Freie Universität Berlin

MAX-PLANCK-GESELLSCHAFT

What is RNA Bioinformatics?

Annalisa Marsico
OWL RNA Bioinformatics, MPI Molgen Berlin

High-Throughput Genomics (FU)
14.10.15

Goals of this course - I

Soft skills

- Learn how to evaluate a research paper
- Learn what makes a paper good
- Learn how to get your paper published
- Learn how to give a scientific talk
- Learn to be critical / evaluate

Goals of this course - II

Hard skills

- Get an overview of the RNA bioinformatics field
- Learn how basic concepts / algorithms/ statistical methods are applied and extended in this field
- Learn how to ask the right biological question and choose the right computational methods ‚to solve it‘

Course Design

- **Today** -> overview on the topics, assignment of papers
- Student presentations
  - Each student will choose a paper and will give a presentation
  - One presentation per term (40-50 minutes + 15 minutes questions)
  - Discussion: questions, critical assessmnet. One scientific question per person + 1 good comment and 1 comment on what can be improved

Presentation Guidelines

Compression with minimal loss of information

1. Understand the context & data used
2. Identify the important question/motivation
3. Focus on the method
4. Summarize shortly the main findings
   - Forget about unimportant details
5. Evaluate and think about possible future directions

Advices / Help

- Read your paper twice before saying ‚I don‘t understand it‘
- Read the supplementary material
- Do not try to understand every detail but the general idea has to be clear
- Main objective: lively interesting talk that promotes discussion
- Come anytime to me with questions (write me 3-4 days before)

marsico@molgen.mpg.de
Tel: +49 30 8413 1843
where: MPI for Molecular Genetics, Ihnestrasse 63-73, Room 1.3.07

- send me your presentation one week before your talk
- Get feedback and give feedback (also to me ☺)

What happens if I miss a session?

- Write a small report about the topic you have missed (2 pages – latex)
  - Abstract, Introduciton, Material & Methods, Results & Discussion
  - Re-phrase it in your own words
  - By the 31th of March

- What happens if I miss two sessions?
  - Write two of such reports..

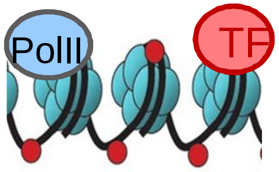
The schedule

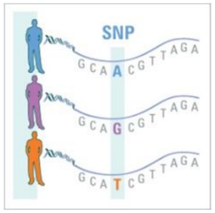
Day	Talk	Topic
October 14	Annalisa	Introduction to Rna Bioinformatics
October 21		
October 28		
November 04		
November 11		
~~November 18~~	~~Annalisa in Köln~~	
November 25		
December 02	backup	
December 09	backup	
December 16	backup	
January 06 (‘16)	backup	

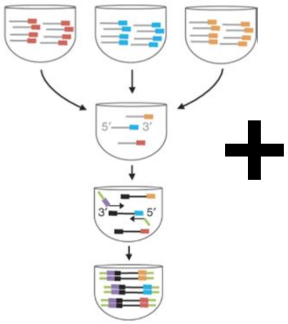
High-throughput genomics

DNA

Chromatin modifications & packing, TF binding, Genetic variation





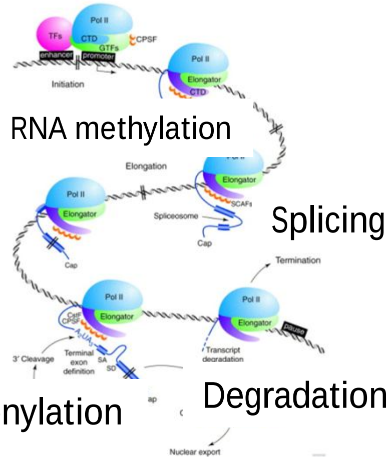


+

Illumina HiSeq 2000

RNA

Synthesis and control, Splicing, Modifications & Degradation



RNA methylation

Splicing

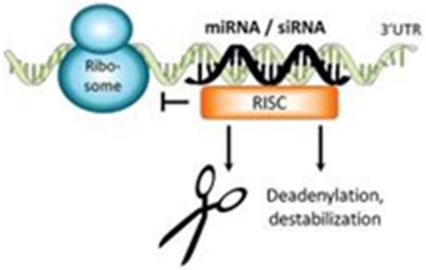
Polyadenylation

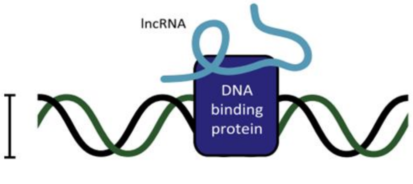
Degradation

(Adapted from Proudfoot et al. 2002)

Non-coding RNA

miRNA transcription, miRNA regulation, lincRNA

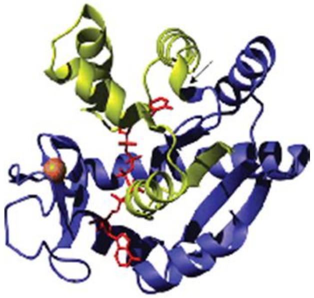




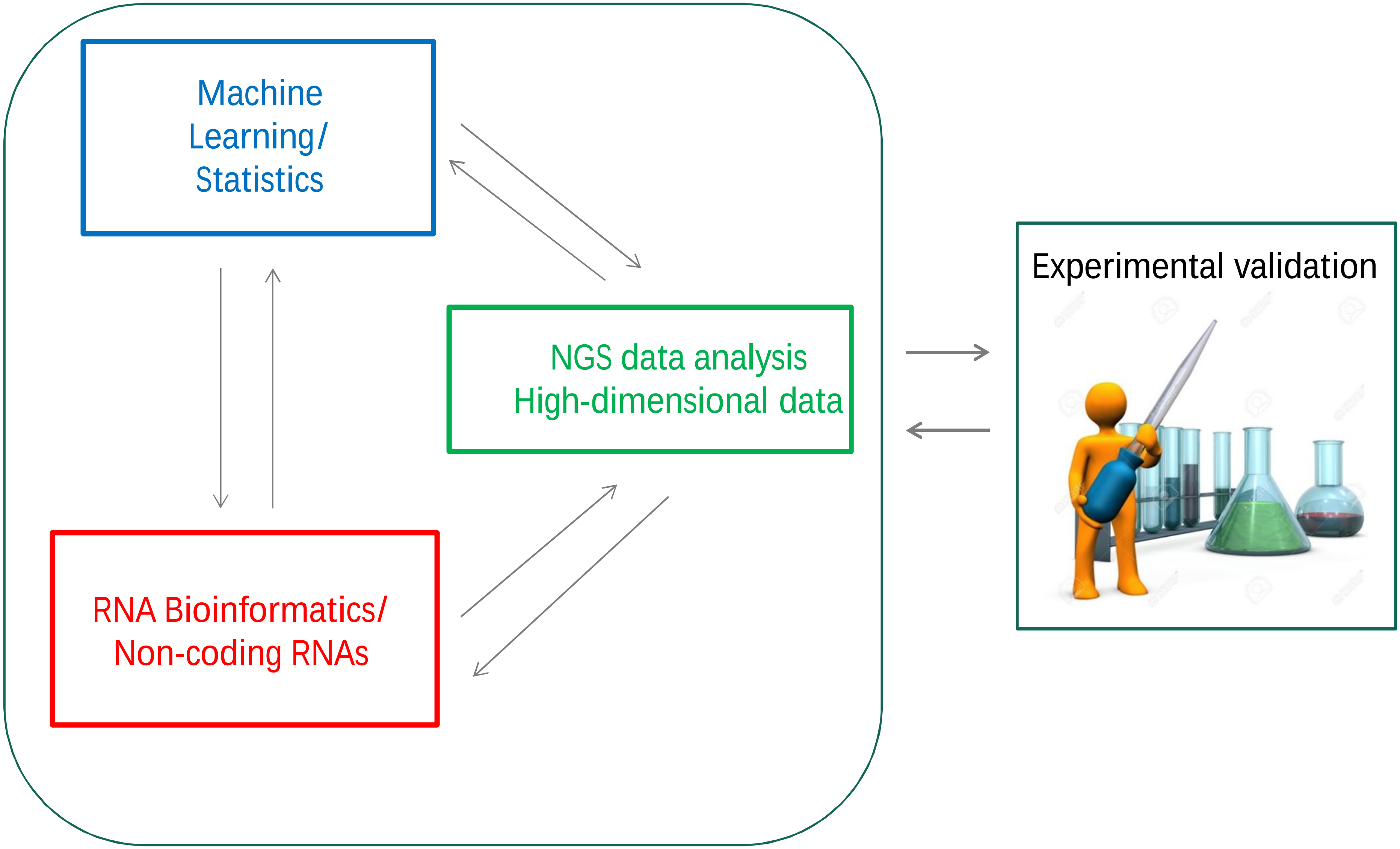
Indirect RNA–protein–DNA associations

Protein

Synthesis, Modifications, Function



Focus of the RNA Bioinformatics group



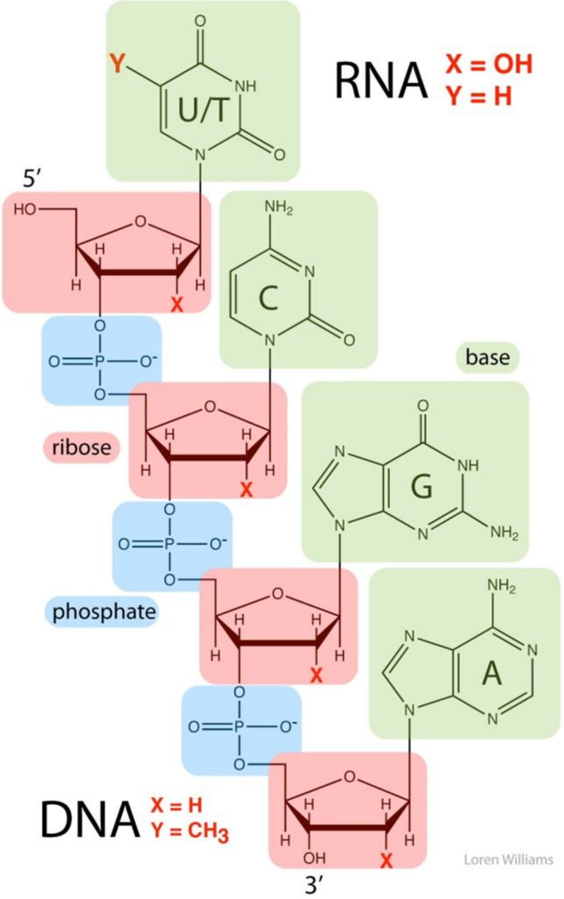
Machine Learning/ Statistics

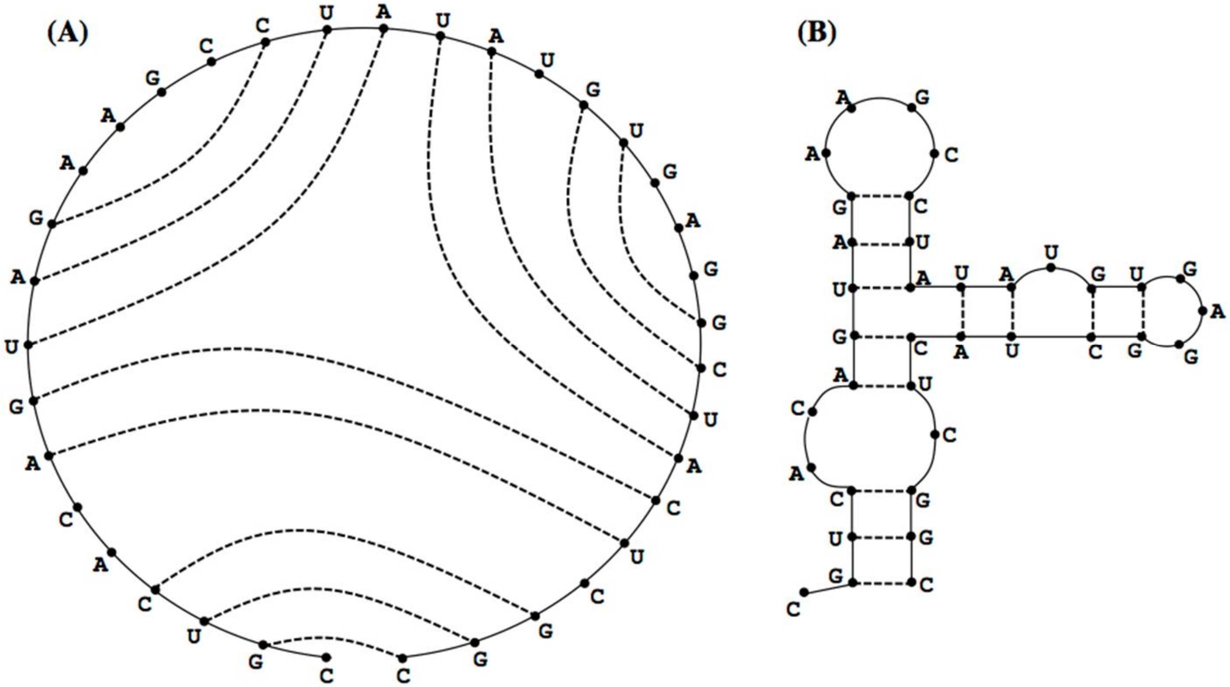
NGS data analysis
High-dimensional data

RNA Bioinformatics/ Non-coding RNAs

Experimental validation

What is an RNA?



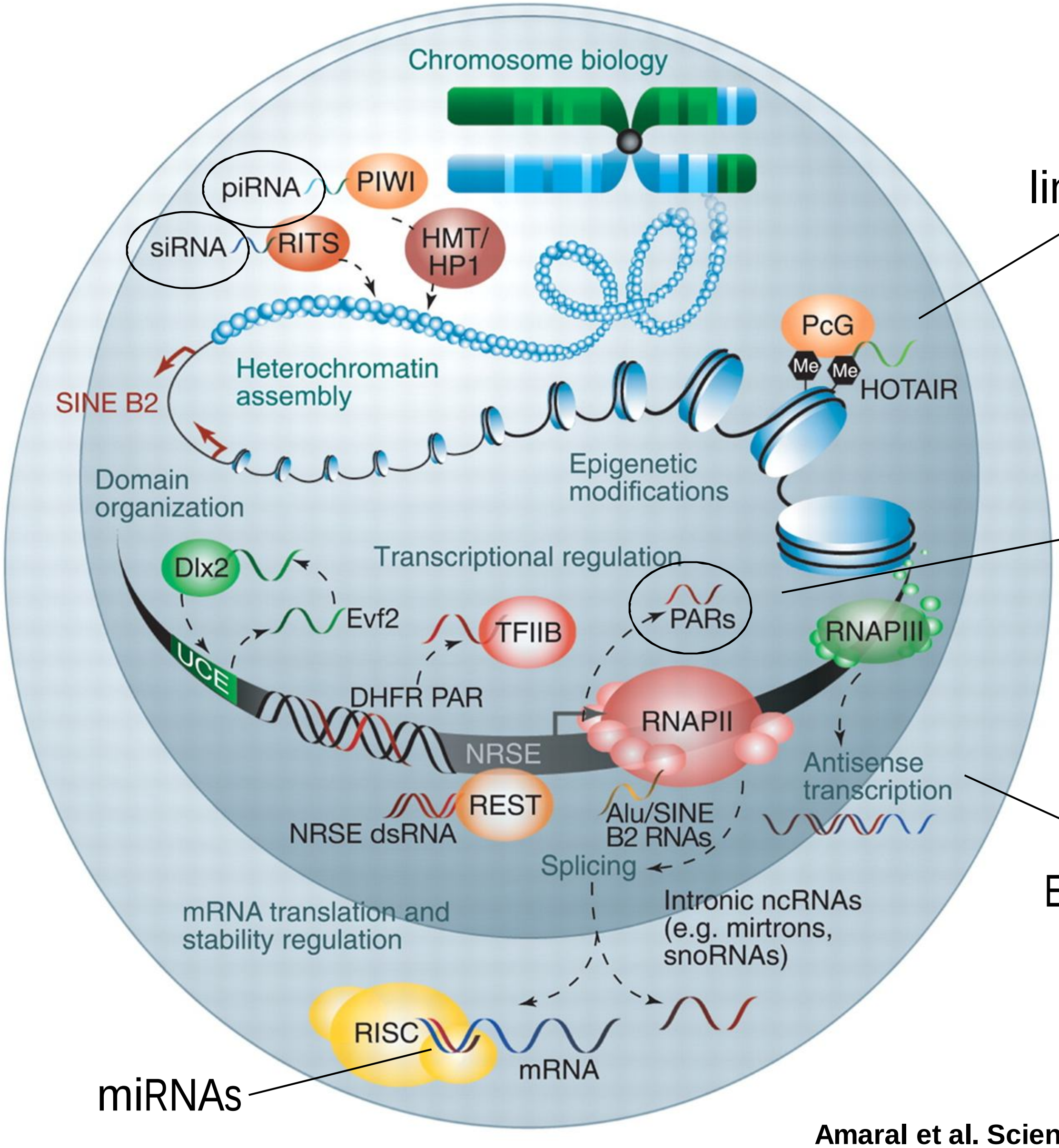

(A)
(B)
C G U C A C A G U A G A A G C C U A U A U G U G A G G C U A C U C G G C

Secondary structure: set of base pairs which can be mapped into a plane

The RNA revolution

- Not only intermediates between DNA and proteins, but informational molecules (enzymes)
- The first primitive form of life? (Woese CR 1967)
- Ability to function as molecular machines (e.g. tRNA, RNAs in splicesosome complex)
- **Ability to to function as regulators of gene expression** (miRNA, sRNAs, piRNAs, lincRNA, eRNAs, ceRNAs..)
- Different sizes and functions (e.g. miRNAs 22nt, lincRNAs > 200nt)
- 1.5 % of the human genome codes for protein, the rest is ‚junk'
- Since ten years **junk has become really important** -> transcribed in ncRNAs
- More than 80% of human disease loci are within non-coding regions
- A lot of tools developed to identify ncRNA genes
- E.g. Rfam – database which collect RNA families and their potential functions

The Eukaryotic Genome as an RNA machine
The ‘RNA world’



lincRNAs

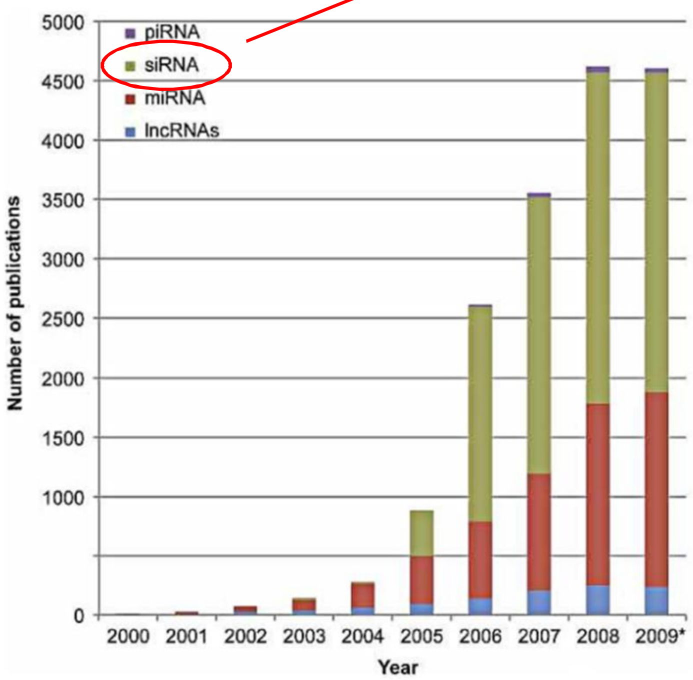
Promoter-associated
RNAs

E(enhancer-like)RNAs

miRNAs

Amaral et al. Science 2008;319:1787-1789

Non-coding RNAs: hot stuff

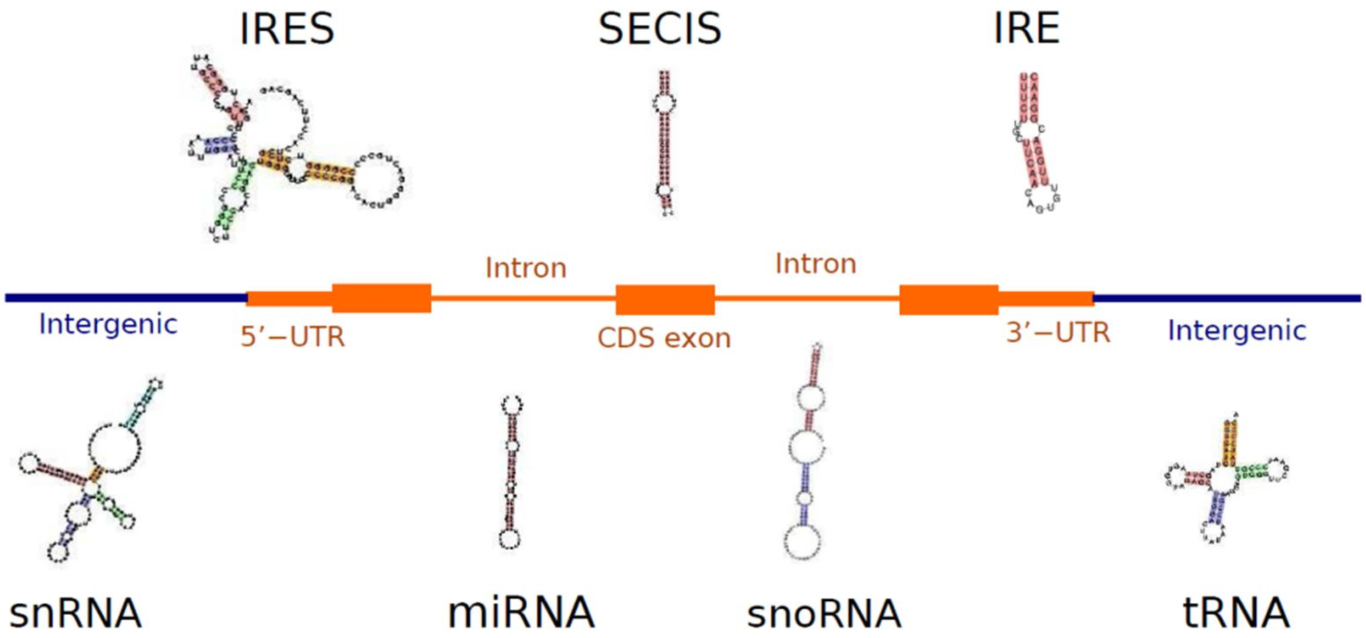


Nobel Prize in Physiology
or Medicine 2006

Taft *et al.*, J Pathol, 2010

Structured RNAs examples

Structured RNAs: examples



Research in RNA Bionformatics and Perspectives

- Initially focus on folding of single RNA molecules, but further improvements:
  - Nussinov algorithm
  - Zuker algorithm and partition function
  - Fold many sequence togehter -> exploiting comparative information
  - More complex models for finding RNA motifs (Covariance models, Rfam database)

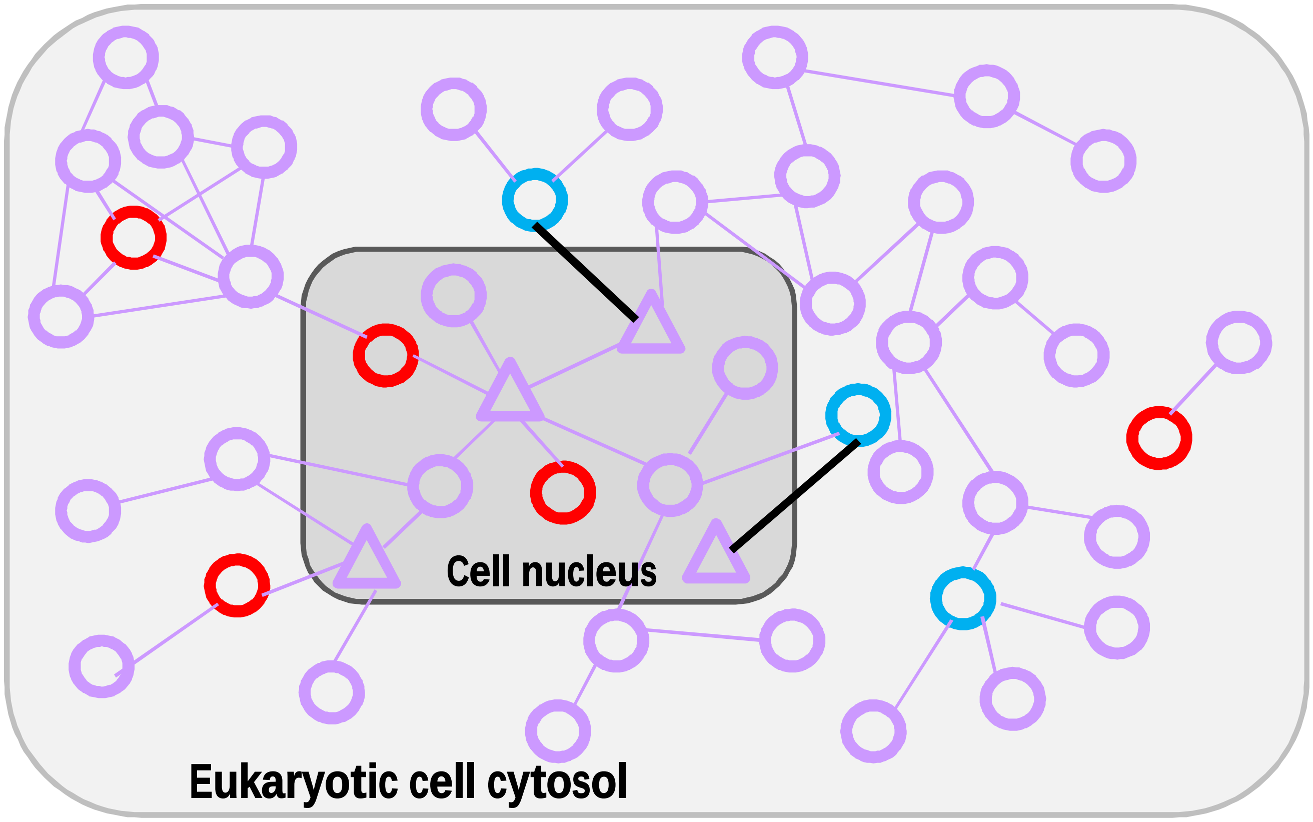
- Searching for ncRNAs

- **miRNA identification and role in gene-regulatory networks**

- **lncRNA (~13000 in the human genome) new challenge: poorly annotated, poorly conserved, strucures unkown**

- Focus RNA-RNA interactions and RNA-protein interactions
  - miRNA target prediction
  - lncRNA target prediction (indirect methods)
  - RNA Binding Proteins (RBPs)

Non-coding RNAs in gene regulatory networks

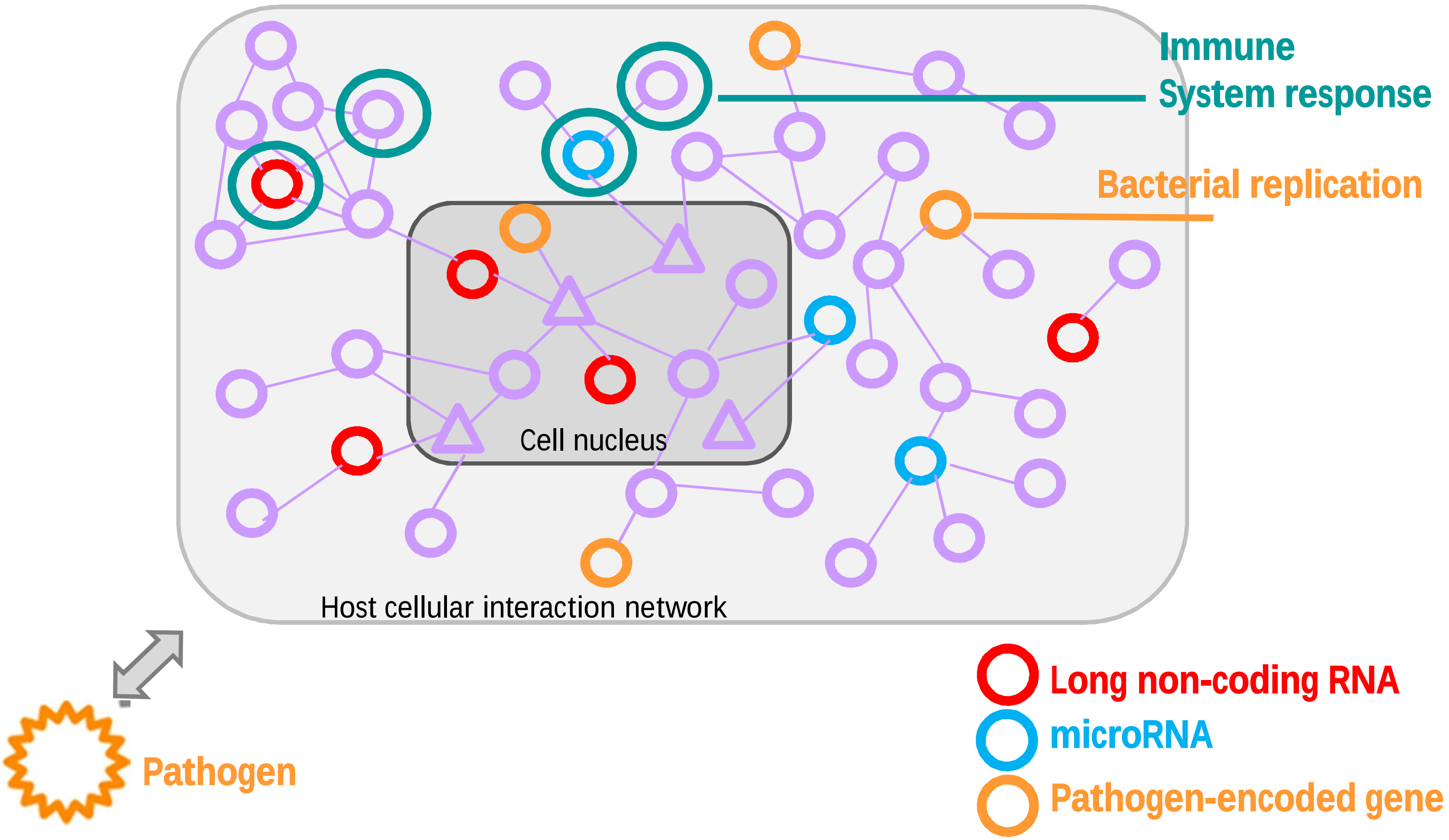


○ Long non-coding RNA

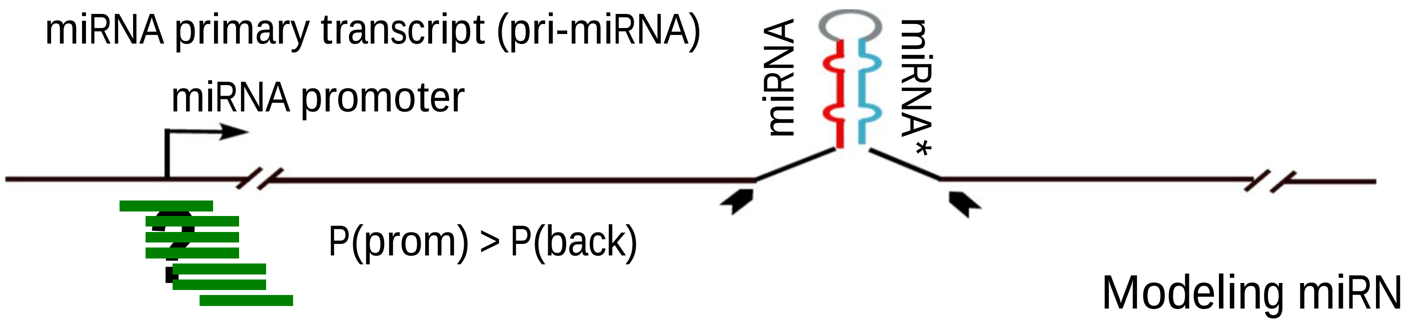
○ microRNA

△ Transcription factor

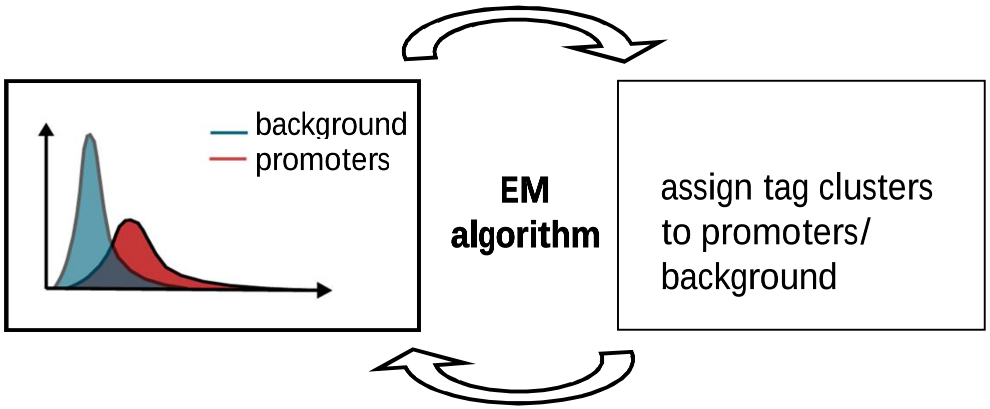
Non-coding RNA-mediated networks in bacterial infections



Statistical modeling of miRNA transcriptional regulation

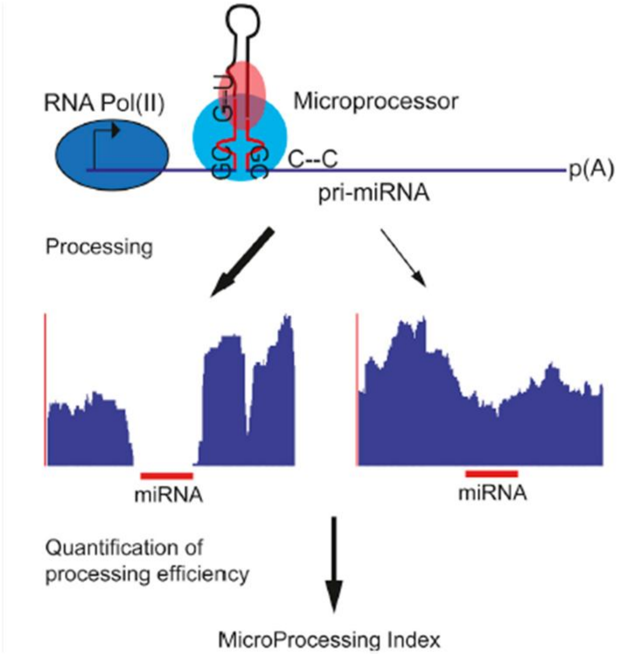


PROmiRNA : machine learning method for miRNA promoter prediction. Semi-supervised hierarchical model



A. Marsico et al. Genome Biol 2013

Modeling miRNA processing: String kernel regression



T. Conrad, A. Marsico* et al. Cell Reports 2014*

Research in RNA Bionformatics and Perspectives

- Initially focus on folding of single RNA molecules, but further improvements:
  - Nussinov algorithm
  - Zuker algorithm and partition function
  - Fold many sequence togehter -> exploiting comparative information
  - More complex models for finding RNA motifs (Covariance models, Rfam database)

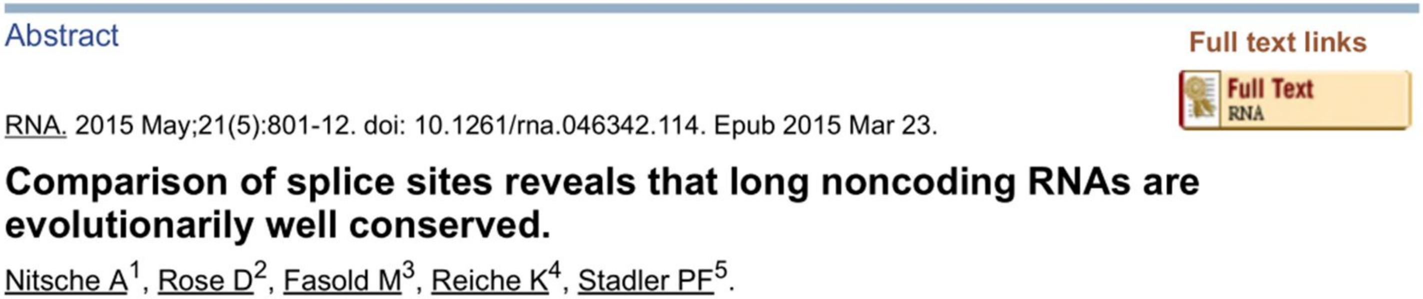
- Searching for ncRNAs

- miRNA identification and role in gene-regulatory networks

- **lncRNA (~13000 in the human genome) new challenge: poorly annotated, poorly conserved, strucures unkown**

- Focus RNA-RNA interactions and RNA-protein interactions
  - miRNA target prediction
  - lncRNA target prediction (indirect methods)
  - RNA Binding Proteins (RBPs)

Evolution of long-non coding RNAs and implication for their functional classification (largely unknown so far!)



Abstract

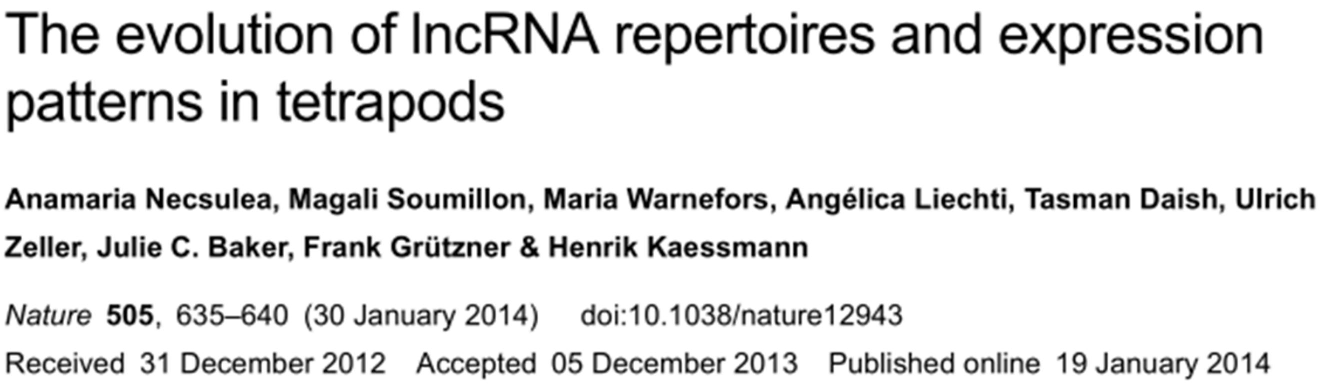
Full text links

Full Text
RNA

RNA. 2015 May;21(5):801-12. doi: 10.1261/rna.046342.114. Epub 2015 Mar 23.

Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved.

Nitsche A[1], Rose D[2], Fasold M[3], Reiche K[4], Stadler PF[5].



The evolution of lncRNA repertoires and expression patterns in tetrapods

Anamaria Necsulea, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C. Baker, Frank Grützner & Henrik Kaessmann

Nature **505**, 635–640 (30 January 2014) doi:10.1038/nature12943

Received 31 December 2012 Accepted 05 December 2013 Published online 19 January 2014

Research in RNA Bionformatics and Perspectives

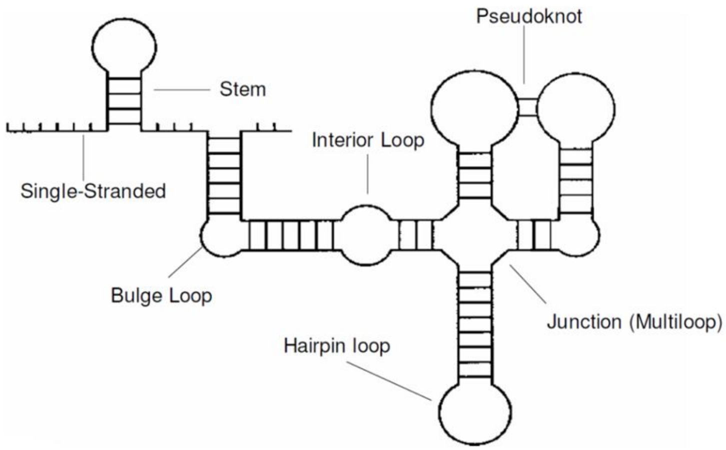
- Initially focus on folding of single RNA molecules, but further improvements:
  - **Nussinov algorithm**
  - **Zuker algorithm and partition function**
  - Fold many sequence togehter -> exploiting comparative information
  - More complex models for finding RNA motifs (Covariance models, Rfam database)
- Searching for ncRNAs
- miRNA identification and role in gene-regulatory networks
- lncRNA (~13000 in the human genome) new challenge: poorly annotated, poorly conserved, strucures unkown
- Focus RNA-RNA interactions and RNA-protein interactions
  - miRNA target prediction
  - lncRNA target prediction (indirect methods)
  - RNA Binding Proteins (RBPs)

Research in RNA Bionformatics and Perspectives

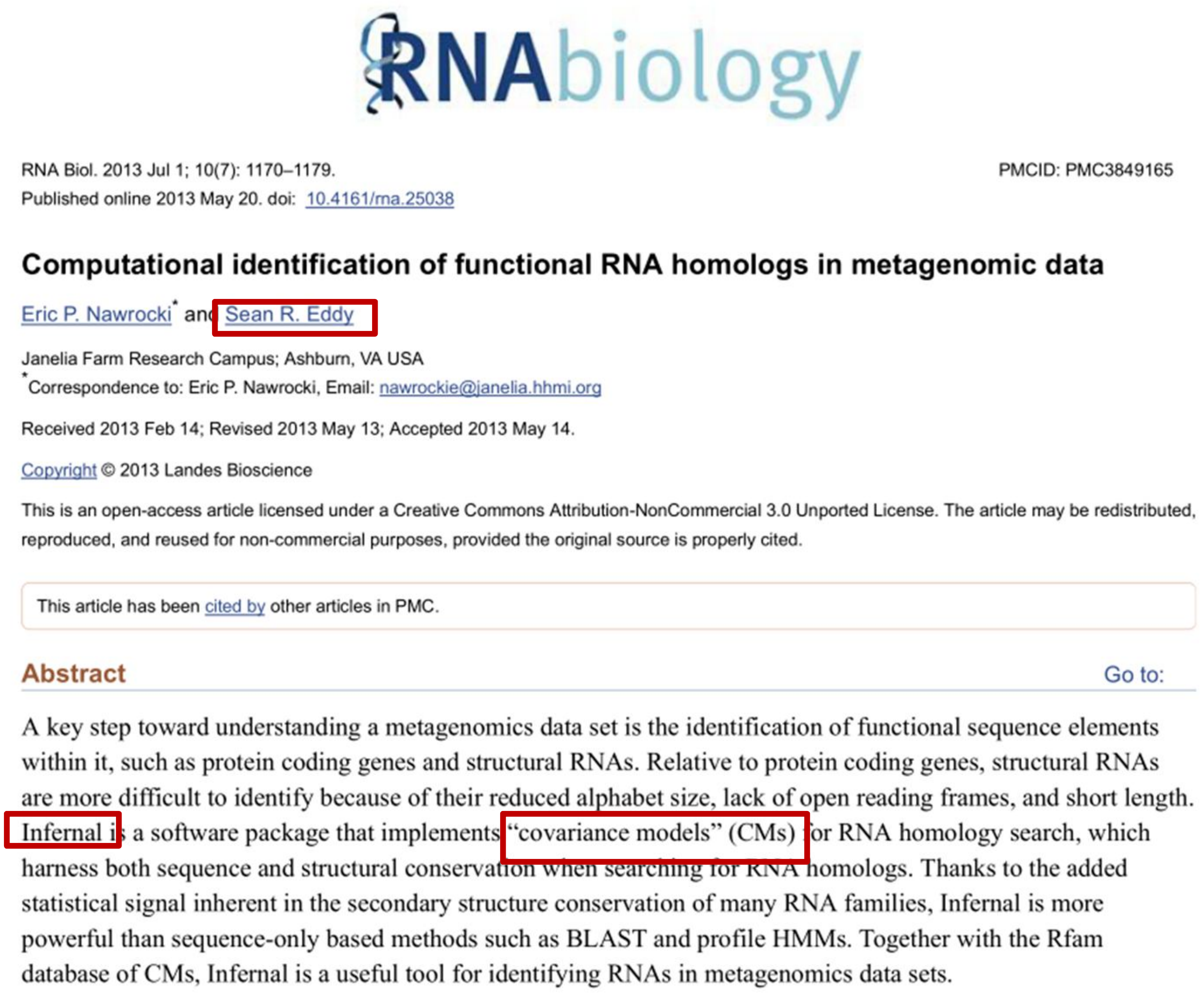
- Approximation: prediction of RNA secondary structure

```
RNAfold < trna.fa

>AF041468

GGGGGUAUAGCUCAGUUGGUAGAGCGCUGCCUUUGCACGGCAGAUGUCAGGGGUUCGAGUCCCCUUACCUCCA

(((((((..((((........)))).(((((.......))))).....(((((.......))))))))))))).

-31.10 kcal/mol
```

RNA secondary structure elements



Models and databases to represent RNA structure and sequence consensus



RNAbiology

RNA Biol. 2013 Jul 1; 10(7): 1170–1179.
Published online 2013 May 20. doi: 10.4161/rna.25038

PMCID: PMC3849165

Computational identification of functional RNA homologs in metagenomic data

Eric P. Nawrocki* and Sean R. Eddy

Janelia Farm Research Campus; Ashburn, VA USA
*Correspondence to: Eric P. Nawrocki, Email: nawrockie@janelia.hhmi.org

Received 2013 Feb 14; Revised 2013 May 13; Accepted 2013 May 14.

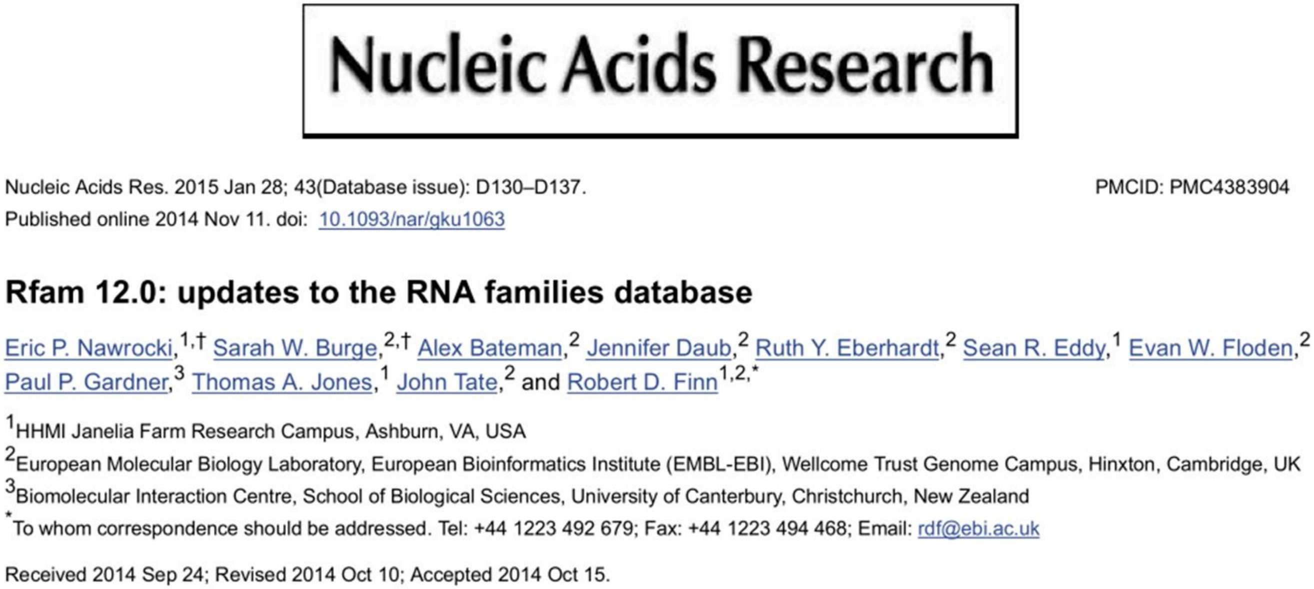
This article has been cited by other articles in PMC.

Abstract

Go to:

A key step toward understanding a metagenomics data set is the identification of functional sequence elements within it, such as protein coding genes and structural RNAs. Relative to protein coding genes, structural RNAs are more difficult to identify because of their reduced alphabet size, lack of open reading frames, and short length. Infernal is a software package that implements "covariance models" (CMs) for RNA homology search, which harness both sequence and structural conservation when searching for RNA homologs. Thanks to the added statistical signal inherent in the secondary structure conservation of many RNA families, Infernal is more powerful than sequence-only based methods such as BLAST and profile HMMs. Together with the Rfam database of CMs, Infernal is a useful tool for identifying RNAs in metagenomics data sets.

Models and databases to represent RNA structure and sequence consensus



Nucleic Acids Research

Nucleic Acids Res. 2015 Jan 28; 43(Database issue): D130–D137. PMCID: PMC4383904
Published online 2014 Nov 11. doi: 10.1093/nar/gku1063

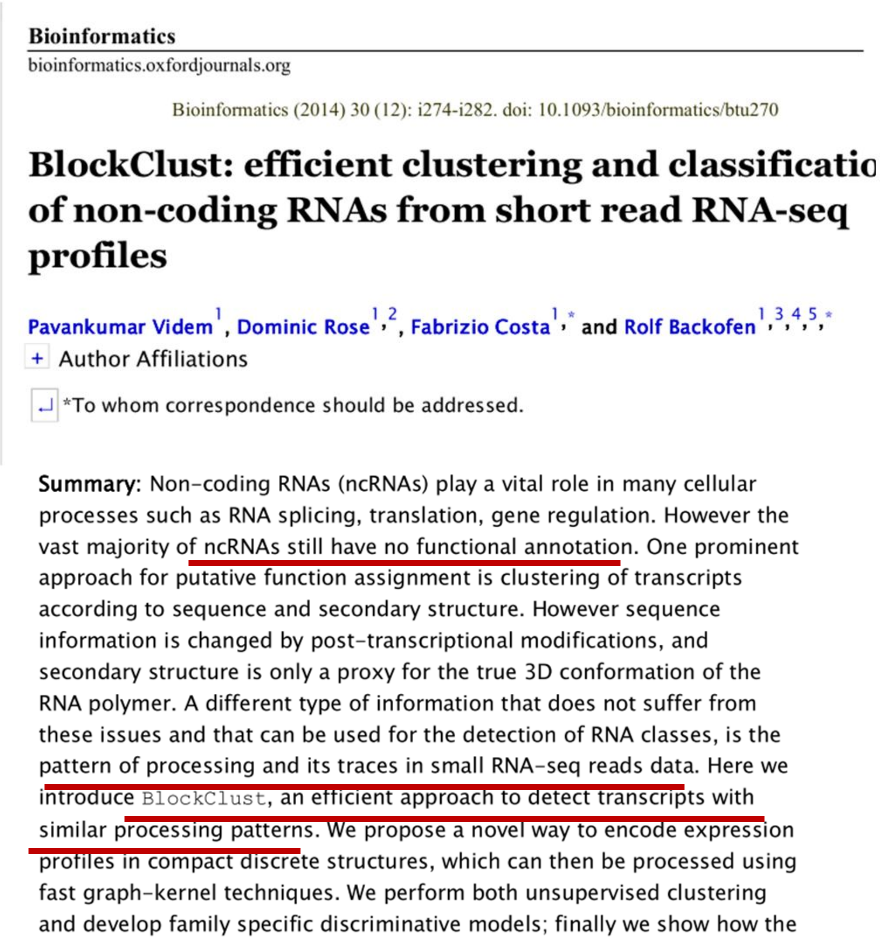
Rfam 12.0: updates to the RNA families database

Eric P. Nawrocki,[1,†] Sarah W. Burge,[2,†] Alex Bateman,[2] Jennifer Daub,[2] Ruth Y. Eberhardt,[2] Sean R. Eddy,[1] Evan W. Floden,[2] Paul P. Gardner,[3] Thomas A. Jones,[1] John Tate,[2] and Robert D. Finn[1,2,*]

[1]HHMI Janelia Farm Research Campus, Ashburn, VA, USA
[2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[3]Biomolecular Interaction Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand
[*]To whom correspondence should be addressed. Tel: +44 1223 492 679; Fax: +44 1223 494 468; Email: rdf@ebi.ac.uk

Received 2014 Sep 24; Revised 2014 Oct 10; Accepted 2014 Oct 15.

Sequence-structure alignment and folding of non-coding RNA - clustering



Bioinformatics
bioinformatics.oxfordjournals.org

Bioinformatics (2014) 30 (12): i274-i282. doi: 10.1093/bioinformatics/btu270

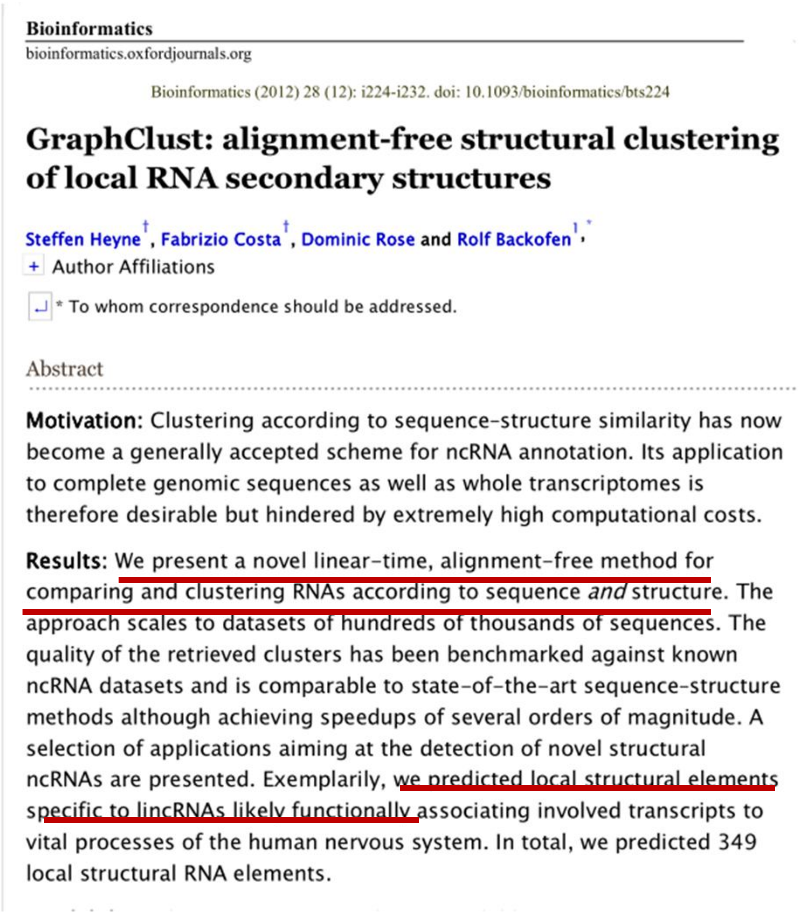
BlockClust: efficient clustering and classificatio of non-coding RNAs from short read RNA-seq profiles

Pavankumar Videm[1], Dominic Rose[1,2], Fabrizio Costa[1,*] and Rolf Backofen[1,3,4,5,*]

+ Author Affiliations

*To whom correspondence should be addressed.

Summary: Non-coding RNAs (ncRNAs) play a vital role in many cellular processes such as RNA splicing, translation, gene regulation. However the vast majority of ncRNAs still have no functional annotation. One prominent approach for putative function assignment is clustering of transcripts according to sequence and secondary structure. However sequence information is changed by post-transcriptional modifications, and secondary structure is only a proxy for the true 3D conformation of the RNA polymer. A different type of information that does not suffer from these issues and that can be used for the detection of RNA classes, is the pattern of processing and its traces in small RNA-seq reads data. Here we introduce `BlockClust`, an efficient approach to detect transcripts with similar processing patterns. We propose a novel way to encode expression profiles in compact discrete structures, which can then be processed using fast graph-kernel techniques. We perform both unsupervised clustering and develop family specific discriminative models; finally we show how the



Bioinformatics
bioinformatics.oxfordjournals.org

Bioinformatics (2012) 28 (12): i224-i232. doi: 10.1093/bioinformatics/bts224

GraphClust: alignment-free structural clustering of local RNA secondary structures

Steffen Heyne[†], Fabrizio Costa[†], Dominic Rose and Rolf Backofen[1,*]

+ Author Affiliations

* To whom correspondence should be addressed.

Abstract

Motivation: Clustering according to sequence-structure similarity has now become a generally accepted scheme for ncRNA annotation. Its application to complete genomic sequences as well as whole transcriptomes is therefore desirable but hindered by extremely high computational costs.

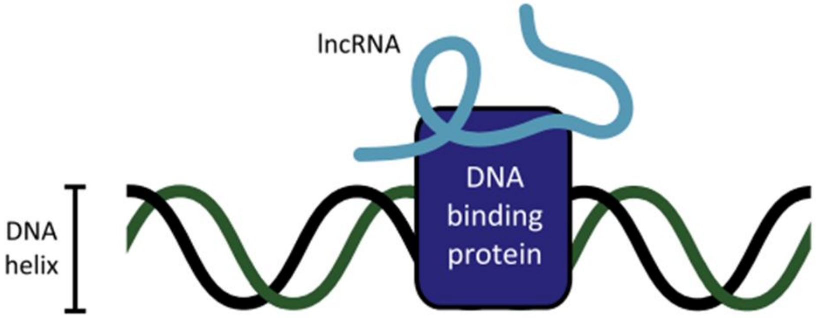
Results: We present a novel linear-time, alignment-free method for comparing and clustering RNAs according to sequence *and* structure. The approach scales to datasets of hundreds of thousands of sequences. The quality of the retrieved clusters has been benchmarked against known ncRNA datasets and is comparable to state-of-the-art sequence-structure methods although achieving speedups of several orders of magnitude. A selection of applications aiming at the detection of novel structural ncRNAs are presented. Exemplarily, we predicted local structural elements specific to lincRNAs likely functionally associating involved transcripts to vital processes of the human nervous system. In total, we predicted 349 local structural RNA elements.

Research in RNA Bionformatics and Perspectives

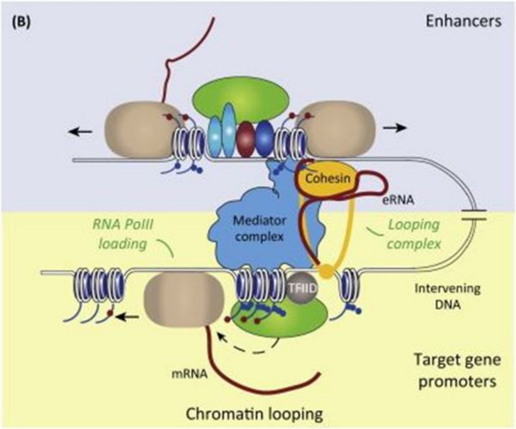
- Initially focus on folding of single RNA molecules, but further improvements:
  - Nussinov algorithm
  - Zuker algorithm and partition function
  - Fold many sequence togehter -> exploiting comparative information
  - More complex models for finding RNA motifs (Covariance models, Rfam database)
- Searching for ncRNAs
- miRNA identification and role in gene-regulatory networks
- lncRNA (~13000 in the human genome) new challenge: poorly annotated, poorly conserved, strucures unkown
- **Focus RNA-RNA interactions and RNA-protein interactions**
  - **miRNA target prediction**
  - **lncRNA target prediction (indirect methods)**
  - **RNA Binding Proteins (RBPs)**

Prediction of RNA Binding Protein (RBP) sites genome-wide

- Proteins are involved in RNA processing, e.g. Splicing
- When RNAs work in gene regulation they do it through protein-binding



Indirect RNA–protein–DNA associations



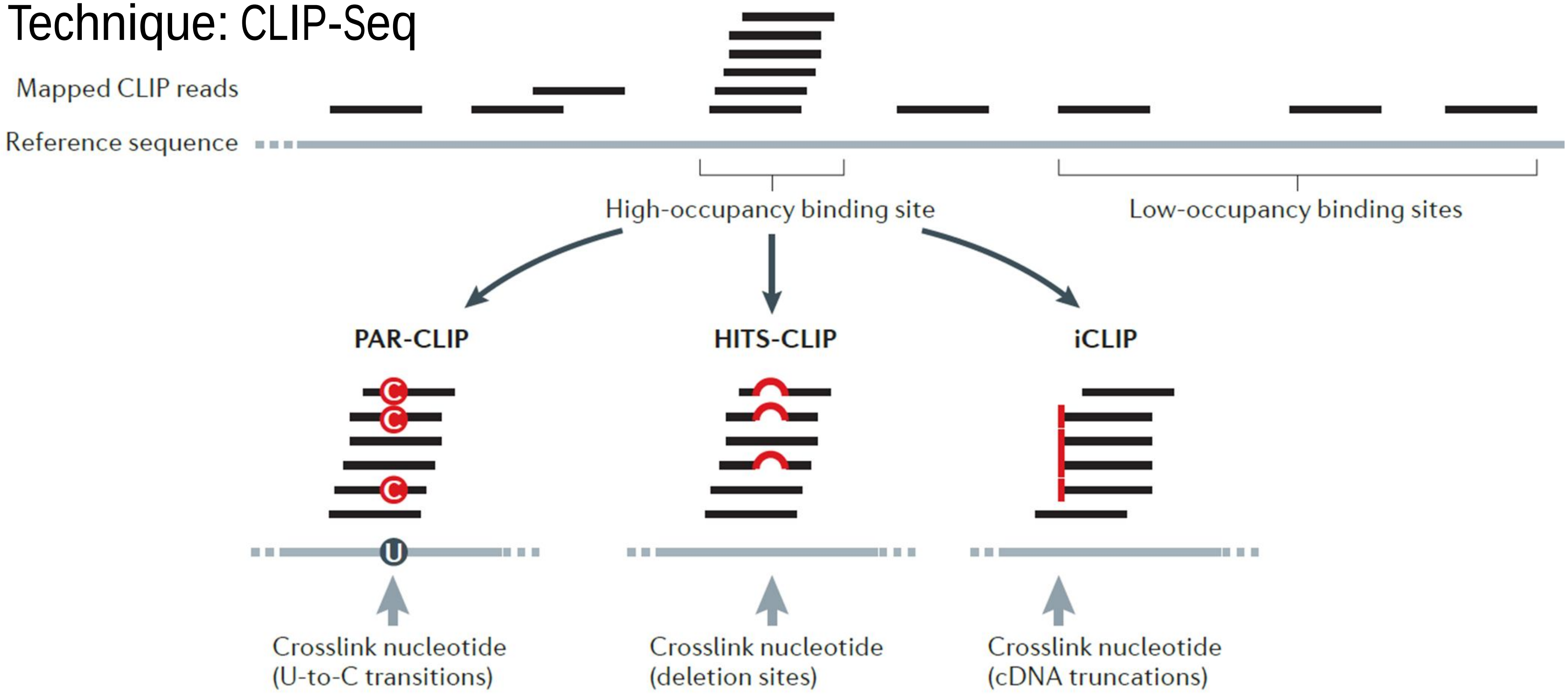
KW Vance & C. Ponting, Trends in Genetics 2014

M.T. Lam et al., Nature 2013, F. Lai et al, Nature 2013

Prediction of RNA Binding Protein (RBP) sites genome-wide

(Source: Nature Reviews Genetics 13, 77-83)

Technique: CLIP-Seq

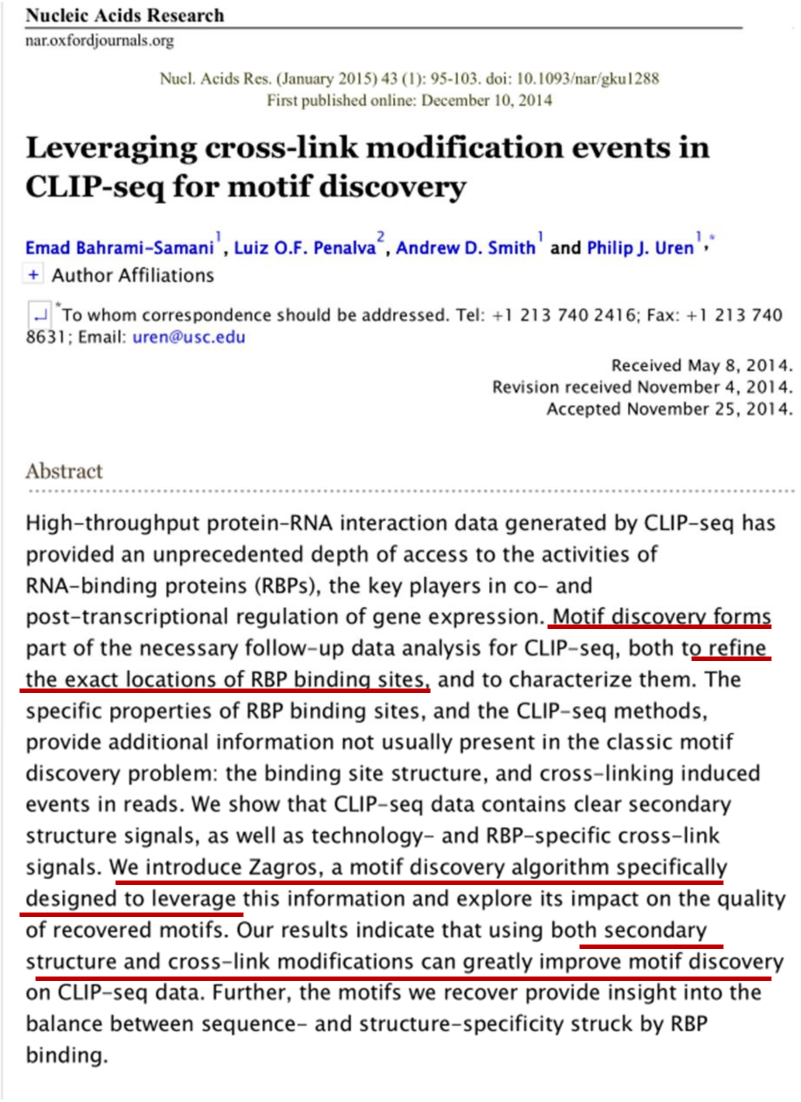


→ diagnostic events (DEs)

Read counts ⟺ RBP affinity ?

Sabrina Krakau

Investigation of RNA function through binding with RNA Binding Proteins - CLIP-seq and hiCLIP-seq data



Nucleic Acids Research
nar.oxfordjournals.org

Nucl. Acids Res. (January 2015) 43 (1): 95-103. doi: 10.1093/nar/gku1288
First published online: December 10, 2014

Leveraging cross-link modification events in CLIP-seq for motif discovery

Emad Bahrami-Samani[1], Luiz O.F. Penalva[2], Andrew D. Smith[1] and Philip J. Uren[1,*]

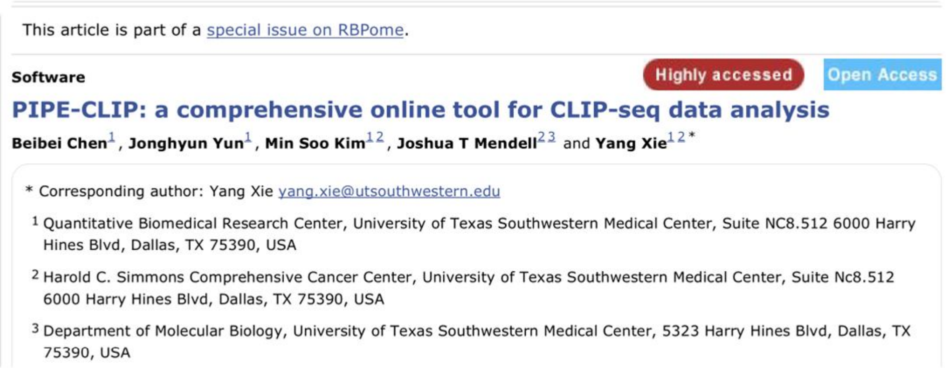
+ Author Affiliations

*To whom correspondence should be addressed. Tel: +1 213 740 2416; Fax: +1 213 740 8631; Email: uren@usc.edu

Received May 8, 2014.
Revision received November 4, 2014.
Accepted November 25, 2014.

Abstract

High-throughput protein–RNA interaction data generated by CLIP-seq has provided an unprecedented depth of access to the activities of RNA-binding proteins (RBPs), the key players in co- and post-transcriptional regulation of gene expression. Motif discovery forms part of the necessary follow-up data analysis for CLIP-seq, both to refine the exact locations of RBP binding sites, and to characterize them. The specific properties of RBP binding sites, and the CLIP-seq methods, provide additional information not usually present in the classic motif discovery problem: the binding site structure, and cross-linking induced events in reads. We show that CLIP-seq data contains clear secondary structure signals, as well as technology- and RBP-specific cross-link signals. We introduce Zagros, a motif discovery algorithm specifically designed to leverage this information and explore its impact on the quality of recovered motifs. Our results indicate that using both secondary structure and cross-link modifications can greatly improve motif discovery on CLIP-seq data. Further, the motifs we recover provide insight into the balance between sequence- and structure-specificity struck by RBP binding.



This article is part of a special issue on RBPome.

Software

Highly accessed Open Access

PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis

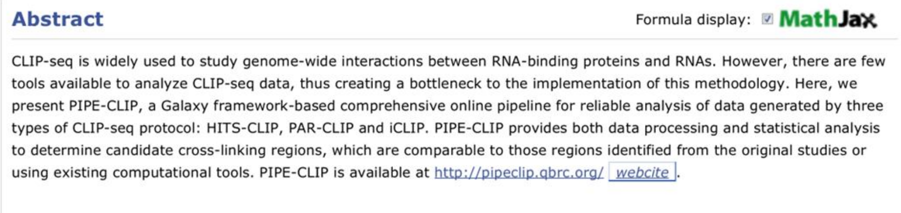
Beibei Chen[1], Jonghyun Yun[1], Min Soo Kim[1 2], Joshua T Mendell[2 3] and Yang Xie[1 2 *]

* Corresponding author: Yang Xie yang.xie@utsouthwestern.edu

[1] Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, Suite NC8.512 6000 Harry Hines Blvd, Dallas, TX 75390, USA

[2] Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Suite Nc8.512 6000 Harry Hines Blvd, Dallas, TX 75390, USA

[3] Department of Molecular Biology, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, USA



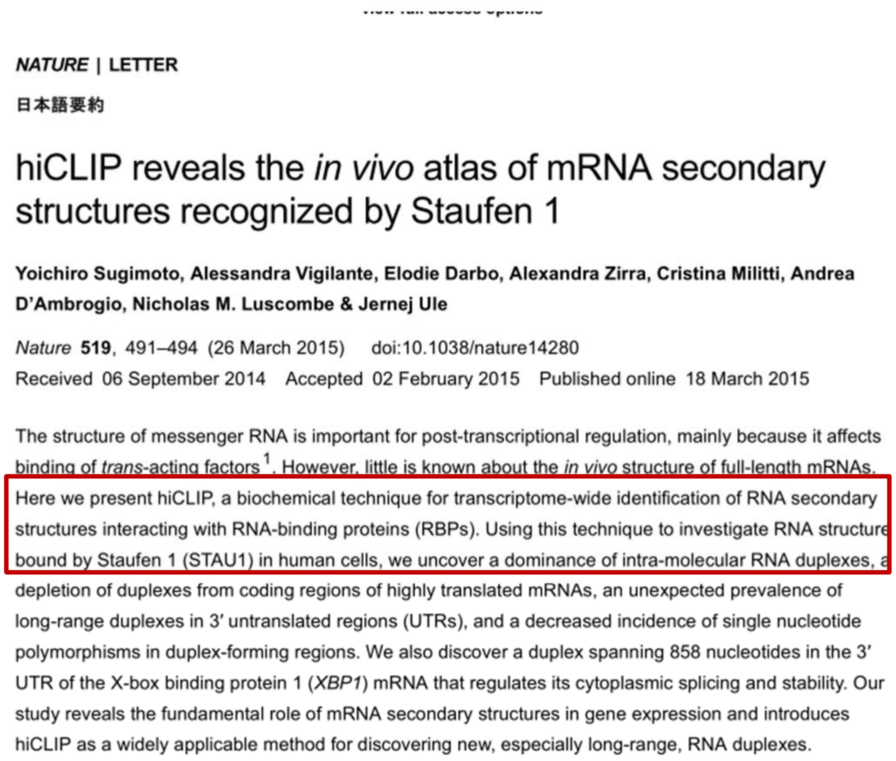
Abstract

Formula display: ☑ MathJax

CLIP-seq is widely used to study genome-wide interactions between RNA-binding proteins and RNAs. However, there are few tools available to analyze CLIP-seq data, thus creating a bottleneck to the implementation of this methodology. Here, we present PIPE-CLIP, a Galaxy framework-based comprehensive online pipeline for reliable analysis of data generated by three types of CLIP-seq protocol: HITS-CLIP, PAR-CLIP and iCLIP. PIPE-CLIP provides both data processing and statistical analysis to determine candidate cross-linking regions, which are comparable to those regions identified from the original studies or using existing computational tools. PIPE-CLIP is available at http://pipeclip.qbrc.org/ webcite.

A statistical method for peak calling based on a generalized linear model

Investigation of RNA function through binding with RNA Binding Proteins - CLIP-seq and hiCLIP-seq data



View full access options

NATURE | LETTER

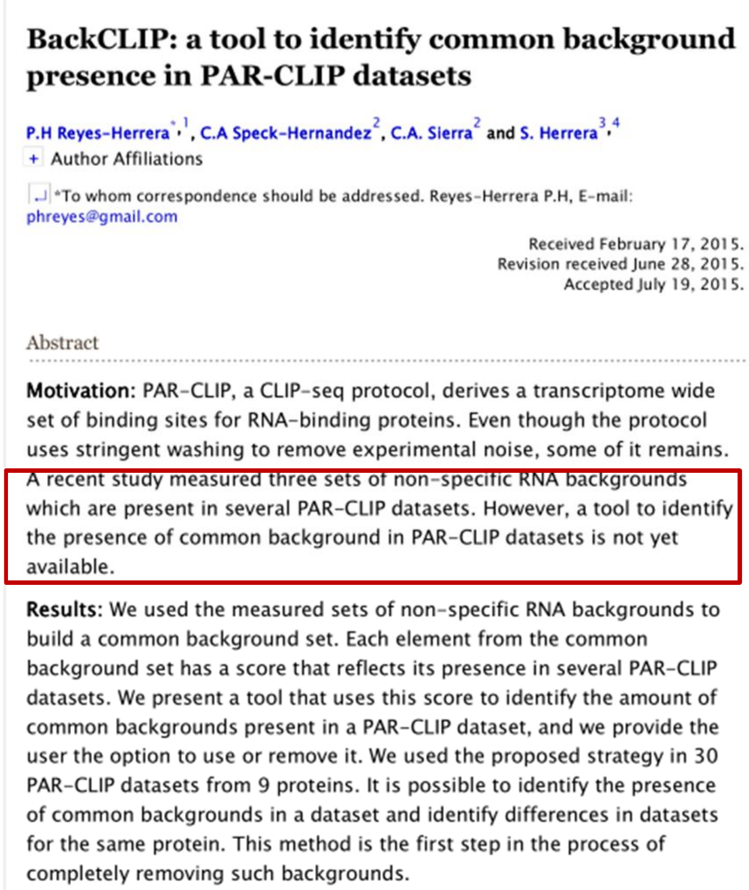
日本語要約

hiCLIP reveals the *in vivo* atlas of mRNA secondary structures recognized by Staufen 1

Yoichiro Sugimoto, Alessandra Vigilante, Elodie Darbo, Alexandra Zirra, Cristina Militti, Andrea D'Ambrogio, Nicholas M. Luscombe & Jernej Ule

Nature **519**, 491–494 (26 March 2015) doi:10.1038/nature14280
Received 06 September 2014 Accepted 02 February 2015 Published online 18 March 2015

The structure of messenger RNA is important for post-transcriptional regulation, mainly because it affects binding of *trans*-acting factors[1]. However, little is known about the *in vivo* structure of full-length mRNAs. Here we present hiCLIP, a biochemical technique for transcriptome-wide identification of RNA secondary structures interacting with RNA-binding proteins (RBPs). Using this technique to investigate RNA structure bound by Staufen 1 (STAU1) in human cells, we uncover a dominance of intra-molecular RNA duplexes, a depletion of duplexes from coding regions of highly translated mRNAs, an unexpected prevalence of long-range duplexes in 3′ untranslated regions (UTRs), and a decreased incidence of single nucleotide polymorphisms in duplex-forming regions. We also discover a duplex spanning 858 nucleotides in the 3′ UTR of the X-box binding protein 1 (*XBP1*) mRNA that regulates its cytoplasmic splicing and stability. Our study reveals the fundamental role of mRNA secondary structures in gene expression and introduces hiCLIP as a widely applicable method for discovering new, especially long-range, RNA duplexes.



BackCLIP: a tool to identify common background presence in PAR-CLIP datasets

P.H Reyes-Herrera*[1], C.A Speck-Hernandez[2], C.A. Sierra[2] and S. Herrera[3,4]
+ Author Affiliations

*To whom correspondence should be addressed. Reyes-Herrera P.H, E-mail: phreyes@gmail.com

Received February 17, 2015.
Revision received June 28, 2015.
Accepted July 19, 2015.

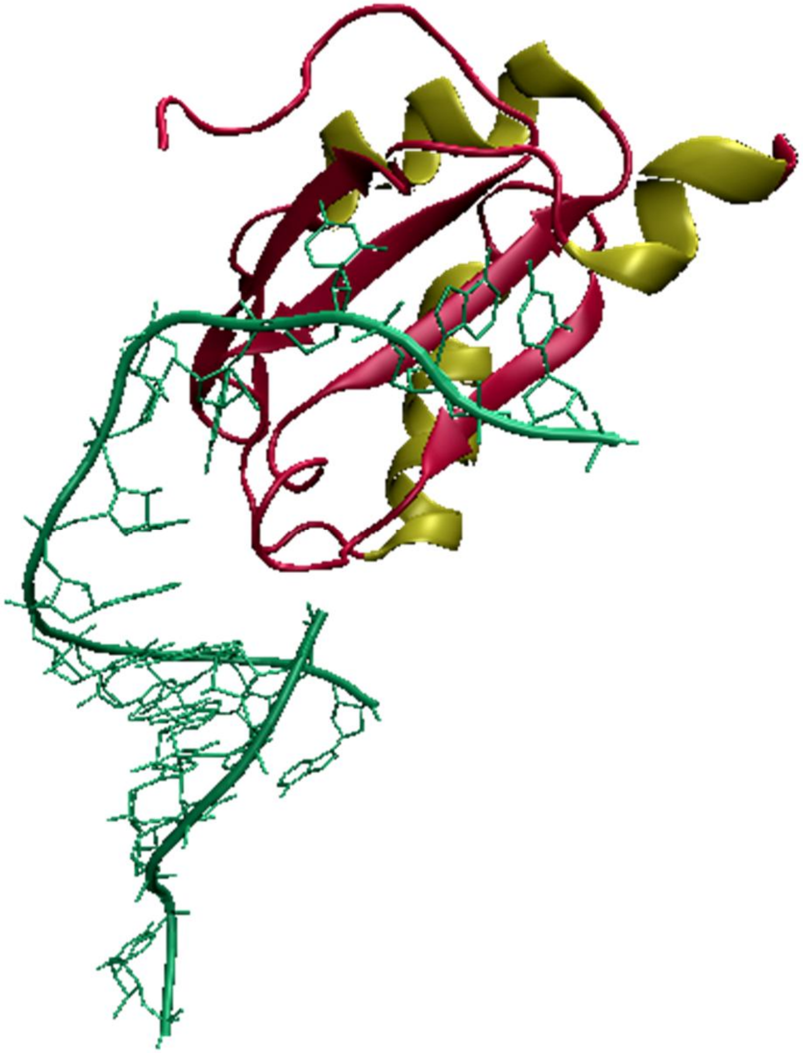
Abstract

Motivation: PAR-CLIP, a CLIP-seq protocol, derives a transcriptome wide set of binding sites for RNA-binding proteins. Even though the protocol uses stringent washing to remove experimental noise, some of it remains. A recent study measured three sets of non-specific RNA backgrounds which are present in several PAR-CLIP datasets. However, a tool to identify the presence of common background in PAR-CLIP datasets is not yet available.

Results: We used the measured sets of non-specific RNA backgrounds to build a common background set. Each element from the common background set has a score that reflects its presence in several PAR-CLIP datasets. We present a tool that uses this score to identify the amount of common backgrounds present in a PAR-CLIP dataset, and we provide the user the option to use or remove it. We used the proposed strategy in 30 PAR-CLIP datasets from 9 proteins. It is possible to identify the presence of common backgrounds in a dataset and identify differences in datasets for the same protein. This method is the first step in the process of completely removing such backgrounds.

Modeling and prediction of RNA-protein Binding Sites

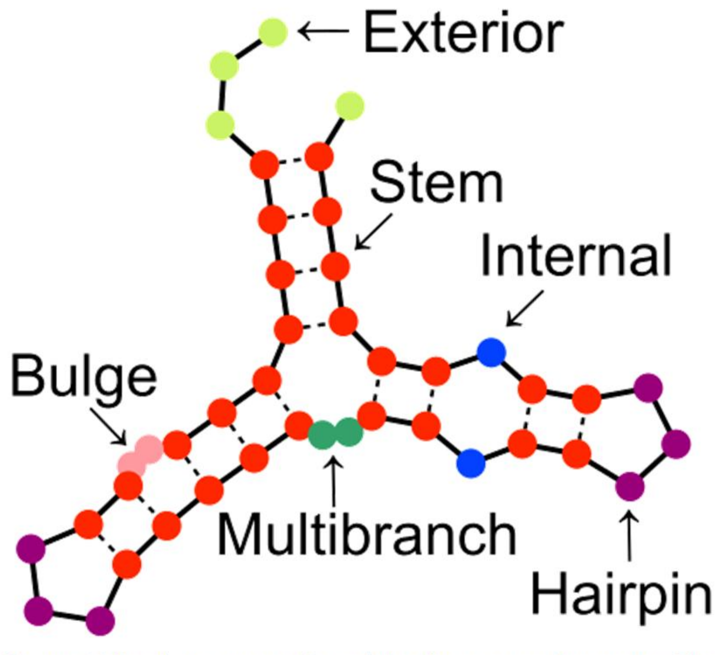
- RBPs process RNAs (e.g. Splicing, editing, stability)
- Help them to carry out their function
- Human genome has ~424 known and predicted RBPs
- Recognize their targets at sequence and structural level

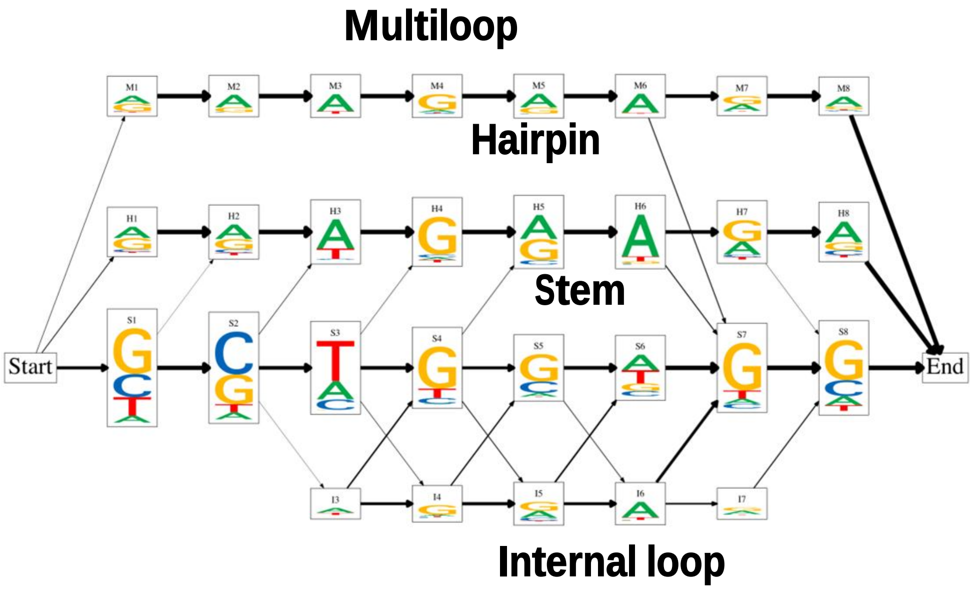


De novo discovery of RBPs motifs

HMM + Gibbs optimization
to capture sequence and structure preferences

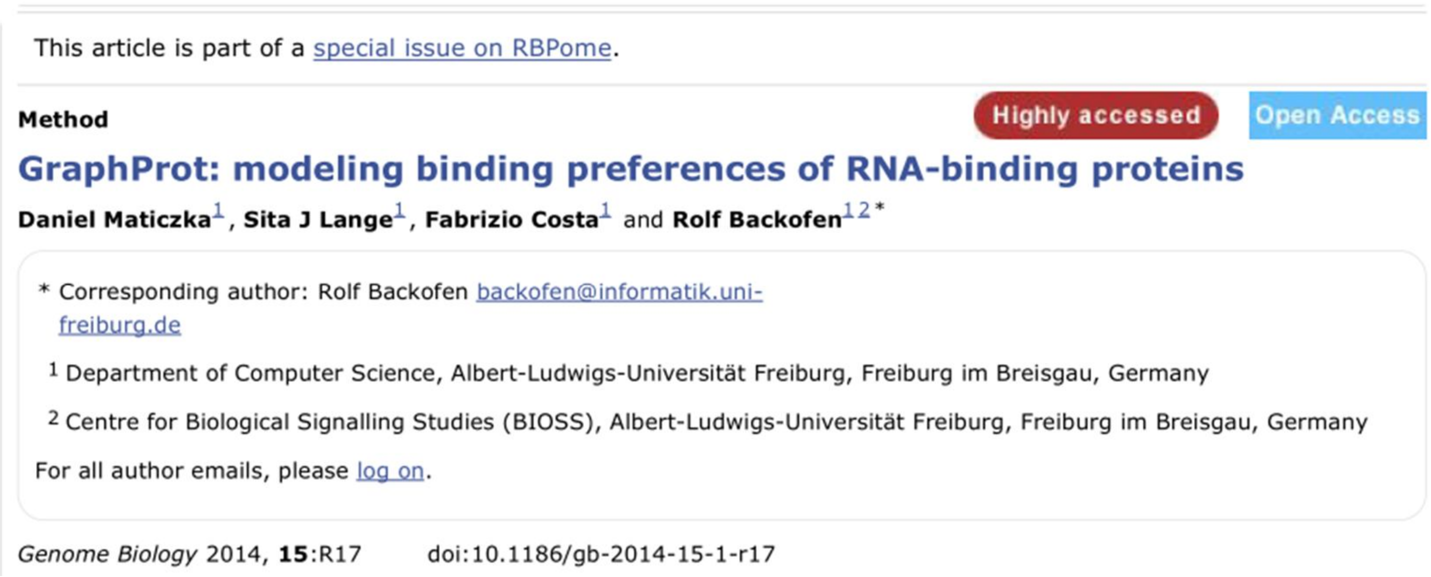
SFRS1 splicing factor





David Heller

Modeling and prediction of RNA-protein Binding Sites



This article is part of a special issue on RBPome.

Method

Highly accessed Open Access

GraphProt: modeling binding preferences of RNA-binding proteins

Daniel Maticzka[1], **Sita J Lange**[1], **Fabrizio Costa**[1] and **Rolf Backofen**[1 2 *]

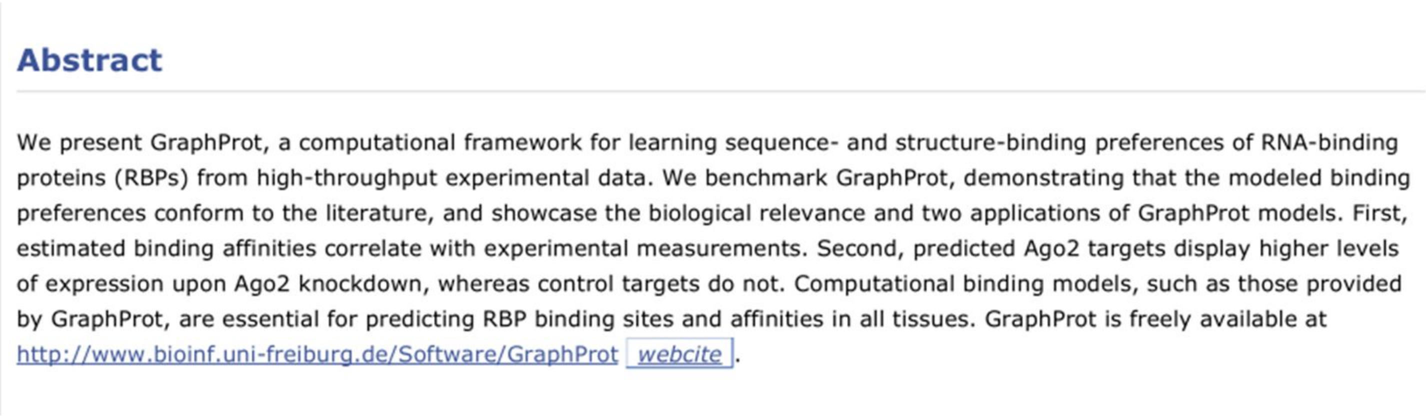
* Corresponding author: Rolf Backofen backofen@informatik.uni-freiburg.de

[1] Department of Computer Science, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany

[2] Centre for Biological Signalling Studies (BIOSS), Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany

For all author emails, please log on.

Genome Biology 2014, **15**:R17 doi:10.1186/gb-2014-15-1-r17



Abstract

We present GraphProt, a computational framework for learning sequence- and structure-binding preferences of RNA-binding proteins (RBPs) from high-throughput experimental data. We benchmark GraphProt, demonstrating that the modeled binding preferences conform to the literature, and showcase the biological relevance and two applications of GraphProt models. First, estimated binding affinities correlate with experimental measurements. Second, predicted Ago2 targets display higher levels of expression upon Ago2 knockdown, whereas control targets do not. Computational binding models, such as those provided by GraphProt, are essential for predicting RBP binding sites and affinities in all tissues. GraphProt is freely available at http://www.bioinf.uni-freiburg.de/Software/GraphProt webcite.

Research in RNA Bionformatics and Perspectives

- Initially focus on folding of single RNA molecules, but further improvements:
  - Nussinov algorithm
  - Zuker algorithm and partition function
  - Fold many sequence togehter -> exploiting comparative information
  - More complex models for finding RNA motifs (Covariance models, Rfam database)

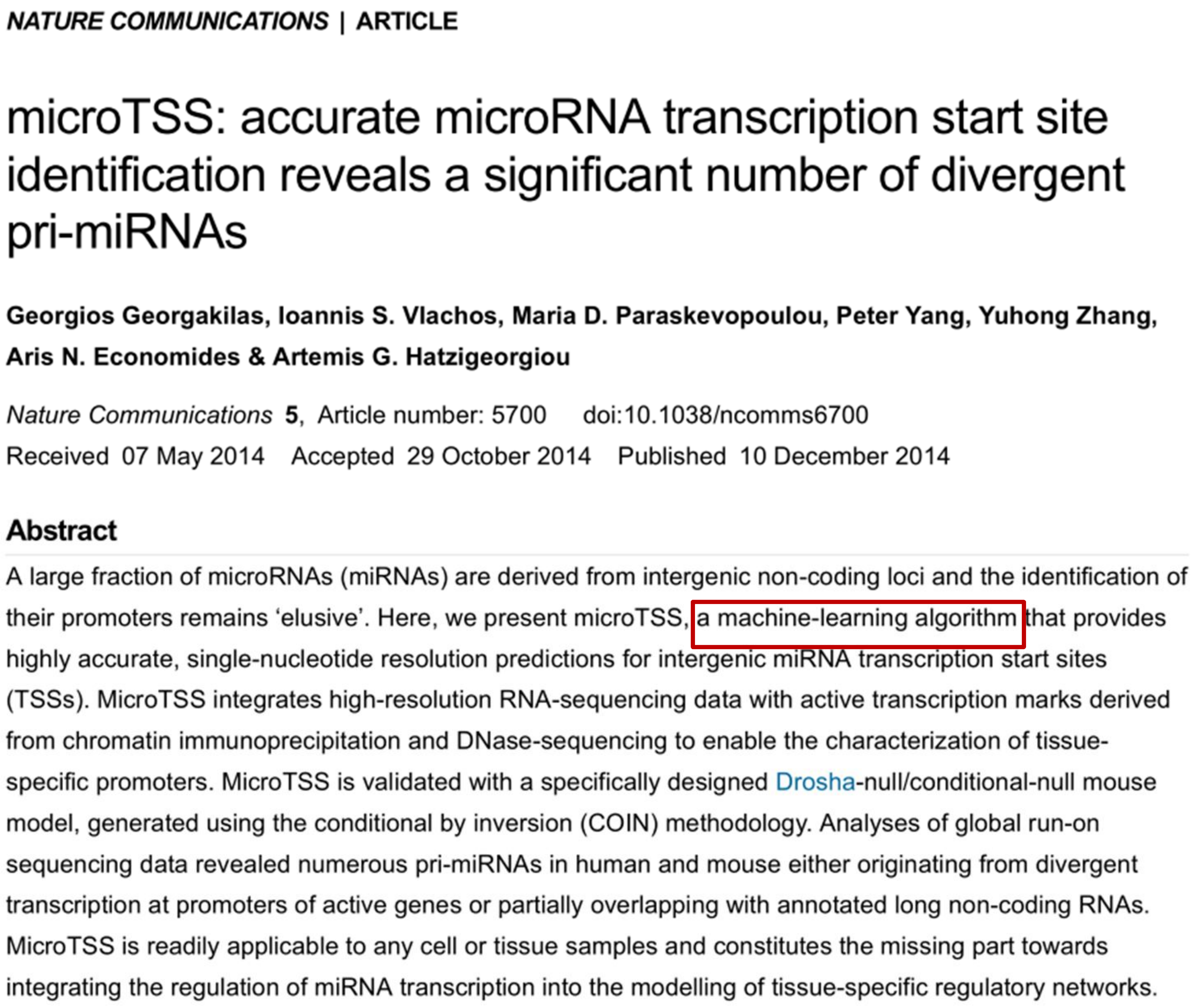
- Searching for ncRNAs

- **miRNA identification and role in gene-regulatory networks**

- lncRNA (~13000 in the human genome) new challenge: poorly annotated, poorly conserved, strucures unkown

- Focus RNA-RNA interactions and RNA-protein interactions
  - miRNA target prediction
  - lncRNA target prediction (indirect methods)
  - RNA Binding Proteins (RBPs)

Accurate annotation of microRNA genes from high-throughput data



NATURE COMMUNICATIONS | ARTICLE

microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs

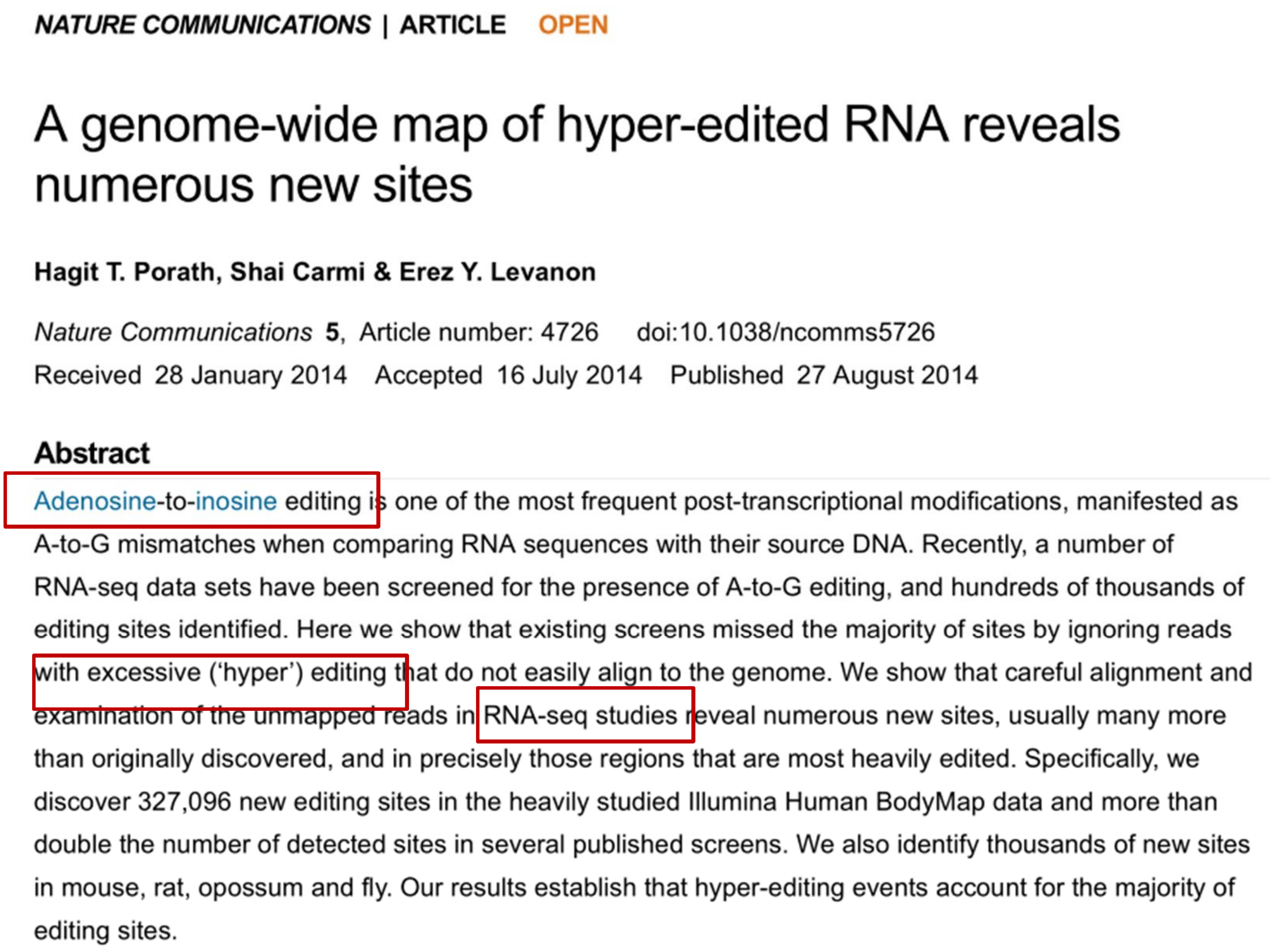
Georgios Georgakilas, Ioannis S. Vlachos, Maria D. Paraskevopoulou, Peter Yang, Yuhong Zhang, Aris N. Economides & Artemis G. Hatzigeorgiou

Nature Communications **5**, Article number: 5700 doi:10.1038/ncomms6700
Received 07 May 2014 Accepted 29 October 2014 Published 10 December 2014

Abstract

A large fraction of microRNAs (miRNAs) are derived from intergenic non-coding loci and the identification of their promoters remains 'elusive'. Here, we present microTSS, a machine-learning algorithm that provides highly accurate, single-nucleotide resolution predictions for intergenic miRNA transcription start sites (TSSs). MicroTSS integrates high-resolution RNA-sequencing data with active transcription marks derived from chromatin immunoprecipitation and DNase-sequencing to enable the characterization of tissue-specific promoters. MicroTSS is validated with a specifically designed Drosha-null/conditional-null mouse model, generated using the conditional by inversion (COIN) methodology. Analyses of global run-on sequencing data revealed numerous pri-miRNAs in human and mouse either originating from divergent transcription at promoters of active genes or partially overlapping with annotated long non-coding RNAs. MicroTSS is readily applicable to any cell or tissue samples and constitutes the missing part towards integrating the regulation of miRNA transcription into the modelling of tissue-specific regulatory networks.

RNA post-transcriptional modifications important for RNA functional studies



NATURE COMMUNICATIONS | ARTICLE OPEN

A genome-wide map of hyper-edited RNA reveals numerous new sites

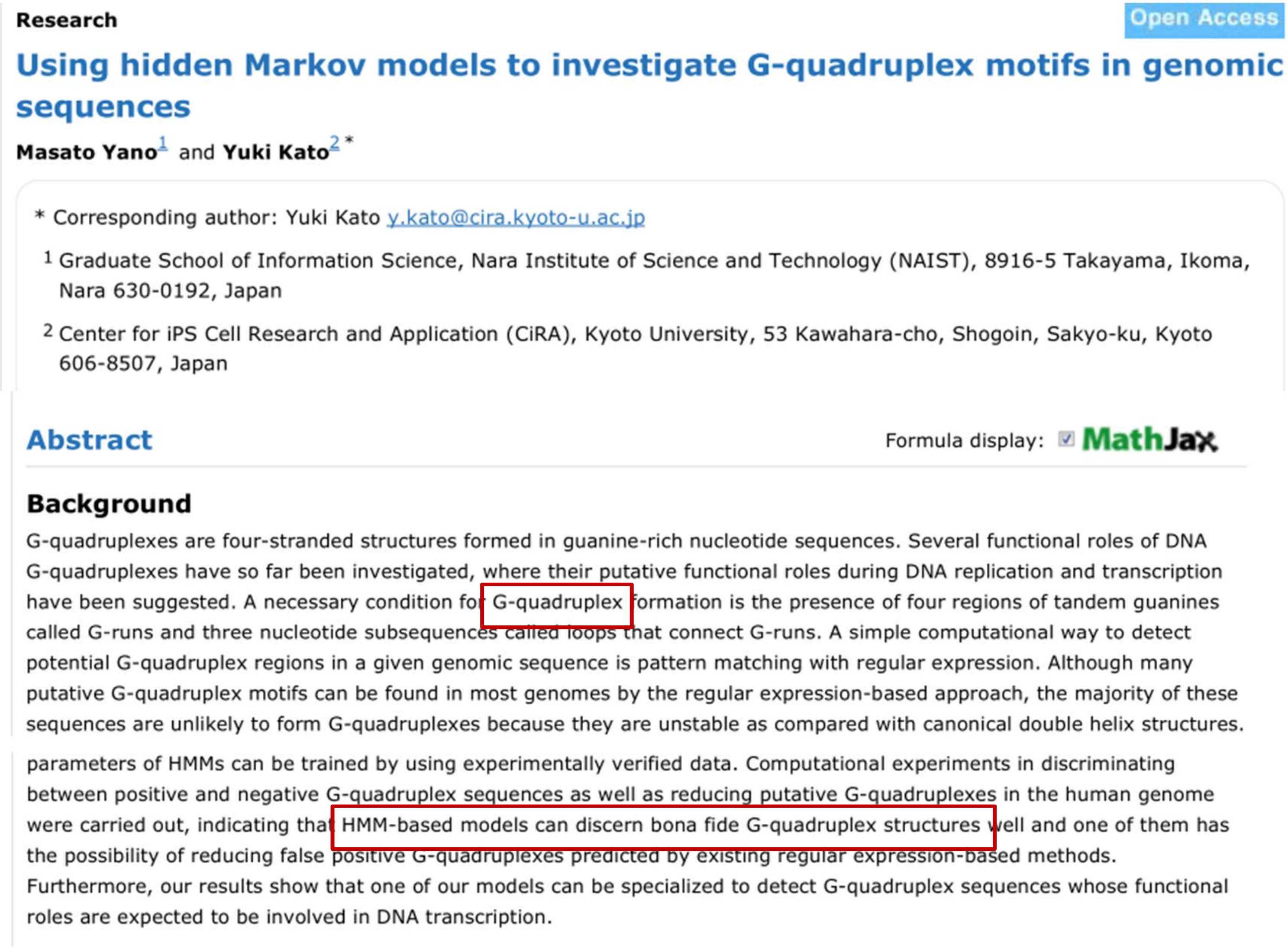
Hagit T. Porath, Shai Carmi & Erez Y. Levanon

Nature Communications **5**, Article number: 4726 doi:10.1038/ncomms5726
Received 28 January 2014 Accepted 16 July 2014 Published 27 August 2014

Abstract

Adenosine-to-inosine editing is one of the most frequent post-transcriptional modifications, manifested as A-to-G mismatches when comparing RNA sequences with their source DNA. Recently, a number of RNA-seq data sets have been screened for the presence of A-to-G editing, and hundreds of thousands of editing sites identified. Here we show that existing screens missed the majority of sites by ignoring reads with excessive ('hyper') editing that do not easily align to the genome. We show that careful alignment and examination of the unmapped reads in RNA-seq studies reveal numerous new sites, usually many more than originally discovered, and in precisely those regions that are most heavily edited. Specifically, we discover 327,096 new editing sites in the heavily studied Illumina Human BodyMap data and more than double the number of detected sites in several published screens. We also identify thousands of new sites in mouse, rat, opossum and fly. Our results establish that hyper-editing events account for the majority of editing sites.

RNA post-transcriptional modifications important for RNA functional studies



Research

Open Access

Using hidden Markov models to investigate G-quadruplex motifs in genomic sequences

Masato Yano[1] and **Yuki Kato**[2]*

* Corresponding author: Yuki Kato y.kato@cira.kyoto-u.ac.jp

[1] Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), 8916-5 Takayama, Ikoma, Nara 630-0192, Japan

[2] Center for iPS Cell Research and Application (CiRA), Kyoto University, 53 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan

Abstract

Formula display: ☑ MathJax

Background

G-quadruplexes are four-stranded structures formed in guanine-rich nucleotide sequences. Several functional roles of DNA G-quadruplexes have so far been investigated, where their putative functional roles during DNA replication and transcription have been suggested. A necessary condition for G-quadruplex formation is the presence of four regions of tandem guanines called G-runs and three nucleotide subsequences called loops that connect G-runs. A simple computational way to detect potential G-quadruplex regions in a given genomic sequence is pattern matching with regular expression. Although many putative G-quadruplex motifs can be found in most genomes by the regular expression-based approach, the majority of these sequences are unlikely to form G-quadruplexes because they are unstable as compared with canonical double helix structures.

parameters of HMMs can be trained by using experimentally verified data. Computational experiments in discriminating between positive and negative G-quadruplex sequences as well as reducing putative G-quadruplexes in the human genome were carried out, indicating that HMM-based models can discern bona fide G-quadruplex structures well and one of them has the possibility of reducing false positive G-quadruplexes predicted by existing regular expression-based methods. Furthermore, our results show that one of our models can be specialized to detect G-quadruplex sequences whose functional roles are expected to be involved in DNA transcription.