RNAct: Protein-RNA interactions at Your Fingertips

Iliana Crespin and Marcus Avila
BIOL 367
Loyola Marymount University
October 1, 2019

Objectives to familiarize with RNAct

- Knowing the general information of the database
  - Type, contributors, fundings, and more
- Understanding the scientific quality of the database
  - Comprehensiveness of Content, Possibilities with Database & Similar Works, Updates
- How database has a general utility to the Scientific Community
  - Different databases, Convenience, User-friendliness
- Summary of key points of the database
  - Overall usage, Future works

Objectives to familiarize with RNAct

- **Knowing the general information of the database**
    - **Type, contributors, fundings, and more**
- Understanding the scientific quality of the database
    - Comprehensiveness of Content, Possibilities with Database & Similar Works, Updates
- How database has a general utility to the Scientific Community
    - Different databases, Convenience, User-friendliness
- Summary of key points of the database
    - Overall usage, Future works

RNAct is a Protein-RNA Interactome

- Interactome: a whole set of physical interactions among proteins
  - Mapping protein interactions
- Combines “experimentally identified” interactions with predictions
- Uses catRAPID algorithm
  - Predicts protein-RNA interactions with 89% accuracy
  - Helps with the identification of these interactions
- Metabolic/Signalling
  - Involves genomes of mouse, human, and yeast
  - Secondary source that is electronically curated

Created by Tartaglia Lab

- Private organization that focuses on RNA molecules
  - Research in understanding protein homeostasis
- Funded by Horizon 2020
  - EU research and innovation program (2014-2020)
  - Main security for Europe's global competitiveness

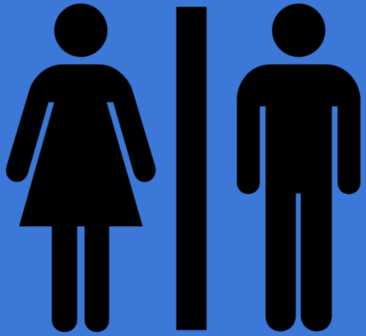
(Tartaglia Lab, 2012)

Objectives to familiarize with RNAct

- Knowing the general information of the database
  - Type, contributors, fundings, and more
- **Understanding the scientific quality of the database**
  - **Comprehensiveness of Content, Possibilities with Database & Similar Works, Updates**
- How database has a general utility to the Scientific Community
  - Different Databases, Convenience, User-friendliness
- Summary of key points of the database
  - Overall usage, Future works

Comprehensiveness of Content

- 5963 ***Saccharomyces cerevisiae*** proteins, 20778 ***Homo sapien*** proteins, and 22080 ***Mus musculus***
- Studying **gene regulatory events** and **networks at the post-transcriptional level**



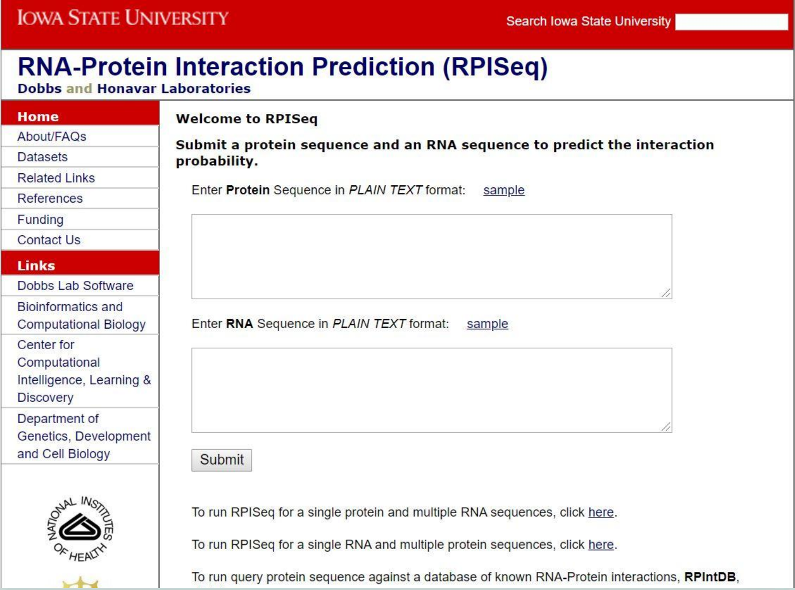
(Wikipedia, 2007)

(Wikipedia, 2015)

(Flickr, 2008)

Possibilities with Database & Similar Works

- Predicting the RNA-protein interactions, along with RNA-RNA and protein-protein interactions, to create a **map of molecular interactions** in an organism
- Other RNA-protein interaction databases: **PRD & PRIDB**



RNA-Protein Interaction Prediction (RPISeq)

Dobbs and Honavar Laboratories

Welcome to RPISeq

Submit a protein sequence and an RNA sequence to predict the interaction probability.

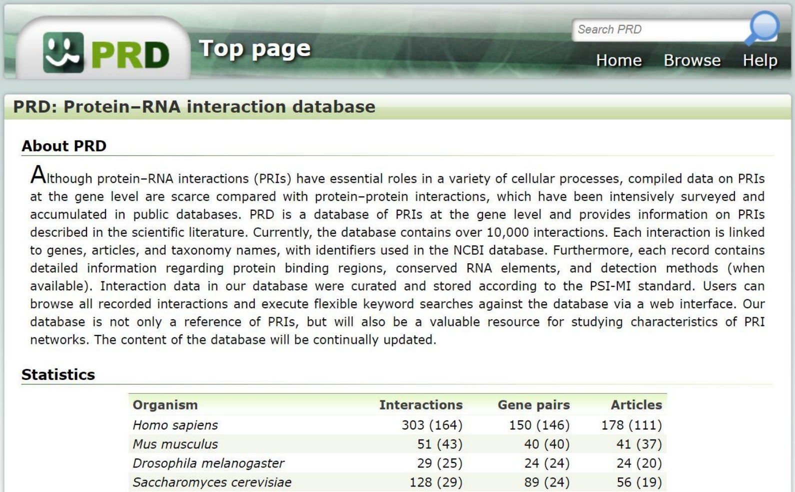
Enter **Protein** Sequence in *PLAIN TEXT* format: sample

Enter **RNA** Sequence in *PLAIN TEXT* format: sample

To run RPISeq for a single protein and multiple RNA sequences, click here.

To run RPISeq for a single RNA and multiple protein sequences, click here.

To run query protein sequence against a database of known RNA-Protein interactions, **RPIntDB**,



PRD: Protein–RNA interaction database

About PRD

Although protein–RNA interactions (PRIs) have essential roles in a variety of cellular processes, compiled data on PRIs at the gene level are scarce compared with protein–protein interactions, which have been intensively surveyed and accumulated in public databases. PRD is a database of PRIs at the gene level and provides information on PRIs described in the scientific literature. Currently, the database contains over 10,000 interactions. Each interaction is linked to genes, articles, and taxonomy names, with identifiers used in the NCBI database. Furthermore, each record contains detailed information regarding protein binding regions, conserved RNA elements, and detection methods (when available). Interaction data in our database were curated and stored according to the PSI-MI standard. Users can browse all recorded interactions and execute flexible keyword searches against the database via a web interface. Our database is not only a reference of PRIs, but will also be a valuable resource for studying characteristics of PRI networks. The content of the database will be continually updated.

Statistics

Organism	Interactions	Gene pairs	Articles
Homo sapiens	303 (164)	150 (146)	178 (111)
Mus musculus	51 (43)	40 (40)	41 (37)
Drosophila melanogaster	29 (25)	24 (24)	24 (20)
Saccharomyces cerevisiae	128 (29)	89 (24)	56 (19)

Objectives to familiarize with RNAct

- Knowing the general information of the database
    - Type, contributors, fundings, and more
- Understanding the scientific quality of the database
    - Comprehensiveness of Content, Possibilities with Database & Similar Works, Updates
- **How database has a general utility to the Scientific Community**
    - **Different Databases, Convenience, User-friendliness**
- Summary of key points of the database
    - Overall usage, Future works

Different Databases Used in RNAct

- Extracts from **UniProt**, **Ensembl**, **GENCODE**, **ENCODE**, and **NCBI**
- Has over 5.87 billion pairwise interactions

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

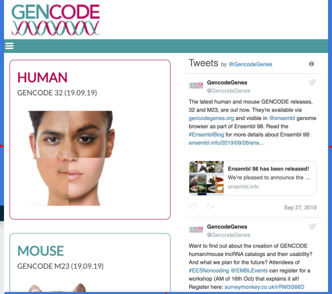
Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects



HUMAN
GENCODE 32 (19.09.19)

MOUSE
GENCODE M23 (19.09.19)

Tweets by @GencodeGenes

GencodeGenes
@GencodeGenes

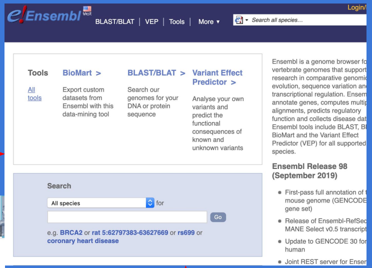
The latest human and mouse GENCODE releases, 32 and M23, are out now. They're available via gencodegenes.org and visible in @ensembl genome browser as part of Ensembl 98. Read the #EnsemblBlog for more details about Ensembl 98: ensembl.info/2019/09/26/ens…

Ensembl 98 has been released!
We're pleased to announce the …
ensembl.info

Sep 27, 2019

GencodeGenes
@GencodeGenes

Want to find out about the creation of GENCODE human/mouse lncRNA catalogs and their usability? And what we plan for the future? Attendees of #EESNoncoding @EMBLEvents can register for a workshop (AM of 16th Oct) that explains it all! Register here: surveymonkey.co.uk/r/RM3S66D



Tools
All tools

BioMart >
Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >
Search our genomes for your DNA or protein sequence

Variant Effect Predictor >
Analyse your own variants and predict the functional consequences of known and unknown variants

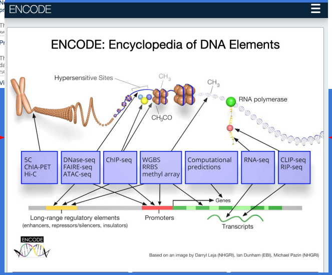
Search
All species for
Go

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

Ensembl is a genome browser fo vertebrate genomes that support research in comparative genomi evolution, sequence variation an transcriptional regulation. Ensem annotate genes, computes multip alignments, predicts regulatory function and collects disease dat Ensembl tools include BLAST, B BioMart and the Variant Effect Predictor (VEP) for all supported species.

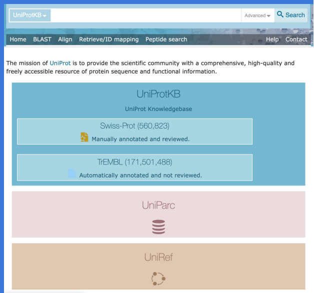
Ensembl Release 98 (September 2019)

- First-pass full annotation of mouse genome (GENCODE gene set)
- Release of Ensembl-RefSe MANE Select v0.5 transcript
- Update to GENCODE 30 fo human
- Joint REST server for Ense



ENCODE: Encyclopedia of DNA Elements

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB
UniProt Knowledgebase

Swiss-Prot (560,823)
Manually annotated and reviewed.

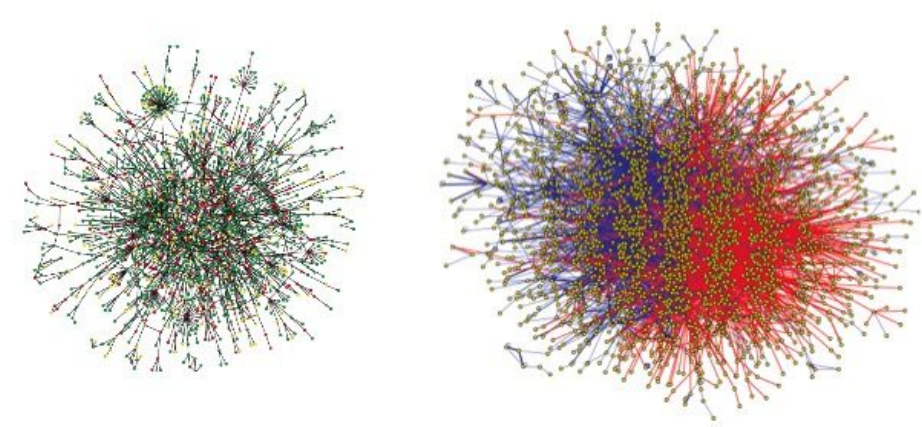
TrEMBL (171,501,488)
Automatically annotated and not reviewed.

UniParc

UniRef

Convenient for a Specific Audience

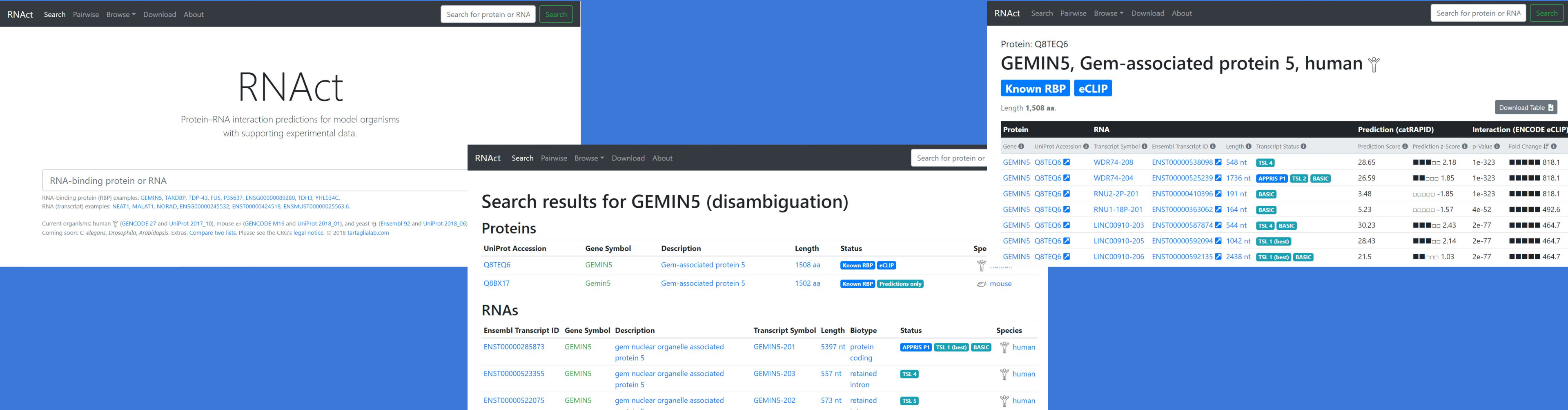
- Focusing on protein-RNA interactions
- Not solely based on experimental data
  - Contains predictions
- Not convenient for those who are uncertain on what they are looking for



(EMBL-EBI, 2018)

Database User-Friendliness

- **Home page** is clear & includes **example terms** to help user
- **Disambiguation list** of proteins and RNAs allows for specification
- **Prediction score** with **p-value** to inform user about accuracy
- **Fold change** to provide experimental interaction evidence



RNAct

Protein–RNA interaction predictions for model organisms
with supporting experimental data.

RNA-binding protein or RNA

RNA-binding protein (RBP) examples: GEMIN5, TARDBP, TDP-43, FUS, P35637, ENSG00000089280, TDH3, YHL034C.
RNA (transcript) examples: NEAT1, MALAT1, NORAD, ENSG00000245532, ENST00000424518, ENSMUST00000025563.6.

Current organisms: human (GENCODE 27 and UniProt 2017_10), mouse (GENCODE M16 and UniProt 2018_01), and yeast (Ensembl 92 and UniProt 2018_06)
Coming soon: *C. elegans*, *Drosophila*, *Arabidopsis*. Extras: Compare two lists. Please see the CRG's legal notice. © 2018 tartaglialab.com

Search results for GEMIN5 (disambiguation)

Proteins

UniProt Accession	Gene Symbol	Description	Length	Status	Spe
Q8TEQ6	GEMIN5	Gem-associated protein 5	1508 aa	Known RBP eCLIP	
Q8BX17	Gemin5	Gem-associated protein 5	1502 aa	Known RBP Predictions only	mouse

RNAs

Ensembl Transcript ID	Gene Symbol	Description	Transcript Symbol	Length	Biotype	Status	Species
ENST00000285873	GEMIN5	gem nuclear organelle associated protein 5	GEMIN5-201	5397 nt	protein coding	APPRIS P1 TSL 1 (best) BASIC	human
ENST00000523355	GEMIN5	gem nuclear organelle associated protein 5	GEMIN5-203	557 nt	retained intron	TSL 4	human
ENST00000522075	GEMIN5	gem nuclear organelle associated	GEMIN5-202	573 nt	retained	TSL 5	human

Protein: Q8TEQ6

GEMIN5, Gem-associated protein 5, human

Known RBP **eCLIP**

Length **1,508 aa**.

Download Table

Protein		RNA				Prediction (catRAPID)		Interaction (ENCODE eCLIP)	
Gene	UniProt Accession	Transcript Symbol	Ensembl Transcript ID	Length	Transcript Status	Prediction Score	Prediction z-Score	p-Value	Fold Change
GEMIN5	Q8TEQ6	WDR74-208	ENST00000538098	548 nt	TSL 4	28.65	■■■□□ 2.18	1e-323	■■■■■ 818.1
GEMIN5	Q8TEQ6	WDR74-204	ENST00000525239	1736 nt	APPRIS P1 TSL 2 BASIC	26.59	■■□□□ 1.85	1e-323	■■■■■ 818.1
GEMIN5	Q8TEQ6	RNU2-2P-201	ENST00000410396	191 nt	BASIC	3.48	□□□□□ -1.85	1e-323	■■■■■ 818.1
GEMIN5	Q8TEQ6	RNU1-18P-201	ENST00000363062	164 nt	BASIC	5.23	□□□□□ -1.57	4e-52	■■■■■ 492.6
GEMIN5	Q8TEQ6	LINC00910-203	ENST00000587874	544 nt	TSL 4 BASIC	30.23	■■■□□ 2.43	2e-77	■■■■■ 464.7
GEMIN5	Q8TEQ6	LINC00910-205	ENST00000592094	1042 nt	TSL 1 (best)	28.43	■■■□□ 2.14	2e-77	■■■■■ 464.7
GEMIN5	Q8TEQ6	LINC00910-206	ENST00000592135	2438 nt	TSL 1 (best) BASIC	21.5	■■□□□ 1.03	2e-77	■■■■■ 464.7

License Agreement Rules

- **Proper credit** must be given to the creators along with a link to the licence agreement
- May **not** be used in for **commercial purposes**
- **Changes** or **additions must be noted** and proper credit must be given to the creators

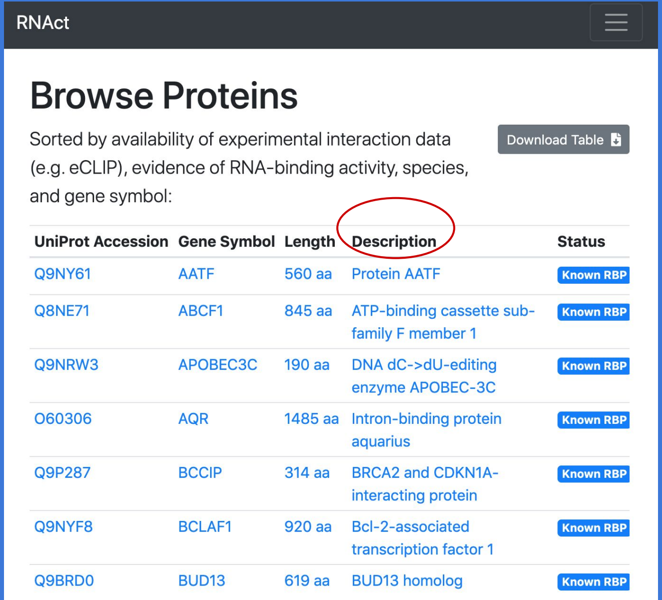
(Creative Commons, 2019)

Objectives to familiarize with RNAct

- Knowing the general information of the database
    - Type, contributors, fundings, and more
- Understanding the scientific quality of the database
    - Comprehensiveness of Content, Possibilities with Database & Similar Works, Updates
- How database has a general utility to the Scientific Community
    - Convenience, User-friendliness, Presentation
- **Summary of key points of the database**
    - **Overall usage, Future works**

Use if Seeking Protein-RNA Interactions

- Can get very confusing if unsure what user is seeking for
- Must understand what each section represents



Browse Proteins

Sorted by availability of experimental interaction data (e.g. eCLIP), evidence of RNA-binding activity, species, and gene symbol:

UniProt Accession	Gene Symbol	Length	Description	Status
Q9NY61	AATF	560 aa	Protein AATF	Known RBP
Q8NE71	ABCF1	845 aa	ATP-binding cassette sub-family F member 1	Known RBP
Q9NRW3	APOBEC3C	190 aa	DNA dC->dU-editing enzyme APOBEC-3C	Known RBP
O60306	AQR	1485 aa	Intron-binding protein aquarius	Known RBP
Q9P287	BCCIP	314 aa	BRCA2 and CDKN1A-interacting protein	Known RBP
Q9NYF8	BCLAF1	920 aa	Bcl-2-associated transcription factor 1	Known RBP
Q9BRD0	BUD13	619 aa	BUD13 homolog	Known RBP

Heading to Something Bigger

- Considered as a professional database
  - Consist of several interactions
- Has room for improvement
  - Planning to cover major model organisms
  - Includes *Drosophila melanogaster*
- Not so complicated to use and will become easier over time
- Opening a path for research on genetic disease risk
- Use predictions to go beyond experimental data

Acknowledgments

Center for Genomic Regulation
Tartaglia Lab
Barcelona, Spain

Dr. Kam Dahlquist

Loyola Marymount University · Founded 1911

https://www.lmu.edu

https://www.crg.eu/

References

- Creative Commons. (2019) Retrieved from https://creativecommons.org/licenses/by-nc-sa/4.0/.
- Lang, B., Armaos, A., & Tartaglia, G. (2017, November). About RNAct. Retrieved from https://rnact.crg.eu/about.
- Mus musculus (2008). Retrieved from https://www.flickr.com/photos/dullhunk/7095792663.
- *Saccharomyces cerevisiae* (2015, December). Retrieved from https://en.wikipedia.org/wiki/Saccharomyces_cerevisiae#/media/File:Saccharomyces_cerevisiae_SEM.jpg.
- Toilets unisex.svg (2007). Retrieved from https://en.wikipedia.org/wiki/File:Toilets_unisex.svg.
- Unknown. (2017). Interactome. Retrieved from https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/interactome.
- Volume 47, Issue D1, 8 January 2019. (2019, January 8). Retrieved from https://academic.oup.com/nar/issue/47/D1.