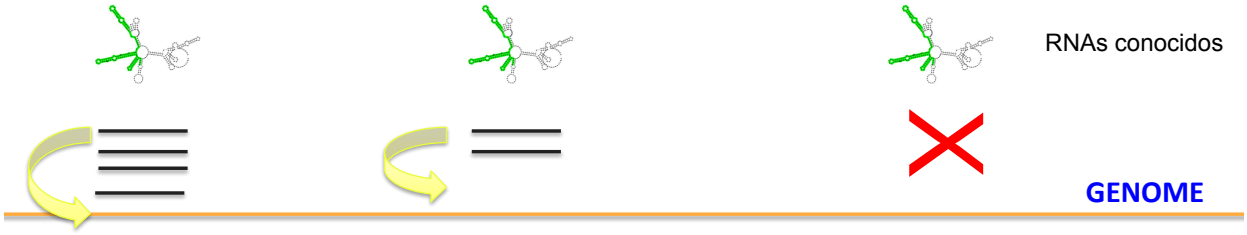
de estructuras secundárias

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

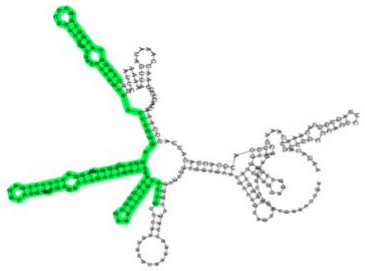
Identificando RNAs a partir de predicción de estructuras secundárias



RNAs conocidos

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

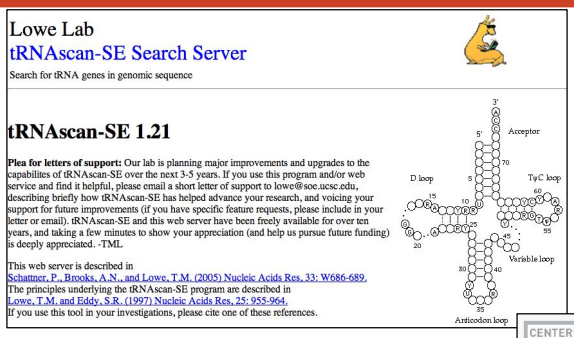


((((……..)))).……..((((((…….))))))

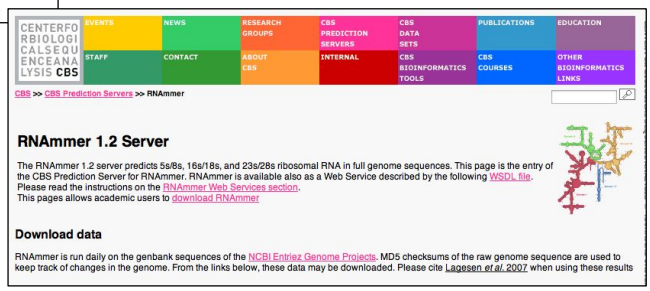
ACGC GCGT
X X
TGGT ACCA
X X
CCTA TAGG

```
>Ec_OxyS,-1-109
aaacggagcggcaccucuuuuaacccuugaagucacugcccguuucgagaguuucucaacucgaauaacuaaagccaacgugaacuuuugcggaucuccaggauccgcu
...((((((((((...((((........))))....)).)))))))).(((((....)))))..........................(((((((...))))))). (-29.40)
>Ec_RhyB,-7-66
aaccugaaagcacgacauugcucacauugcuuccaguauuacuuagccagccgggugcuggcuuuu
...(((.(((((.................))))).))).........(((((((.....)))))))... (-17.40)
>Ec_RprA,-1-106
cgguuauaaaucaacauauugauuuauaagcauggaaaucccugagugaaacaacgaauugcugugugaugcuuugcccaucucccacgaugggcuuuuuuuu
.(.((((((((((....))))))))))..)..((....)).................(((((.....)))))....(((((((......)))))))..... (-24.80)
```

Identificando RNAs a partir de predicción de estructuras secundárias



tRNAs



rRNAs

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

BMC Bioinformatics

SOFTWARE Open Access

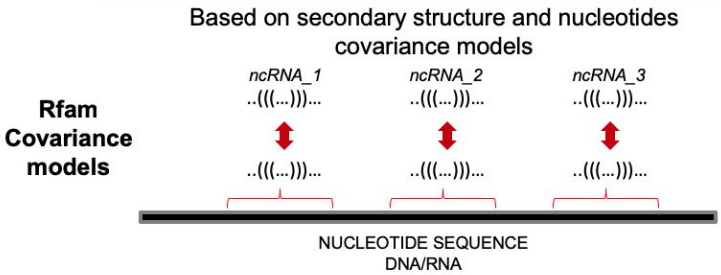
StructRNAfinder: an automated pipeline and web server for RNA families prediction

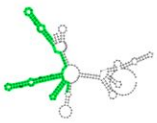
Raúl Arias-Carrasco[1,2†], Yessenia Vásquez-Morán[1†], Helder I. Nakaya[3*] and Vinicius Maracaja-Coutinho[1,4,5,6*]

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 *cis* 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

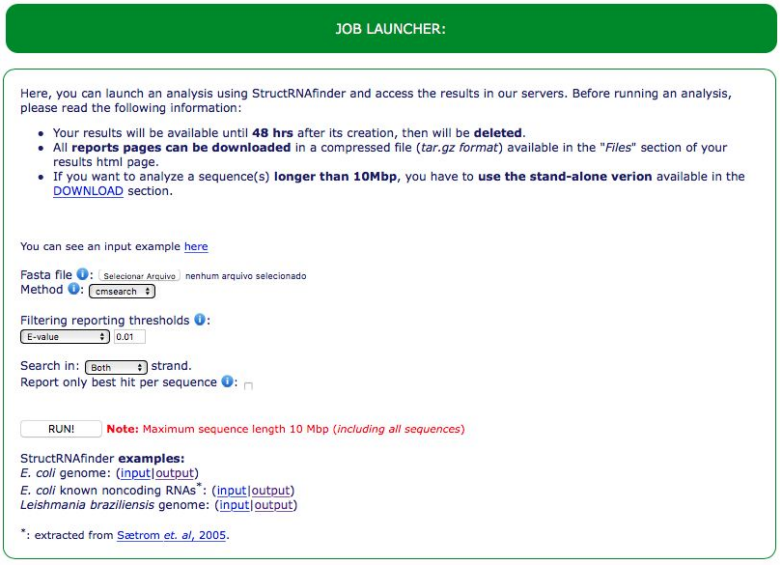
```
StructRNAfinder -i archivo.fasta -d cm
```

-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, cmscan or cmsearch [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: nameinputfile.tab]
-o, --output	direct output to file <f>, not stdout [default: nameinputfile.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):

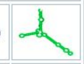
-s, --score	minimun score to each hit [default: 10]
-e, --e-value	maximun e-value to each hit [default: 0.01]

Also available in a user-friendly web server



Struct RNA finder

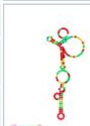
- HOME
- RNA families
  - Cis-reg
    - Others
    - frameshift_element
    - IRES
    - leader
    - riboswitch
    - thermoregulator
  - Gene
  - Intron
- Summary
- Loci distribution
- Taxonomy
- Files
- Contact

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

Struct RNA finder

gi_545778205_gb_U000_273

cmsearch-Rfam

Family : sroC	**Id:** RF00369	**Type:** Gene; sRNA;
Taxonomy: Eukaryota; Bacteria;	**Ontology:** Not found	
Description: This family consists of a number of bacterial sroC RNA genes of around 160 bases in length. sroC is found in several Enterobacterial species but its function is unknown [1].		
Score: 174.8	**E-value:** 1e-38	
From sequence: 686843	**To sequence:** 686681	

Alignments:

```
                                                                                                        NC
                               ::::::::::(((,,,,<<<<<<_____>>->>>><<<<<<<<<<-<<------<<<-<<<-------<<<<<<<<<_____ CS
            RF00369          1 ACUAAuUAaAAccaaaagGGGcGGgAuuUCCugCCCuCuCGauuGuugcucAGCACGGACagACUAUAaGccgGgUGGUCGUU 83
                               ACUAAUUA AA::A+ AGGGGCGG+A+UUCC GCCCUCUCGAUUGUU:C +AGCACGGACAGACUAUA+GCC:G:UGGUCGUU
gi_545778205_gb_U000 686843 ACUAAUUACAAGAACCAGGGGCGGAAAUUCCAGCCCUCUCGAUUGUUACGUAGCACGGACAGACUAUACGCCUGAUGGUCGUU 686761
                               ************************************************************************************ PP

                                                                                                     NC
                               __>>>>>>>>>>>>----->>>------------>>->>>>>--->>>>>,,,,,,,,,,,,,,,,,,,,,,,,,,,))): CS
            RF00369         84 CCCCAcCcggCcuGAAaACCGcAaaacGCuGggcAaCaauCUUCGaGgGUAGCgcUgcuacCcuUUUUUUcuuagugga 163
                               CCCCA:C:GGCCUGAAAACCGCAA+ACGCUGGG:AACAAUCUUCGAGGGUAGC+ U +   C+  U    UCUU++U::
gi_545778205_gb_U000 686760 CCCCAUCGGGCCUGAAAACCGCAAUACGCUGGGUAACAAUCUUCGAGGGUAGCAGUUAACGCUGCUACCCUCUUUUUUCU 686681
                               ********************************************************************************* PP
```

RNAfold

>gi_545778205_gb_U000_273 GAAAAAAGAGGGUAGCAGCGUUAACUGCUACCCUCGAAGAUUGUUACCCA GCGUAUUGCGGUUUUCAGGCCCGAUGGGGAACGACCAUCAGGCGUAUAGU CUGUCCGUGCUACGUAACAAUCGAGAGGGCUGGAAUUUCCGCCCCUGGUU CUUGUAAUUAGU .......((((((((((......))))))))))...((((((((..((((....((((.....(.(((.(((((.......))))).))).)......))))...)))).)))))))))....((((.(((...))))))(((((.....))))). (-63.60)	
Length sequence : 162	**Length match:** 161

Struct RNA finder

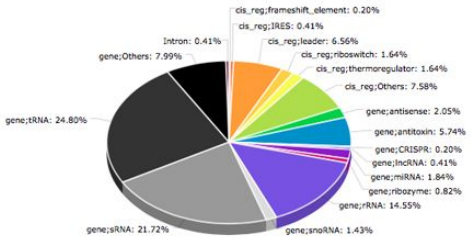
Struct RNA finder

HOME
– RNA families
- Cis-reg
Others
frameshift_element
IRES
leader
riboswitch
thermoregulator
+ Gene
Intron
Summary
Loci distribution
Taxonomy
Files
Contact

Summary RNAs annotated

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



Struct RNA finder

Loci distribution overview

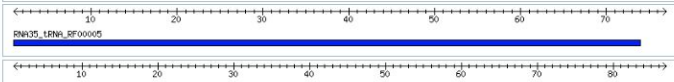
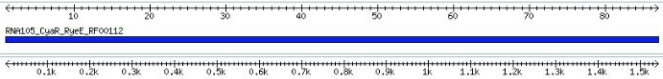
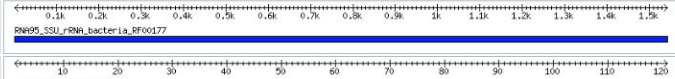
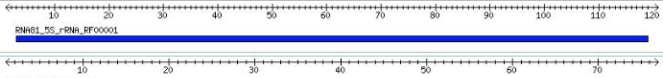
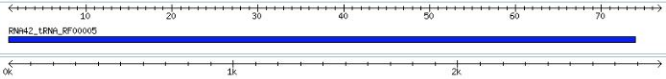
LIB

Struct RNA finder

HOME
– RNA families
- Cis-reg
Others
frameshift_element
IRES
leader
riboswitch
thermoregulator
+ Gene
Intron
Summary
Loci distribution
Taxonomy
Files
Contact

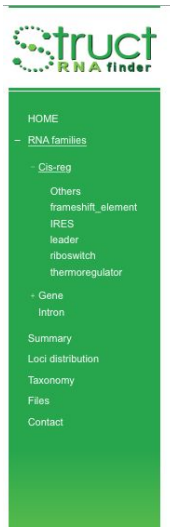
Loci distribution overview

Positive strand
Negative strand

Sequence	Image
asnT	
gltU	
RyeE,	
rrsB	
rrfC	
glyX	
rrlA	

Struct RNA finder

Different file formats useful for downstream analyses



Files:

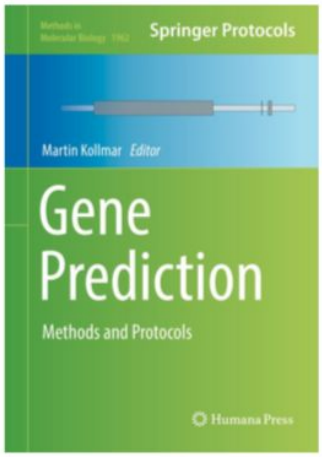
Infernal	new_e_coli.out new_e_coli.tab
StructRNAfinder Annotation	new_e_coli_filtered.tab
Bed format	new_e_coli_filtered.bed
FASTA Sequence	new_e_coli_mature.fasta
RNAfold	new_e_coli_RNAfold.dbn
All Reports	Download web page

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

http://structrnafinder.integrativebioinformatics.me

Arias-Carrasco et al, 2018
BMC Bioinformatics

A guide on how to use StructRNAfinder



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 dispobinle

BOW TIE

HISAT2

BLAST

UCSC Genome Bioinformatics

BLAT

Transcriptoma con genoma disponible

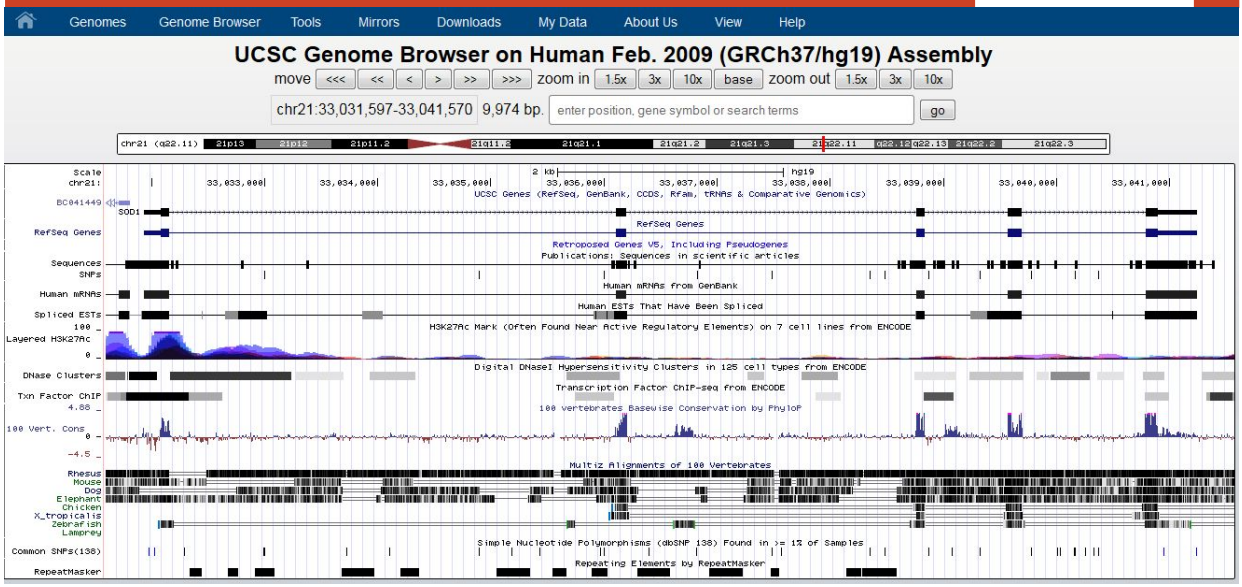
Caracterización usando manipulación de coordenadas genómicas

Coordenadas genómicas son asignadas a genomas

- chr21:33.031.597-33.041.570
- ¿Qué son estas coordinadas?

  UCSC Genome Browser (http://genome.ucsc.edu)

Transcriptoma con genoma dispobinle



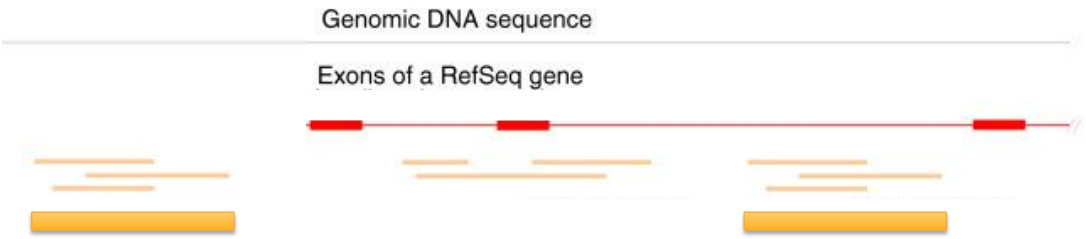
Transcriptoma con genoma disponible

Caracterizar su contexto genómico

Caracterización usando manipulación de coordenadas genómicas

Transcriptoma con genoma dispobinle

Ensamblar su transcrito por overlapping de coordinadas

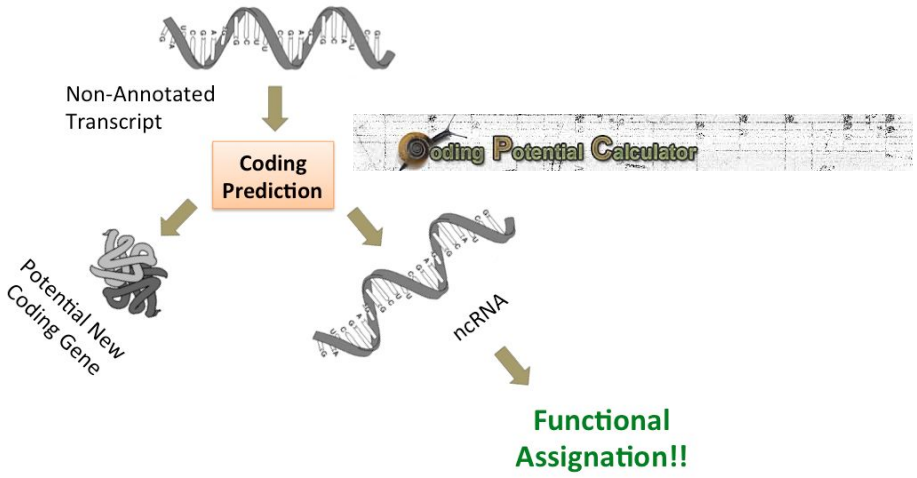


Caracterización usando manipulación de coordinadas genómicas

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

LIB

shortcat

Aliaga-Tobar et al
To be submitted

Automatically annotate ncRNAs according to user-provided references

LIB

shortcat

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

NRDR Non-coding RNA Databases Resources

NR2 Non Redundant Non-coding RNA Sequences

miRBase

Struct RNA finder

Aliaga-Tobar et al
To be submitted

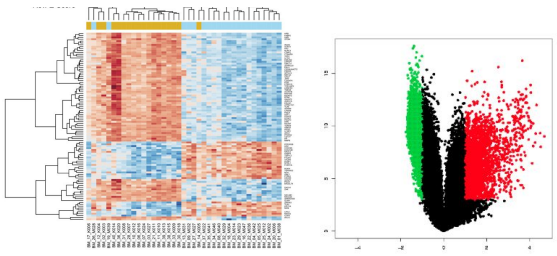
Expression quantification

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[1,2,*], Vanessa Tripp[1], Victor Aliaga-Tobar[3], Sonja-Verena Albers[4], Vinicius Maracaja-Coutinho[2,5] and Lennart Randau[1,*]

genes

MDPI

Article

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

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

ELSEVIER

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

Regular article

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

José Manuel Matamala [a, b, c, d], Raul Arias-Carrasco [d], Carolina Sanchez [d], Markus Uhrig [d], Leslie Bargsted [a, b, f], Soledad Matus [a, f, g], Vinicius Maracaja-Coutinho [d], Sebastian Abarzua [e], Brigitte van Zundert [e], Renato Verdugo [c], Patricio Manque [d], Claudio Hetz [a, b, f, h, i]

DATABASE

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[1,2], Raúl Arias-Carrasco[3], José C. Caris-Maldonado[3], Aldina Barral[1,4,5], Vinicius Maracaja-Coutinho[3,6,7,*] and Artur T. L. De Queiroz[1,2,5]

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[1] · Angelica Cardoso Carlos[1] · Vinicius Maracaja-Coutinho[2,3,4] · Sávio Torres de Farias[1]

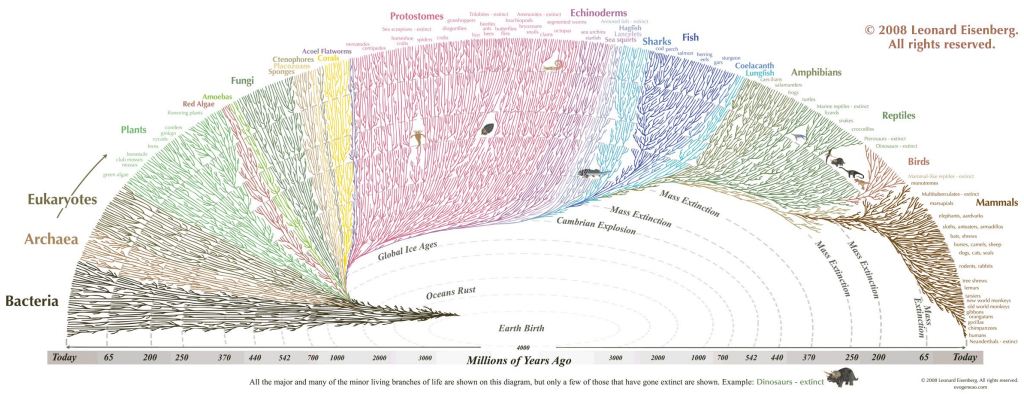
PLOS ONE

RESEARCH ARTICLE

Core non-coding RNAs of *Piscirickettsia salmonis*

Cristopher Segovia[1,2], Raul Arias-Carrasco[2,3], Alejandro J. Yañez[4], Vinicius Maracaja-Coutinho[3,5], Javier Santander[1]*

Non-coding RNAs, complexity, disease and evolution



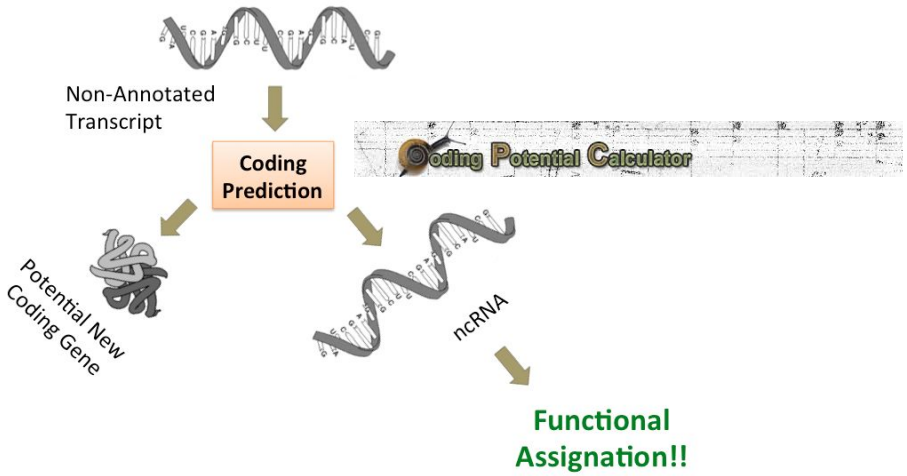
Virus: Hepesvirus; Zika Virus
Archaea: *+250 genomes*
Bacteria: *Piscirickettsia salmonis*
Plants: *Zea mays Lluteño, Cannabis sativa*

Protozoa: *Leishmania ssp.*
Platyhelminthes: *Schistosoma mansoni*
Insecta: *Drosophila melanogaster;*
Apis melifera

Transcriptoma sin genoma disponible – De Novo

Transcriptoma sin genoma dispobinle – De Novo

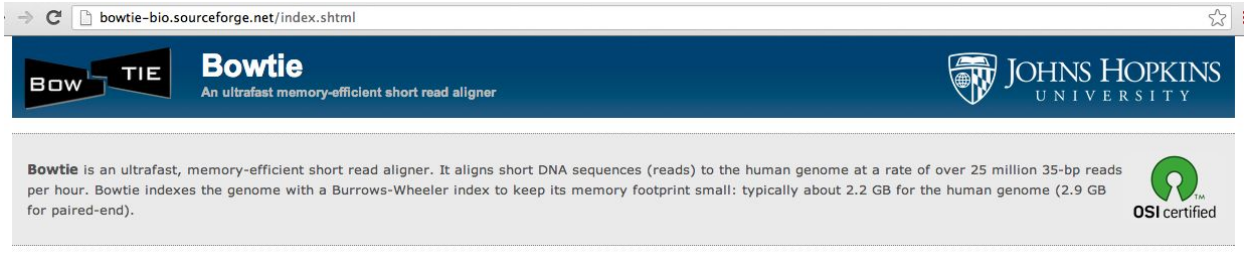
Predicción de su potencial codificante



Asignación funcional a ARNs no codificantes por medio de métodos computacionales

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

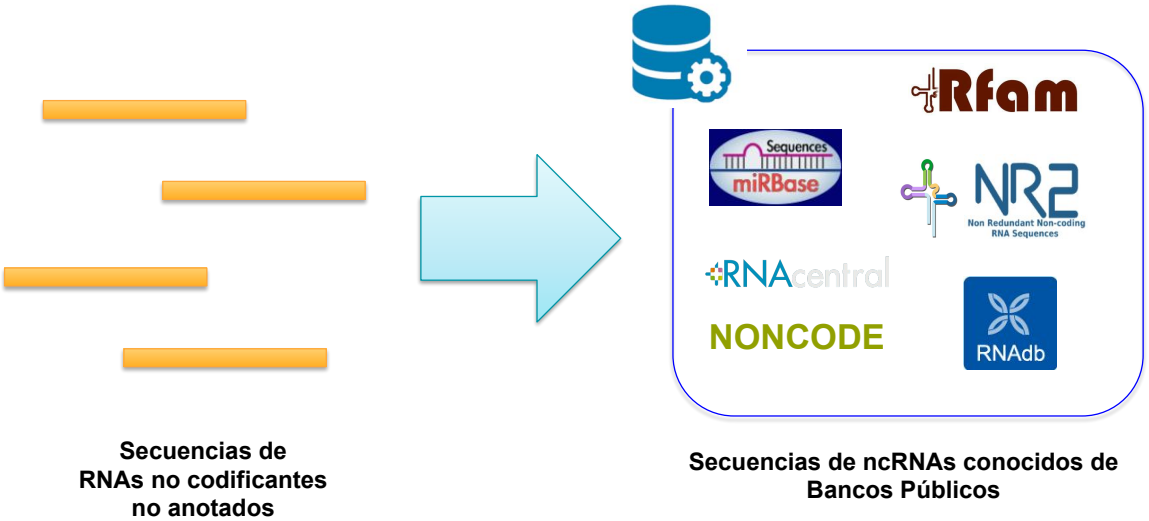
Asignando función a RNAs por medio de comparaciones de secuencias



UCSC Genome Bioinformatics
BLAT

BLASTn

Asignando función a RNAs por medio de comparaciones de secuencias



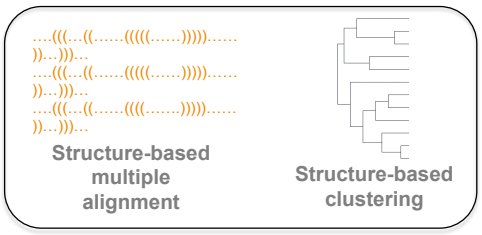
Secuencias de RNAs no codificantes no anotados

Secuencias de ncRNAs conocidos de Bancos Públicos

Asignación funcional a ARNs no codificantes por medio de predictores específicos

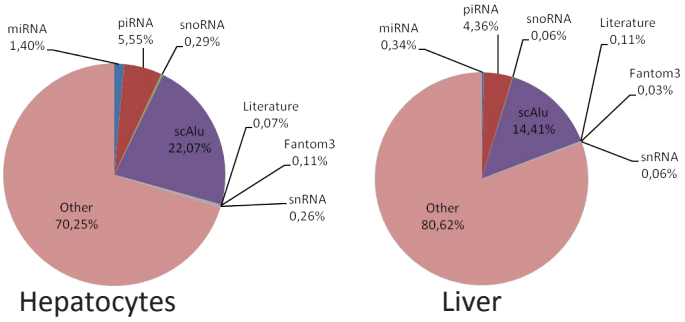
Asignando función a RNAs por medio de otros predictores específicos

SnoReport
HHMMIR
MIPRED
Novomir
Triplet



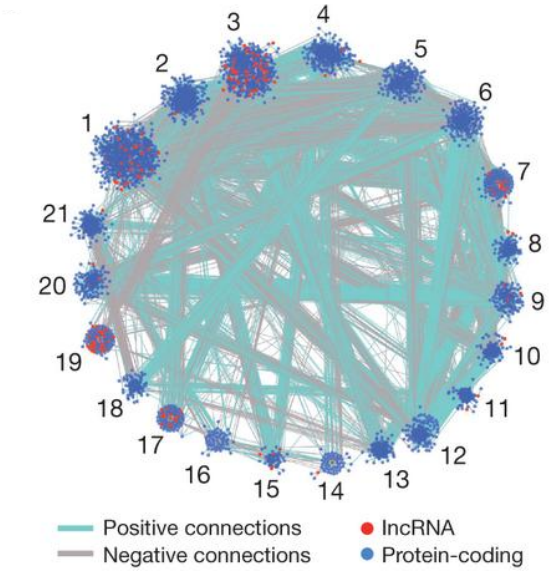
Asignando función a RNAs por medio de co-expresión de transcritos

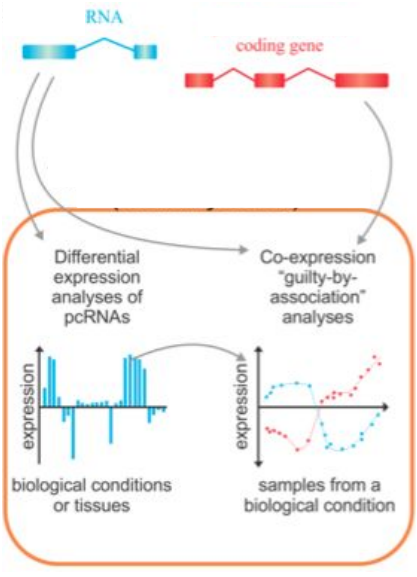
Todavía es muy difícil asignar función a ARNs no codificantes a partir de características en su secuencias



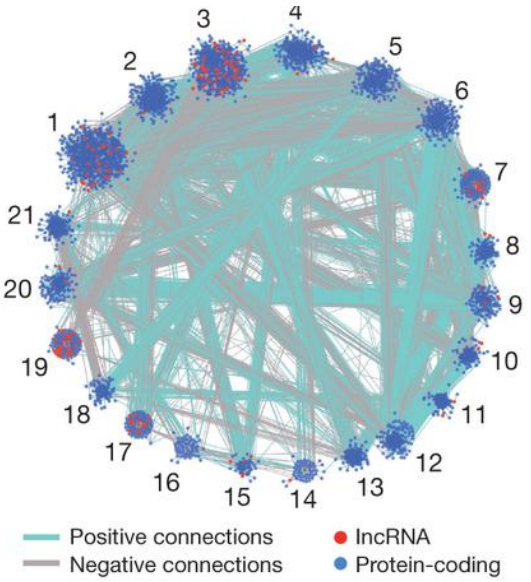
Maracaja-Coutinho, 2013
Tesis Doctoral

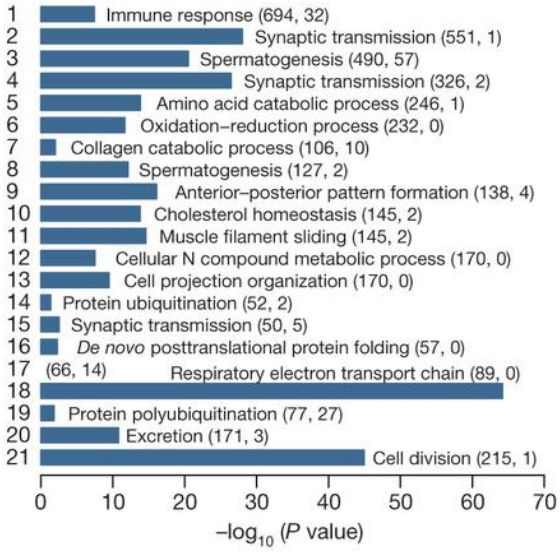
Asignando función a RNAs por medio de co-expresión de transcritos





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

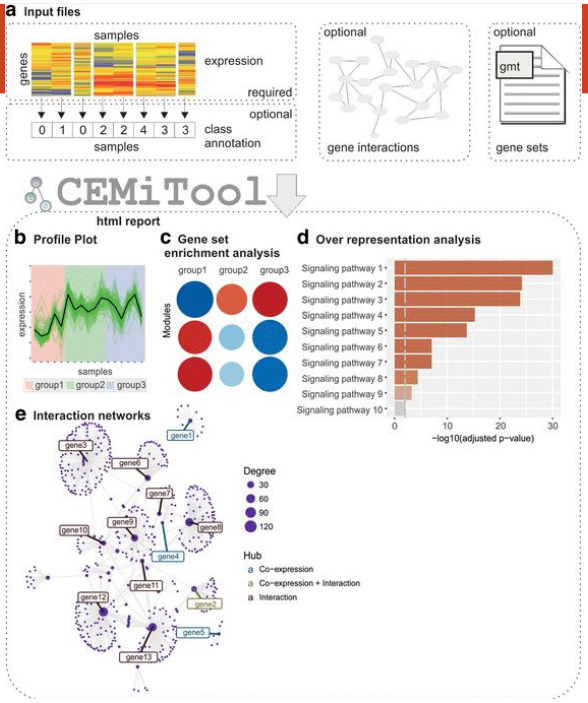
BMC Bioinformatics

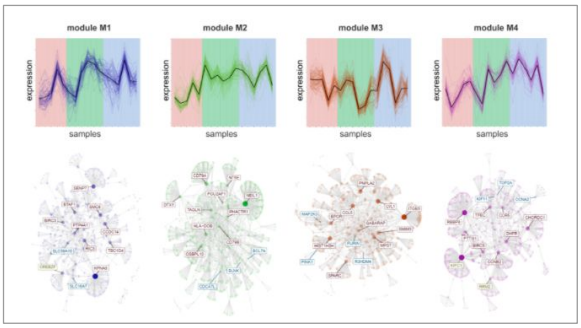
SOFTWARE Open Access

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

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

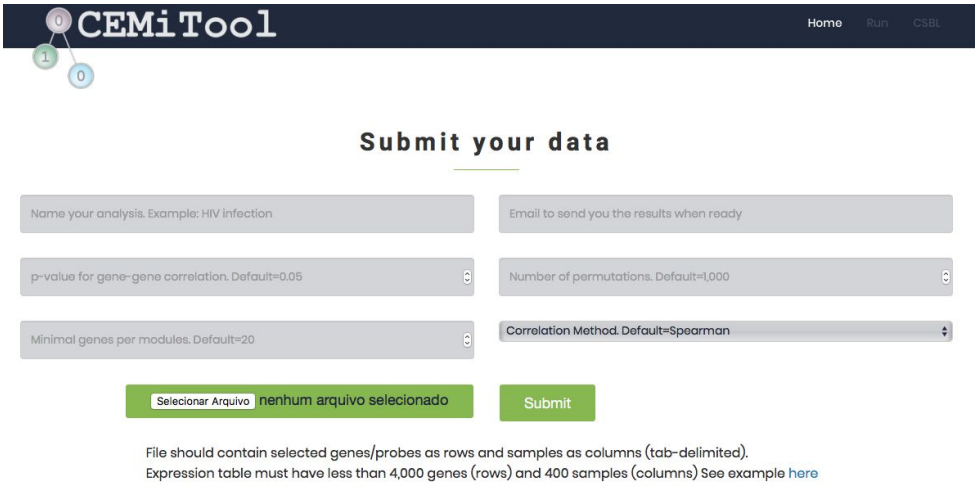
Russo et al., 2018
BMC Bioinformatics





Russo et al., 2018
BMC Bioinformatics

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

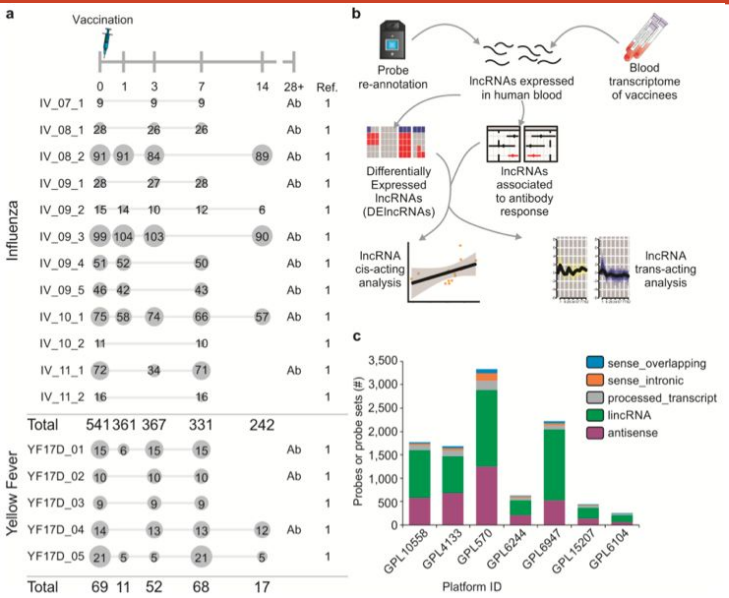


Russo et al. 2019
Frontiers in Genetics

LncRNAs associated to vaccine-induced immunity

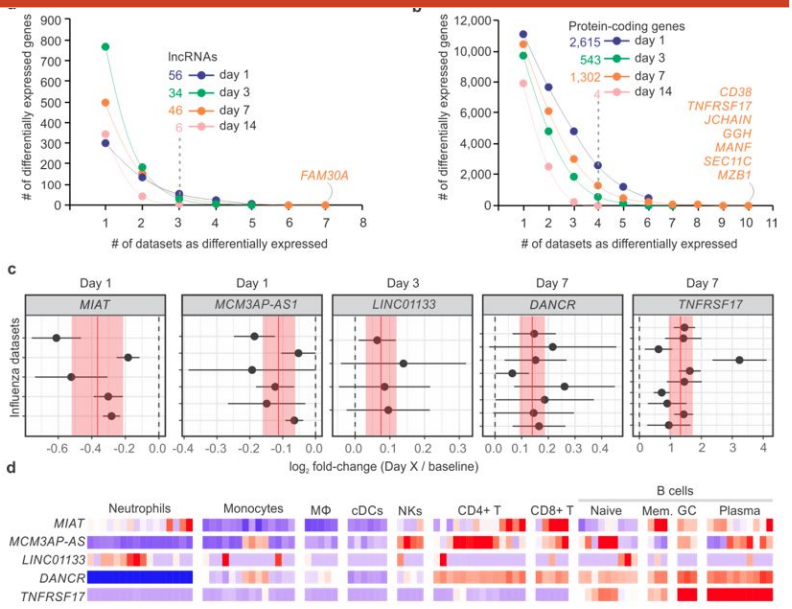
12 cohorts

5 cohorts



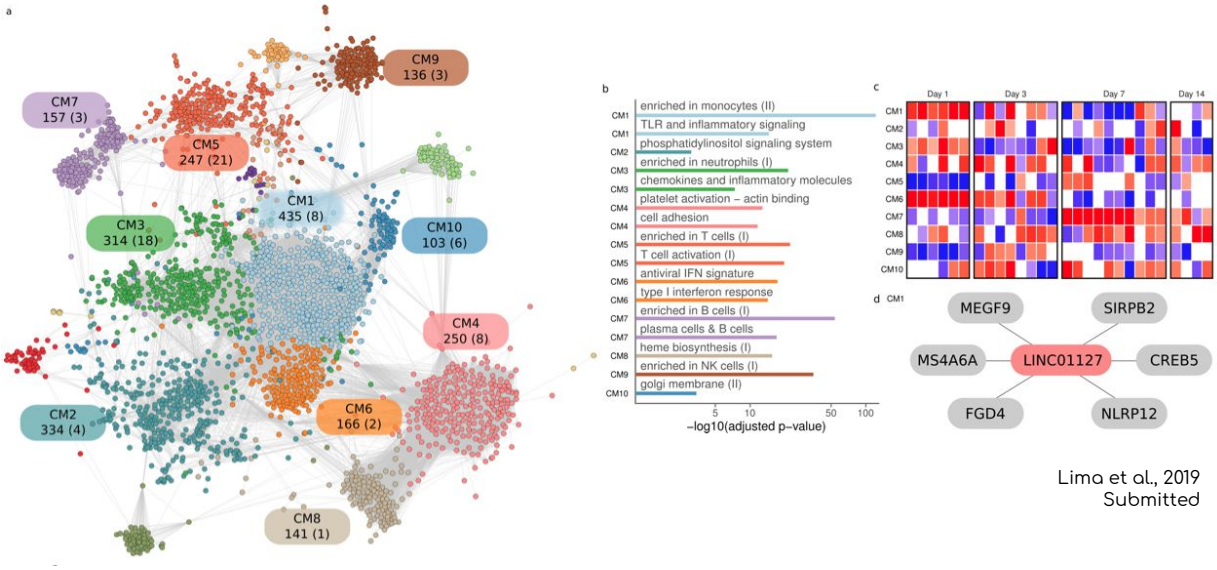
Lima et al., 2019
Submitted

LncRNAs associated to vaccine-induced immunity



Lima et al., 2019
Submitted

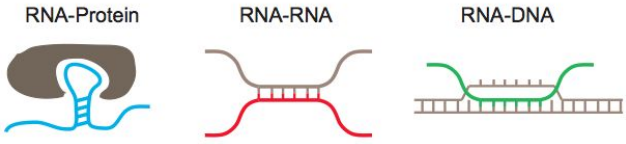
LncRNAs associated to vaccine-induced immunity



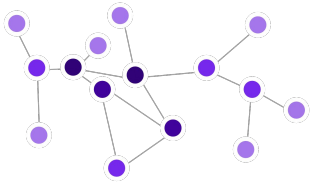
Lima et al., 2019
Submitted

Es una larga jornada....

Interacción con otras moléculas



Redes de interacción



Regulación

THANK YOU!

@vin_maracaja

vinicius.maracaja@uchile.cl