Beth Israel Lahey Health
Beth Israel Deaconess Medical Center

HARVARD MEDICAL SCHOOL
TEACHING HOSPITAL

Non-coding RNA Bioinformatics Workshop

Non-Coding RNA Precision Diagnostics and Therapeutics Core – www.NonCodingRNA.org

Ioannis S. Vlachos, PhD

Assistant Professor, Harvard Medical School
Associate Member, Broad Institute of MIT and Harvard
Co-Director, Bioinformatics Program, Cancer Research Institute
Director of Bioinformatics, ncRNA Institute,
Harvard Initiative for RNA Medicine,
Department of Pathology
Beth Israel Deaconess Medical Center

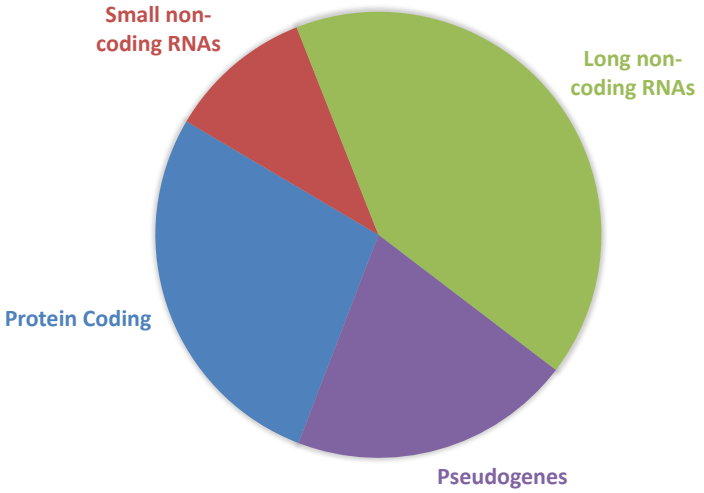
www.non-coding.org

BROAD INSTITUTE

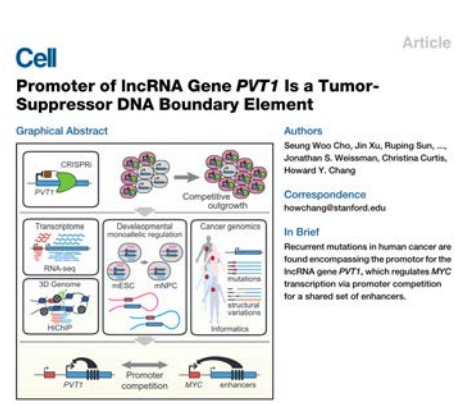
INITIATIVE FOR RNA MEDICINE
HARVARD MEDICAL SCHOOL
Founded and Hosted by
Institute for RNA Medicine,
Cancer Center at Beth Israel Deaconess Medical Center

We are not alone.

HUMAN TRANSCRIPTOME



www.NoncodingRNA.org www.non-coding.org



Article

Cell

Promoter of lncRNA Gene *PVT1* Is a Tumor-Suppressor DNA Boundary Element

Graphical Abstract

Authors

Seung Woo Cho, Jin Xu, Ruping Sun, ..., Jonathan S. Weissman, Christina Curtis, Howard Y. Chang

Correspondence

howchang@stanford.edu

In Brief

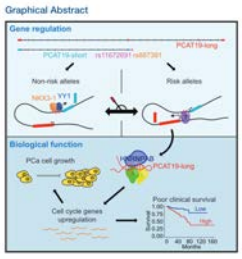
Recurrent mutations in human cancer are found encompassing the promotor for the lncRNA gene *PVT1*, which regulates *MYC* transcription via promoter competition for a shared set of enhancers.

nature

Letter | Published: 27 August 2018

The *NORAD* lncRNA assembles a topoisomerase complex critical for genome stability

Mathias Munschauer, Celina T. Nguyen, Klara Sirokman, Christina R. Hartigan, Larson Hogstrom, Jesse M. Engreitz, Jacob C. Ulirsch, Charles P. Fulco, Vidya Subramanian, Jenny Chen, Monica Schenone, Mitchell Guttman, Steven A. Carr & Eric S. Lander



Article

Cell

Risk SNP-Mediated Promoter-Enhancer Switching Drives Prostate Cancer through lncRNA *PCAT19*

Graphical Abstract

Authors

Junjie Tony Hua, Musaddeque Ahmed, Haiyang Guo, ..., Stephen N. Thibodeau, Paul C. Boutros, Housheng Hansen He

Correspondence

hansenhe@uhnresearch.ca

In Brief

Transcription factor binding site remodeling by a risk allele for aggressive prostate cancer results in conversion of a promoter to an enhancer with downstream consequences on long noncoding RNA isoform expression and oncogenesis.

nature

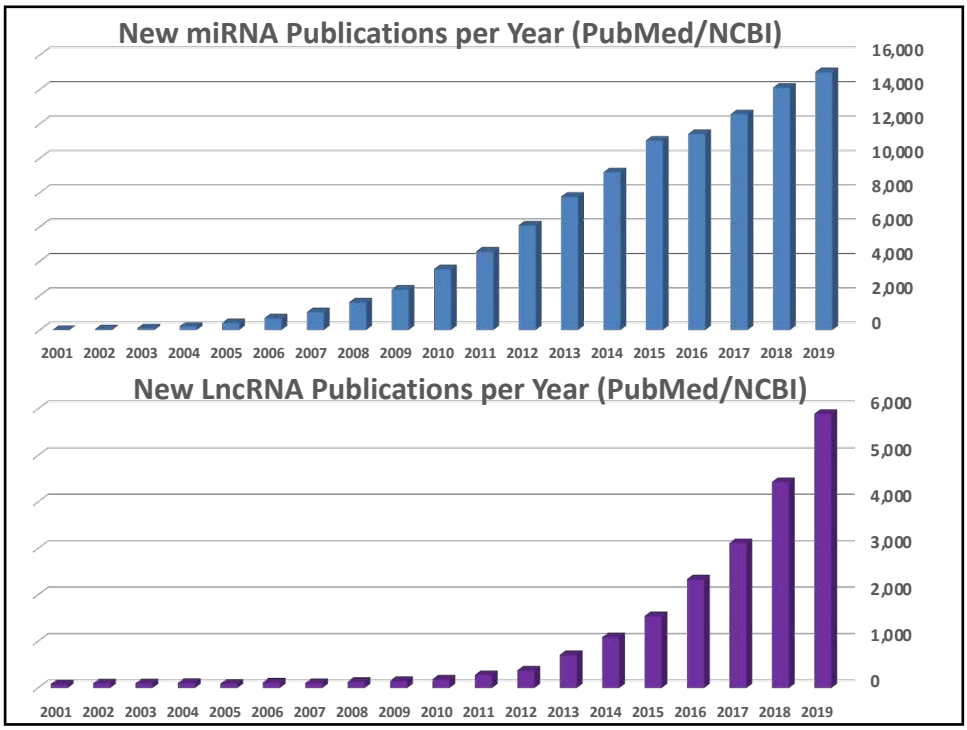
Letter | Published: 27 June 2018

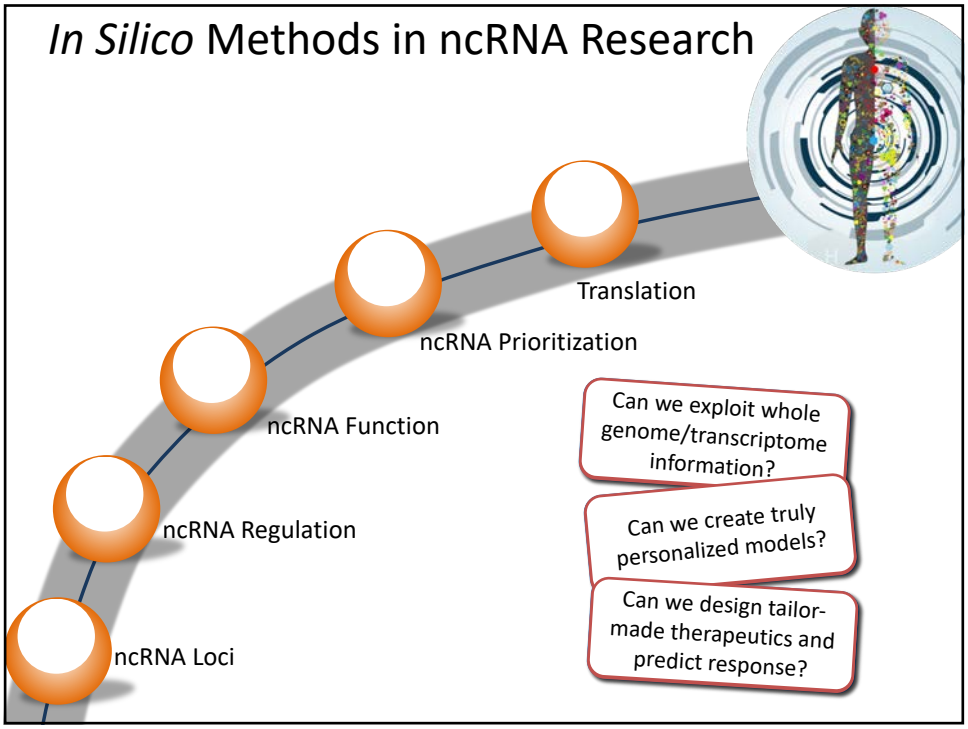
Induction of innate immune memory via microRNA targeting of chromatin remodelling factors

John J. Seeley, Rebecca G. Baker, Ghait Mohamed, Tony Bruns, Matthew S. Hayden, Sachin D. Deshmukh, Daniel E. Freedberg & Sankar Ghosh

Nature **559**, 114–119(2018) | Cite this article

4553 Accesses | **21** Citations | **112** Altmetric | Metrics

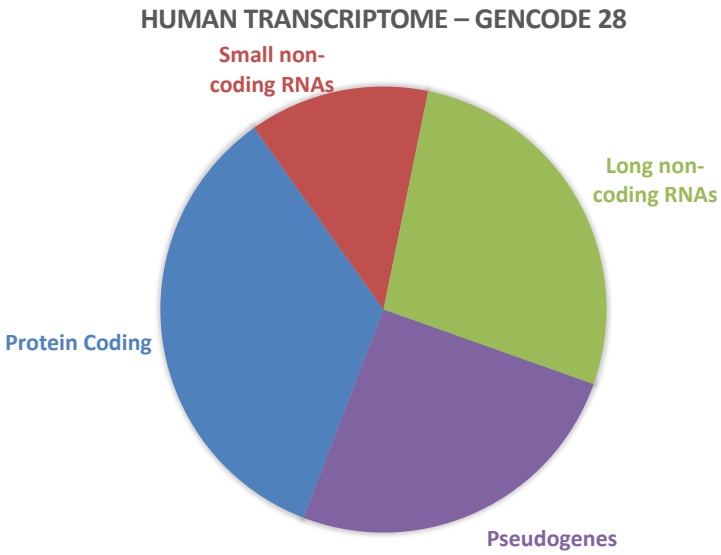

New miRNA Publications per Year (PubMed/NCBI)
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
New LncRNA Publications per Year (PubMed/NCBI)
6,000
5,000
4,000
3,000
2,000
1,000
0
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019


In Silico Methods in ncRNA Research
Translation
ncRNA Prioritization
ncRNA Function
ncRNA Regulation
ncRNA Loci
Can we exploit whole genome/transcriptome information?
Can we create truly personalized models?
Can we design tailor-made therapeutics and predict response?

More Questions

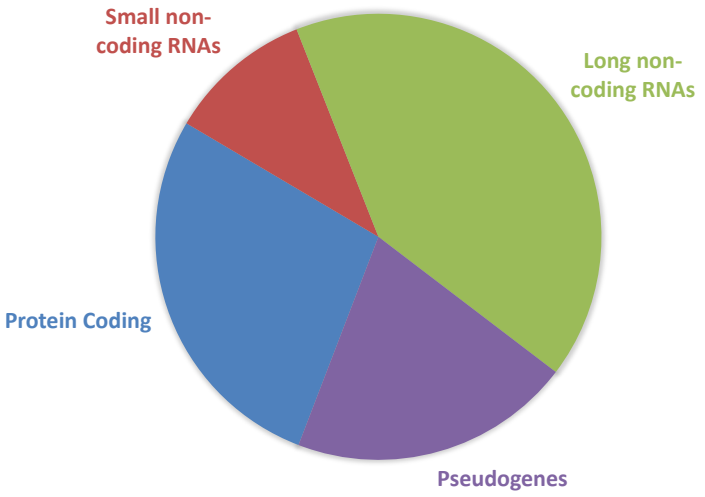
Than Answers

www.NoncodingRNA.org www.non-coding.org



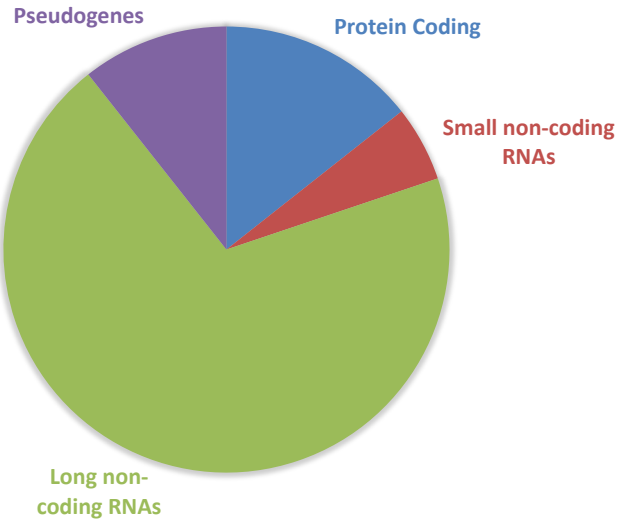
www.NoncodingRNA.org www.non-coding.org

HUMAN TRANSCRIPTOME – FANTOM 5



www.NoncodingRNA.org www.non-coding.org

HUMAN TRANSCRIPTOME – NONCODE V5

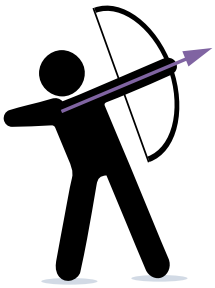


www.NoncodingRNA.org www.non-coding.org

Workshop Topics

- Quantification of microRNAs and Small RNAs using NGS
- Small RNA differential expression analyses
- Identifying microRNA targets
- Uncovering novel and known lncRNAs using NGS
- LncRNA and miRNA functional analyses
- Integrating non-coding RNAs in gene regulatory networks
- Detecting functional Non-coding variants

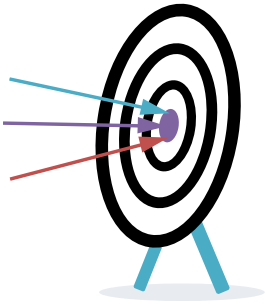
Workshop Aims



01 **Go Wide!**

02 **Provide a Compass to Newcomers**

03 **Discuss what we really know for the Pros**



The HIRM

"The goal of the **HMS Initiative for RNA Medicine** (HIRM) is to make and translate RNA discoveries into novel therapeutics and diagnostics by fostering a climate of outstanding basic, translational and clinical research with world-leading scientists and clinicians"

https://www.bidmc.org/research/research-centers/hms-initiative-for-rna-medicine

INITIATIVE FOR RNA MEDICINE
HARVARD MEDICAL SCHOOL

Non-Coding RNA Core Facility

INITIATIVE FOR RNA MEDICINE
HARVARD MEDICAL SCHOOL

- **ncRNA Detection Unit**
  - ncRNA Quantification
- **ncRNA Bioinformatics Unit**
  - ncRNA (bulk/single cell) Analysis
- **Drug delivery Unit**
  - Nanopackaging of RNA cargo





Dr. Meirav Segal

Jihoon Lim

Leinal Sejour

www.NonCodingRNA.org

Beth Israel Deaconess Medical Center

HARVARD MEDICAL SCHOOL TEACHING HOSPITAL

microRNAs

~2,800 in Humans
~500 detectable in blood
Powerful gene expression regulators
Control Pathways, transcriptional programs

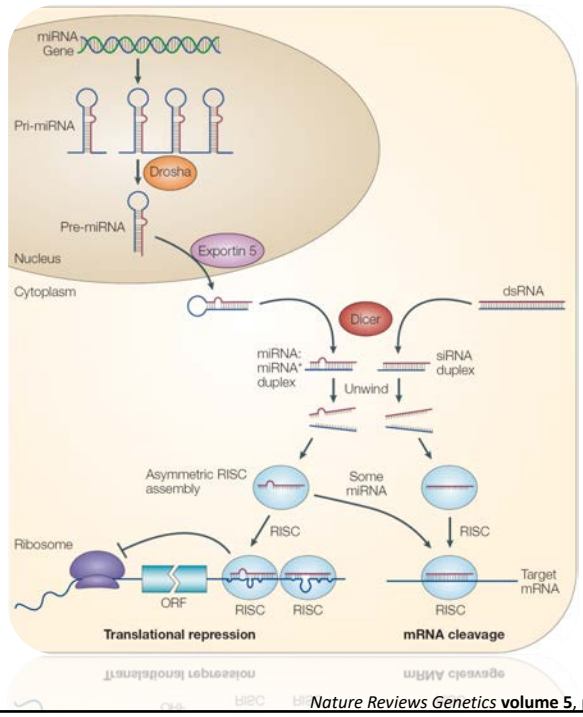
Long Non-coding RNAs

X*10,000+(???) in Humans
High Disease, Tissue, Cell specificity
Functions known for less than 100
Each with unique characteristics

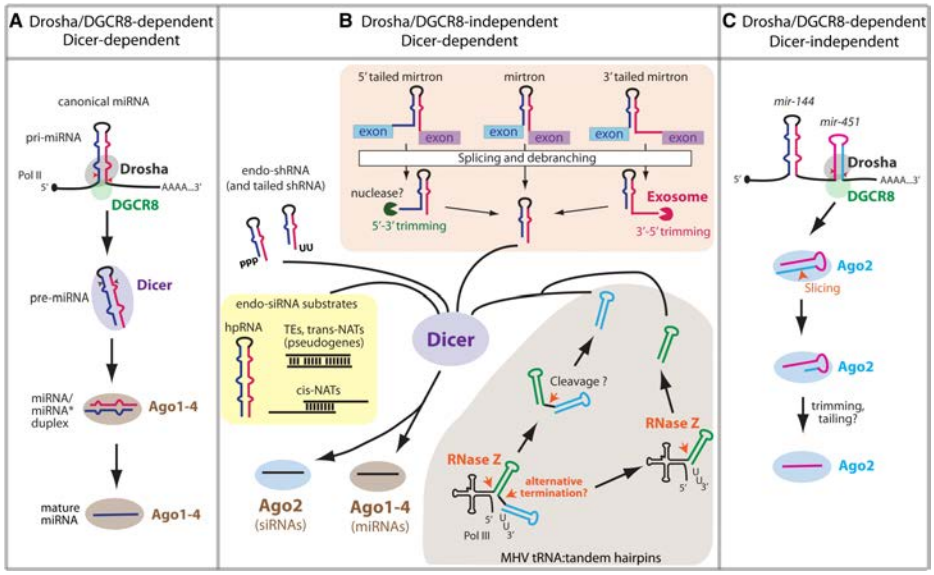
www.NoncodingRNA.org www.non-coding.org

microRNAs (miRNAs)

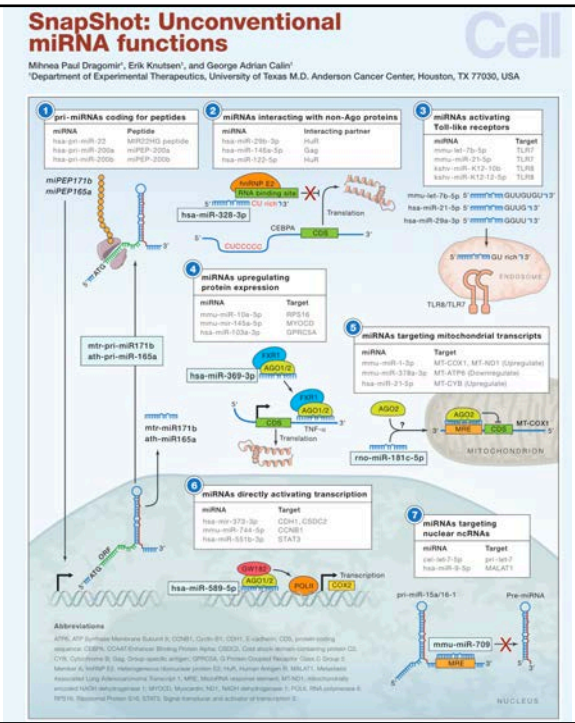
www.NoncodingRNA.org www.non-coding.org



Nature Reviews Genetics **volume 5**, pages 522–531 (2004)



Molecular Cell 2011 43, 892-903DOI: (10.1016/j.molcel.2011.07.024)



At a glance

- **microRNAs are transcribed:**
  - From intergenic or intronic loci
  - As primary miRNA transcripts (pri-miRNAs)
- **Pri-miRNAs (up to 1000nds of nts long)**
  - Are processed into **pre-miRNA hairpins** in the nucleus
- **Pre-miRNAs** (~70-100nts)
  - One arm of the miRNA hairpin turns to mature miRNA
  - The other gets degraded
- **Mature miRNAs (~22nts)**
  - They are the active miRNA form
  - Potent regulators of gene expression:
    - mRNA cleavage
    - mRNA destabilization/degradation
    - Translation Suppression

www.NoncodingRNA.org www.non-coding.org

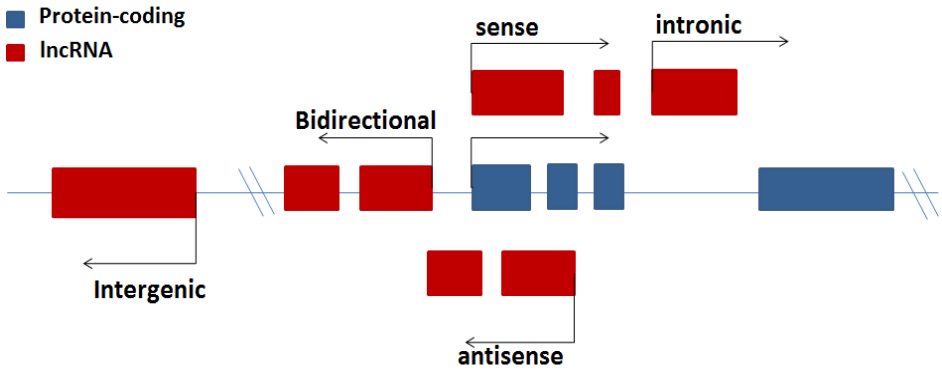
microRNAs in *Homo sapiens*

- >2,800 miRNAs identified in *Homo sapiens*
- ~500 detectable in blood
- 1 miRNA can control dozens of genes
- >60% of genes are estimated to be under miRNA regulation

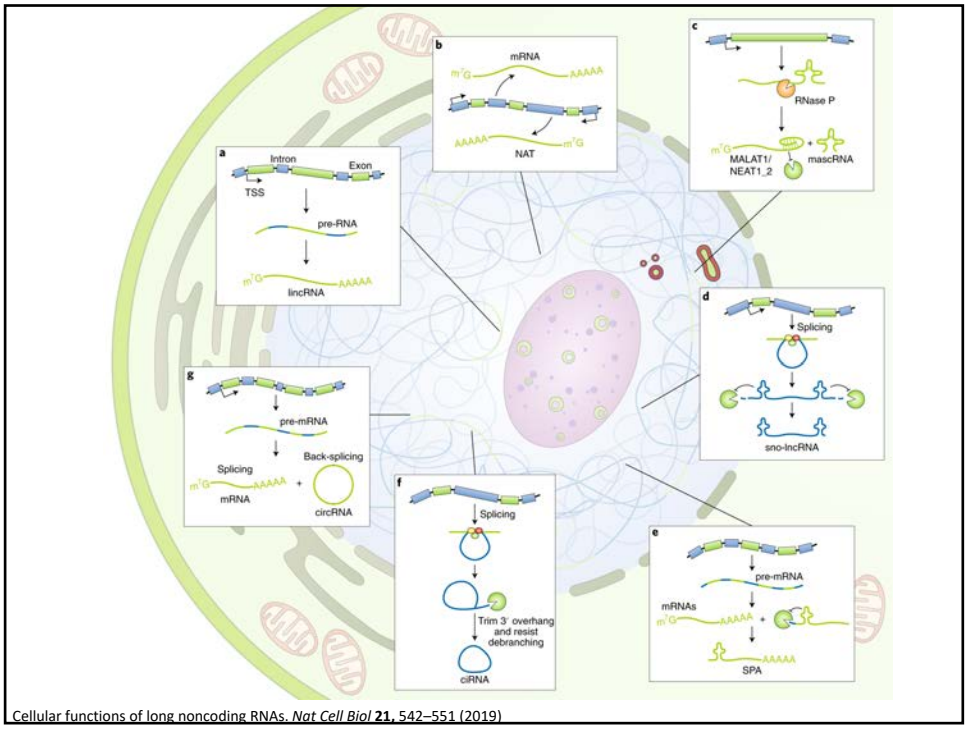
LncRNAs

- Long noncoding transcripts (>200bp)
- 10s of thousands identified in humans
- Can originate:
  - Coding genes (non coding isoforms) [sense]
  - Opposite strand of coding genes [anti-sense]
  - Same strand + opposite direction of coding genes [bi-directional]
  - In coding gene introns [intronic]
  - Far from any coding gene [intergenic]
- Take part in almost any known biological function and mechanism

LncRNA Categorization

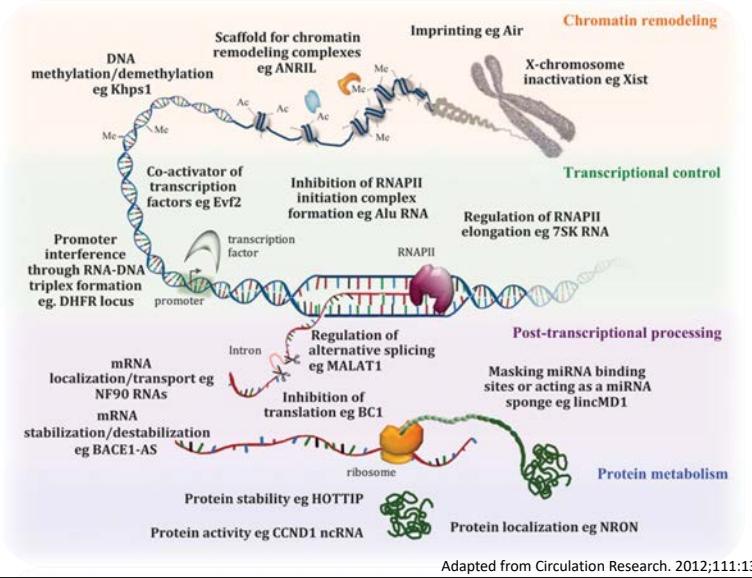


www.NoncodingRNA.org www.non-coding.org

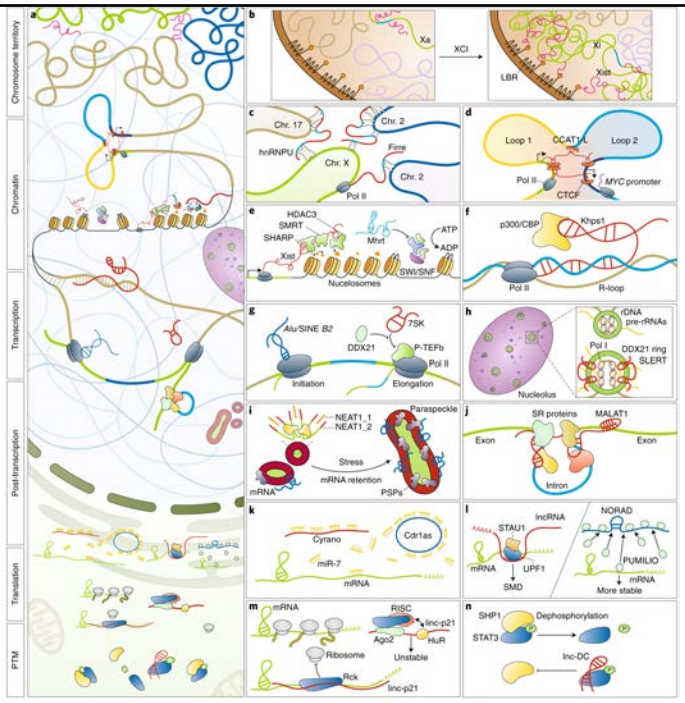


Cellular functions of long noncoding RNAs. *Nat Cell Biol* **21**, 542–551 (2019)

Participation of LncRNAs in



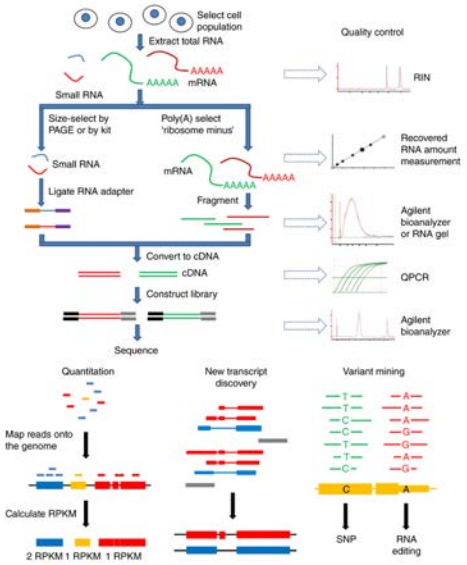
Adapted from Circulation Research. 2012;111:1349–1362



Cellular functions of long noncoding RNAs. *Nat Cell Biol* **21**, 542–551 (2019)

www.NoncodingRNA.org www.non-coding.org

Overview of the RNA-Seq Pipeline



Zeng W, Mortazavi A, Nat Immunol. 2012 Sep;13(9):802-7

RNA-Seq Analysis Pipeline

- QC of raw sequencing data
- Pre-processing
- Quantification
- Differential Expression
- Functional Analysis
- Prioritization

www.NoncodingRNA.org www.non-coding.org

Some rules of thumb

- **Aim vs Sequencing Depth**
  - miRNA Differential Expression: **5-25M** (50bp)
  - mRNA Differential Expression: **15-40M**
    - 70+ nt -long reads
    - Paired-end
    - PolyA Selection
  - mRNA + lncRNA Differential Expression: **50M+**
    - 70+ nt –long reads
    - Paired-end
    - Strand Specific
    - PolyA Selection or Ribosomal RNA Depletion?
  - Differential Exon Analysis
    - **70+M** / 70+nt / PE / SS / PolyA
  - Transcriptome Assembly + Novel Transcript Identification:
    - **100+M** / 100+nt / PE / SS / RiboDepletion

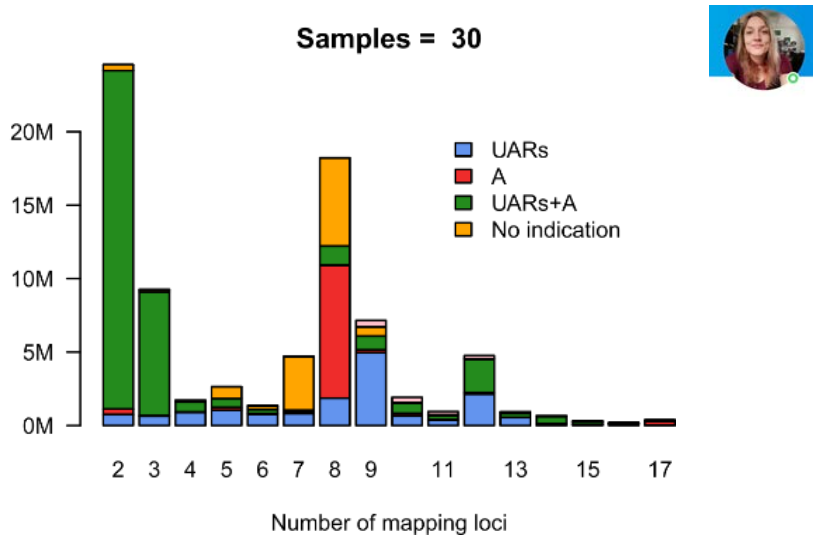
Pre-Processing

- Small RNA-Seq reads have 3' adapters
  - **Solution 1:** Ask the sequencing facility for the kit utilized in the prep
  - **Solution 2:** DNApi https://github.com/jnktsj/DNApi
    - Does not always get the adapters right but more often than not
  - **Solution 3:** Kmer assembly of 3' ends + Adapter Library
    - The assembled 3' enriched sequences are overlapped against a manually curated library of >1K adapter sequences
- **Preprocessing Recipe:**
  - **Cutadapt**
    - Clip identified adapter (12nts suffice)
    - 3nt overlap
    - 10% mismatches
    - Q10 trimming
  - Worry if:
    - You see small % of clipped reads (e.g. <60%)
    - Overrepresentation of a single nucleotide preceding the clipped adapter

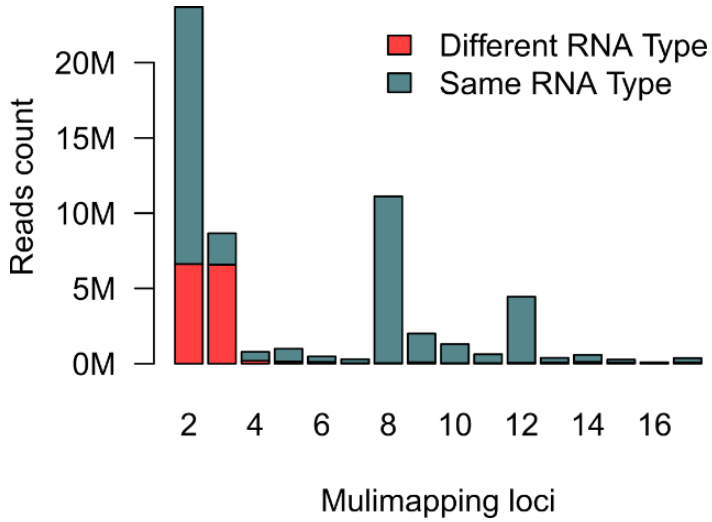
Small RNA Expression analysis of NGS Data

- Enables the detection and quantification of:
  - microRNAs
  - isomiRs
  - Other small RNA Species
  - Variants on small RNA loci
- Small RNA (including miRNA) quantification from NGS is **inherently difficult**
  - Small sequence size
    - Multimapping
  - Numerous small RNAs (including miRNAs) are transcribed from repeated genomic loci
    - Multimapping
  - SNVs, modifications, isoforms
    - Erroneous mapping

www.NoncodingRNA.org www.non-coding.org



Libraries include: hESCs, healthy tissues (heart, liver, colon, gallbladder, lung, pancreas, skin), and neoplastic (thyroid, tongue, MCF7, HepG2), alignment with bowtie v1, all mappings retained.



www.NoncodingRNA.org www.non-coding.org

Available Strategies

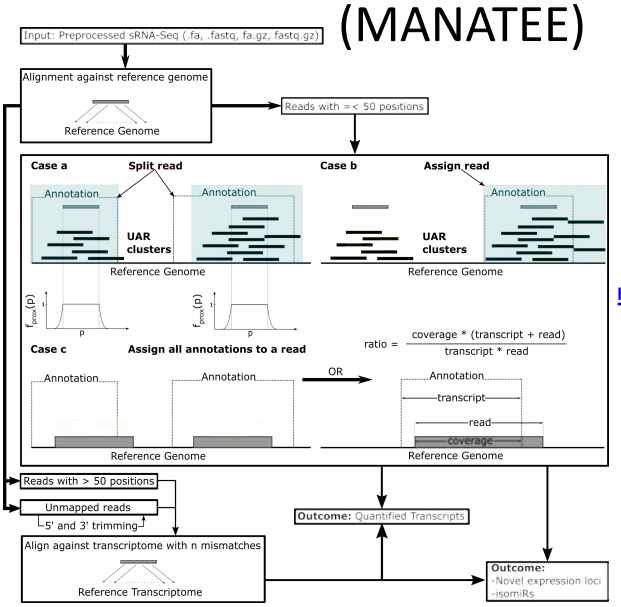
- **(A) Aligning on the Genome**
  - ***Only unique reads are counted***
    - All miRNAs/small RNAs from non unique regions cannot be quantified (very common)
  - ***All reads are counted, even when multimapping***
    - Total number of counts >> total number of reads (very common)
    - Small RNAs (miRNAs) that are not actually expressed have counts > 0 (very common)
  - ***All reads are counted but a fraction (Read/Mapping loci) is assigned***
    - Small RNAs (miRNAs) that are not actually expressed have counts > 0 (very common)
  - ***Mixed approach (e.g. TCGA)***
    - Problems from both categories (common)
- **(B) Aligning on the miRNome**
  - Reads that are from other small RNAs are assigned on miRNAs (very common)
- **(C) Hierarchical alignment on the small RNAome**
  - Same problems as on the miRNome (common in exosomal RNA data)

Representative Pipelines & Tools

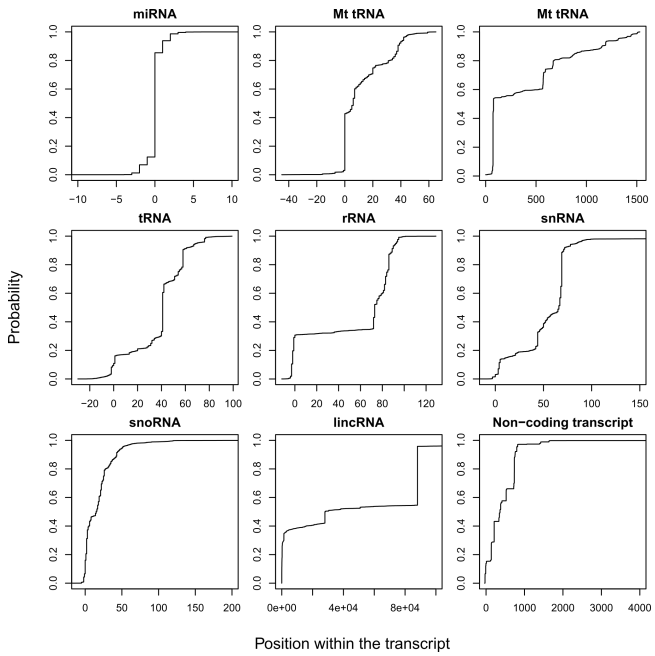
- (A**): Genomic Alignment**
  - **Bowtie v1 || BWA + FeatureCounts** or **HTSeq**
  - **sRNABench** (https://bioinfo2.ugr.es/ceUGR/srnabench/)
  - **ShortStack** (https://github.com/MikeAxtell/ShortStack)
- **(B): miRNAome**
  - **miRDeep2** (https://github.com/rajewsky-lab/mirdeep2)
  - **BLAST** against miRNA annotation
- (C): **Small RNAome**
  - **miRGe** (https://github.com/mhalushka/miRge)
  - **eXCeRpt** (http://github.gersteinlab.org/exceRpt/)

www.NoncodingRNA.org www.non-coding.org

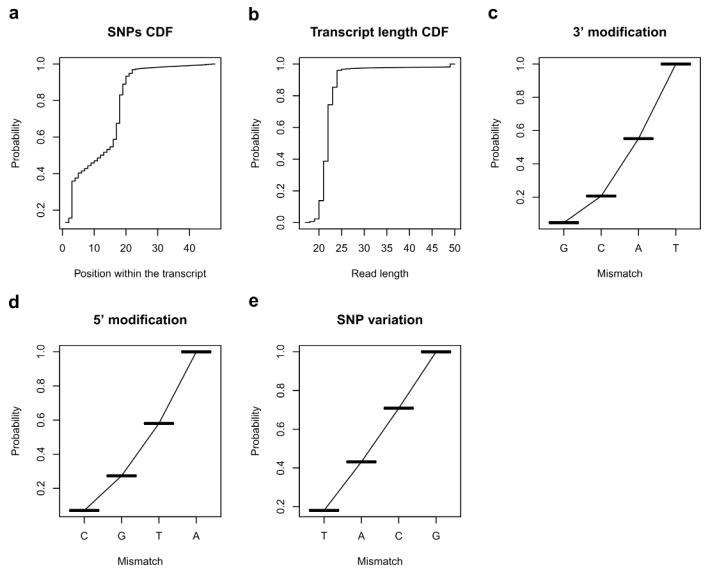
sMAll rNa dATa analysis pipElinE (MANATEE)



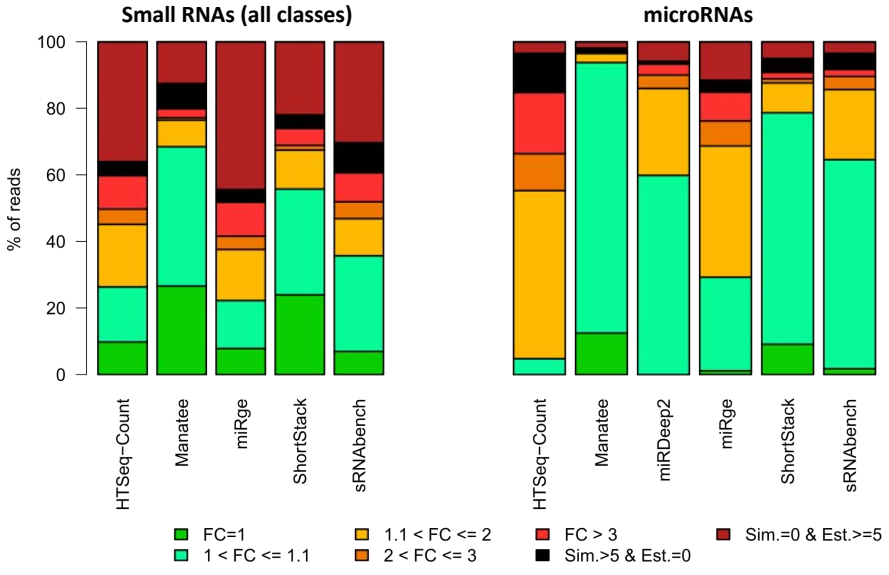
https://github.com/ivlachos/Manatee1.0



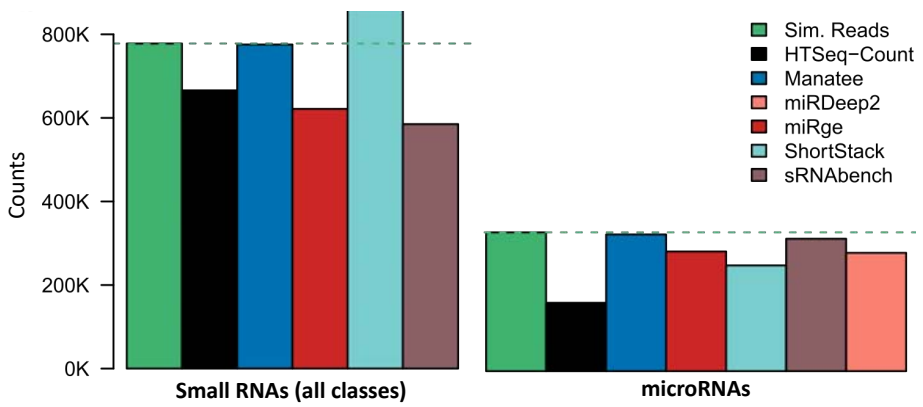
Handzlik, et al., Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data (in press)



Handzlik, et al., Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data (in press)



Handzlik, et al., Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data (in press)

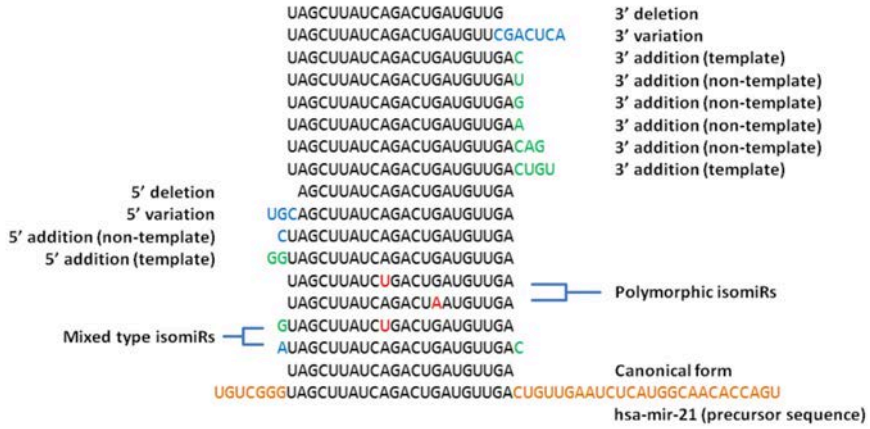


Handzlik, et al., Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data (in press)

How To Run

- https://github.com/ivlachos/Manatee1.0
- **Input:**
  - Clipped and trimmed fastq files
  - Genome
  - Small RNA annotation in GTF format (provided for human)
- **Output:**
  - Small RNA expression counts table
  - Small RNA unique reads table
  - Unknown expressed loci annotation + expression
  - IsomiR sequences and expression

IsomiRs



Wu, et al., BMC Genomics volume 19, Article number: 401 (2018)

IsomiRs Affect

- **miRNA Targeting**
  - 5' Additions (templated / non-templated)
  - 5' SNVs
- **miRNA Fate** (?) (localization, degradation)
  - 3' Additions (templated / non-templated)
- **Experimental Setup**
  - qPCR primers

www.NoncodingRNA.org www.non-coding.org

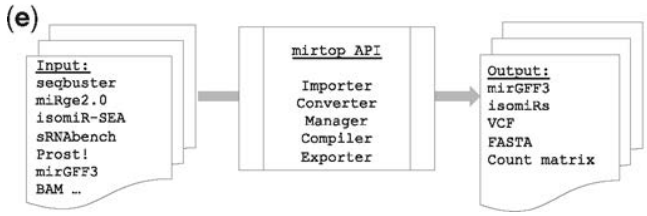
miRTop

Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API

- Community project to:
  - Set the standards for isomiR nomenclature
  - Provide tools for isomiR annotation
  - Create an API compatible with available tools and approaches
  - Provide a uniform solution for all researchers to save miRNA quantification results

https://github.com/miRTop/mirtop

How does it work?



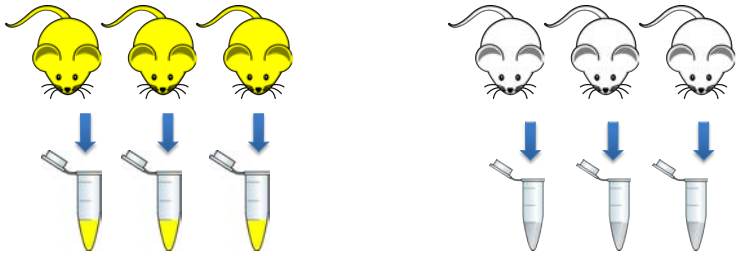
- Steps:
  - Quantify with your favorite method
  - Import the output to miRTop
  - Extract quantification + annotation of all miRNAs at isomiR resolution
    - GFF3
    - Count Matrix
    - VCF
    - FASTA
- https://github.com/miRTop/mirtop

LncRNA Expression Quantification with NGS

- **Different Preps -> different lncRNAs**
  - PolyA
  - RiboDepletion (various flavors)
- **Annotation**
  - GENCODE
  - Expanded annotation
    - GENCODE, LnciPedia, FANTOM, NONCODE
    - Annotation has to be curated
  - *De Novo*
    - *StringTie*
    - *Trinity*
- **Pre-processing**
  - Often low % of adapter content
- **Quantification**
  - STAR + RSEM
  - Salmon (+GC, fragment length bias control)
  - Kallisto
  - StringTie

www.NoncodingRNA.org www.non-coding.org

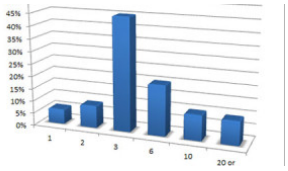
Differential Expression Analysis

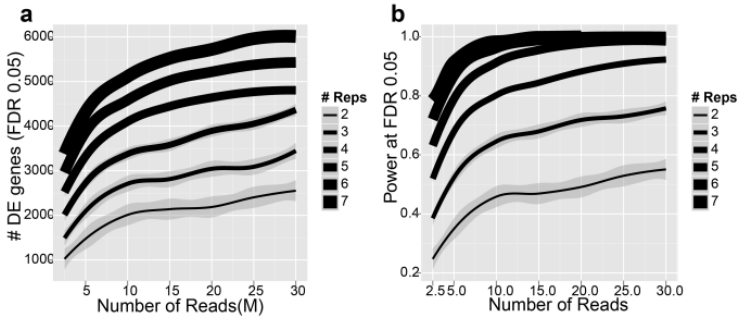


+ Library Prep

+ =?

Replicates





DE Genes at FDR<0.05 **Statistical Power at FDR<0.05**

microRNAs

"microRNAs do not behave similarly to longer RNAS"
Ioannis Vlachos, Dec 12 2019

- A few microRNAs hog all the reads
- The rest have to squeeze in what's left
- Functional impact does not always correlate with expression levels
- Depending on your aim, you might have to dig deeper

Differential Expression: sRNAs

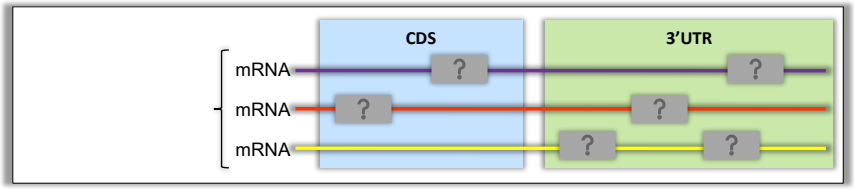
- DE of miRNAs / sRNAs is not as well standardized as for longer transcripts
- There are issues in most steps:
  - **Normalization**
    - miRNA only OR miRNA + small RNAs?
    - Algorithms for Spike-ins or UMIs?
    - TMM / RLE / Voom don't seem to function as planned
  - **Differential Expression**
    - Standard approaches (e.g. DESeq2, limma) seem not to perform optimally
- Suggestions:
  - miRNAs + sRNAs seem to capture better the expression space
  - High Numbers of replicates always increase power
  - DESeq seems to rank properly but too conservative
  - EdgeR shows an inflation in Type I error
  - DESeq2 seems to perform better with miRNAs + small RNAs together but struggles in exosomal RNA or in cases with extremes

Differential Expression: RNAs

- **Tip: *Do Not Remove ncRNAs from the genes table*.**
  - Missing important biological findings
  - Having lncRNAs, increases hits for mRNAs
- **Differential Expression**
  - Limma (when having high numbers of replicates)
  - DESeq2 (when replicate number is small)
  - EdgeR
- **Differential Isoform Expression**
  - Ballgown (for StringTie results)
  - Sleuth (for Kallisto / Salmon results)
  - SUPPA2 (https://github.com/comprna/SUPPA)
  - stageR (https://rdrr.io/bioc/stageR/)
  - DEXSeq (https://rdrr.io/bioc/DEXSeq/)
- **Batch Effect / Covariate Correction**
  - COMBAT (https://rdrr.io/bioc/sva/man/ComBat.html)
  - Surrogate Variable Analysis (https://rdrr.io/bioc/sva)
  - Expression PCs (log2+0.1)
  - PEER (https://www.sanger.ac.uk/science/tools/peer)

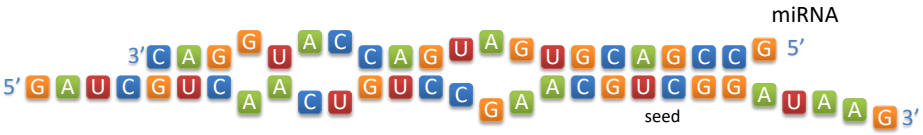
www.NoncodingRNA.org www.non-coding.org

miRNA binding



www.NoncodingRNA.org www.non-coding.org

microRNA Targets



Computational methods for miRNA target prediction usually form the backbone of most experimental or *in silico* miRNA-related pipelines
Vlachos, Hatzigeorgiou, 2013, Clinical Biochemistry 46: 879-890

microRNA Target Predition Algorithms

TargetScan
DIANA-microT-CDS
miRanda
PicTar

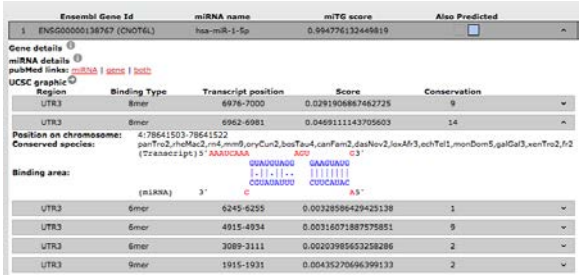
microT-CDS www.microrna.gr/microT-CDS

Target Site Accessibility
Physicochemical Properties
Sequence Properties
MRE conservation
Pair stability
Machine Learning

They nonetheless explain only a small fraction of the mRNA changes observed upon introducing a miRNA ($r^2 = 0.14$) (TargetScan) McGeary, et al, Science, Dev 5 2019

www.NoncodingRNA.org www.non-coding.org

miRNA Targets (Single Interactions)

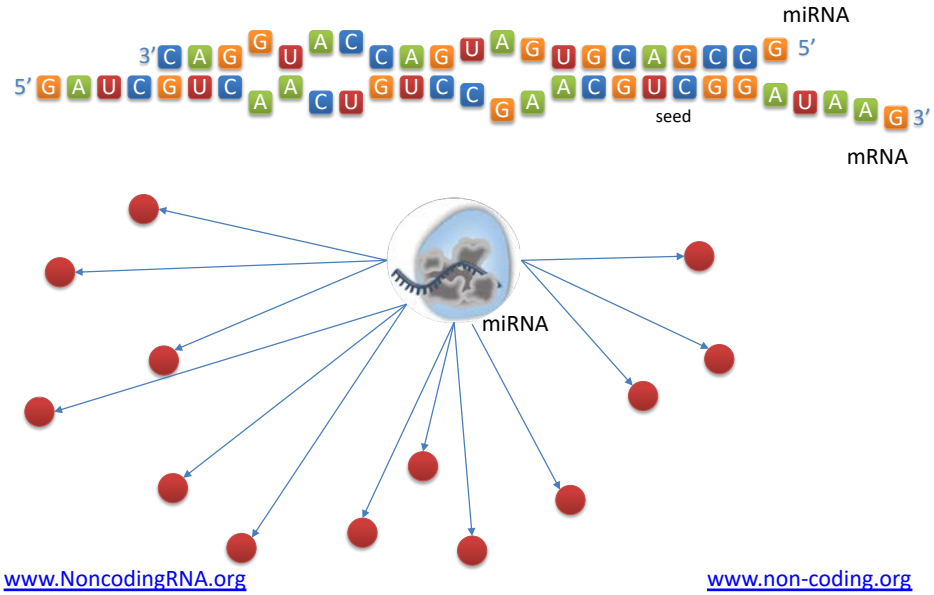


Apart from the prediction score:

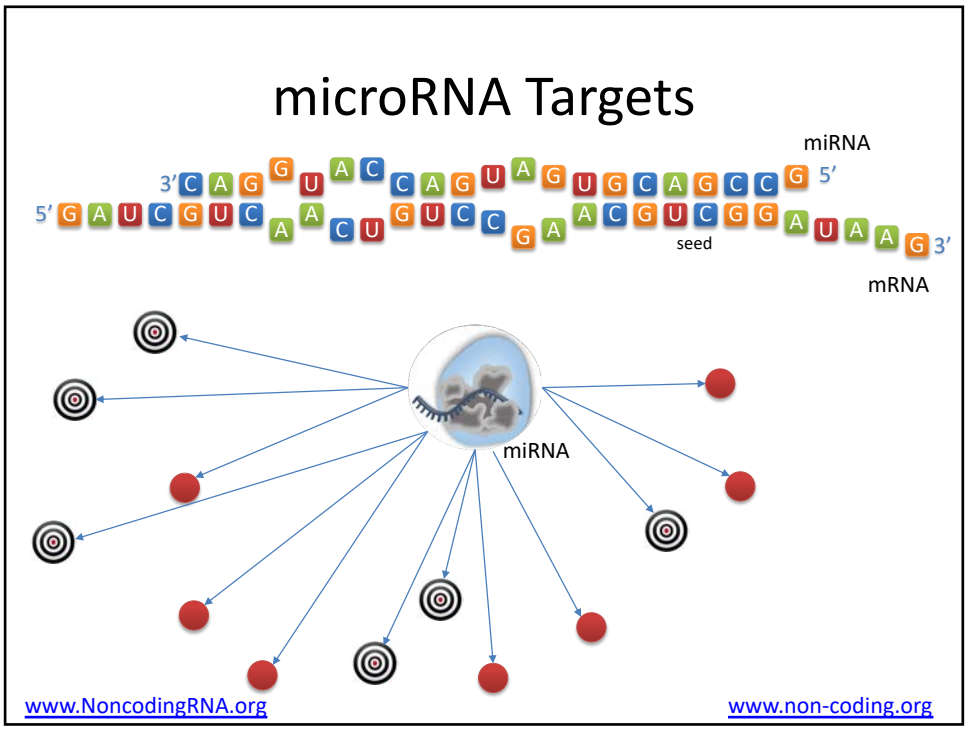
- miRNA binding type (8mer, 7mer, etc)
- Binding site conservation*
- Experimental support from high throughput experiments
- Top interactions from algorithms that prioritize seed-based binding events (e.g. TargetScan) often have higher effect sizes
- Often multiple experimental methods are required to validate a binding event
- miRNAs can also work to stabilize expression and not only to down-regulate

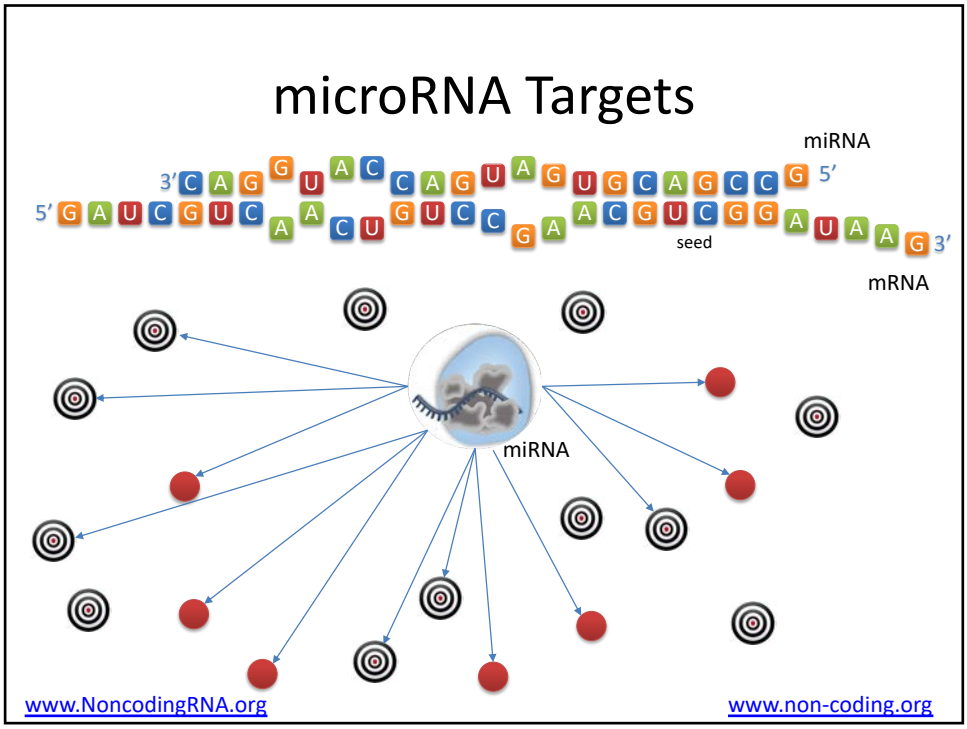
Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. (2013), DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic acids research 41: W169-173.

microRNA Targets



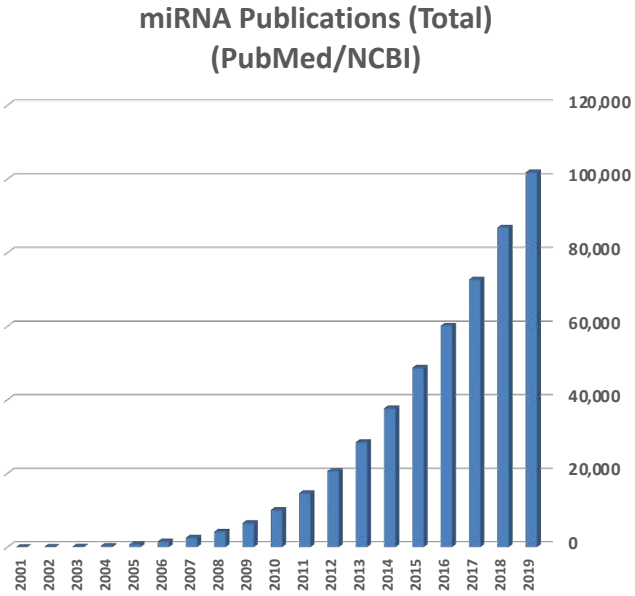
www.NoncodingRNA.org www.non-coding.org


microRNA Targets
miRNA
3′ C A G G U A C C A G U A G U G C A G C C G 5′
5′ G A U C G U C A A C U G U C C G A A C G U C G G A U A A G 3′
seed
mRNA
miRNA
www.NoncodingRNA.org
www.non-coding.org


microRNA Targets
miRNA
3′ C A G G U A C C A G U A G U G C A G C C G 5′
5′ G A U C G U C A A C U G U C C G A A C G U C G G A U A A G 3′
seed
mRNA
miRNA
www.NoncodingRNA.org
www.non-coding.org

Validating miRNA Interactions

Methods
Reporter Genes
Northern Blotting
qPCR
Western Blot
ELISA
5′ RLM-RACE
Microarrays
RNA-Seq
Quantitative Proteomics
AGO-IP
HITS-CLIP / PAR-CLIP
CLASH / PAR-CLIP + Ligation
Biotin miRNA tagging
IMPACT-Seq
PARE / Degradome-Seq
3Life
miTRAP



Method	Throughput	Intended use
Reporter Genes	Low	Validation of miRNA:UTR (or binding region) interaction
Northern Blotting	Low	Relative effect of miRNA on mRNA levels
qPCR	Low	Quantification of miRNA effect on mRNA levels
Western Blot	Low	Relative assessment of miRNA effect on protein concentration
ELISA	Low	Quantification of miRNA effect on protein concentration
5′ RLM-RACE	Low	Identification of cleaved mRNA targets
Microarrays	High	High-throughput assessment of miRNA effect on mRNA expression
RNA-Seq	High	High-throughput assessment of miRNA effect on mRNA expression
Quantitative Proteomics (e.g. pSILAC)	High	High-throughput assessment of miRNA effects on protein concentration
AGO-IP	High	Identification of enriched transcripts (miRNAs and mRNAs) in AGO immunoprecipitates
HITS-CLIP	High	Sequencing of AGO binding regions on targeted transcripts
PAR-CLIP	High	Sequencing of AGO binding regions on targeted transcripts
CLASH / PAR-CLIP + Ligation	High	Sequencing of AGO binding regions on targeted transcripts. Production of chimeric miRNA:mRNA reads for the identification of interacting pairs.
Biotin miRNA tagging	High/Low	Pull-down of biotin-tagged miRNAs and estimation of bound transcript content using qPCR (Low yield), microarrays (High-throughput) and RNA-Seq (High-throughput)
IMPACT-Seq	High	Pull-down of biotin-tagged miRNAs, identification of interacting pairs and binding regions.
PARE / Degradome-Seq	High	High-throughput identification of cleaved mRNA targets
3Life	High	High-throughput reporter gene assay
miTRAP	High	miRNA trapping by RNA baiting

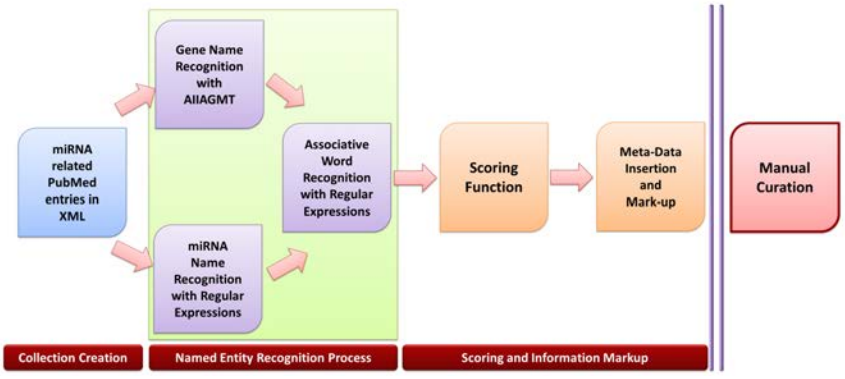
Vlachos IS, *et al.*, (2015). Nucleic Acids Res 43: D153-159

Manually curated Databases of miRNA targets with Experimental Support

- **TarBase** (www.microrna.gr/tarbase)
  - >1M interactions
  - >600K unique miRNA:target pairs
  - ~600 cell types
  - ~451 experimental conditions
- **miRTarbase** (http://mirtarbase.mbc.nctu.edu.tw)
  - 470,000 interactions
  - 11K manuscripts

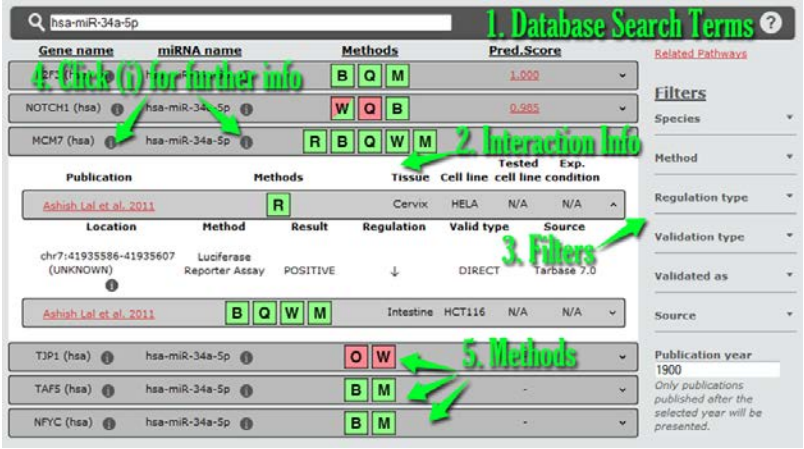
www.NoncodingRNA.org www.non-coding.org

TarBase Curation Pipeline



Vergoulis T*, Vlachos IS*, *et al*, (2012), **Nucleic Acids Research** 40: D222-D229.

TarBase: The DB of miRNA:mRNA interactions with Experimental Support



www.microrna.gr/tarbase

Vlachos IS, *et al.*, (2015). Nucleic Acids Res 43: D153-159

Using TarBase

- Experimentally Supported interactions exhibit:
  - Higher signal/noise ratio vs predictions
  - Provide detailed information regarding experimental conditions
  - Can denote:
    - Effects on expression
    - Protein levels
    - Direct / indirect binding events
- **However:**
  - Literature Bias
  - High Throughput methods favor highly expressed miRNAs
  - CLIP-Seq datasets are notoriously complex to analyze

www.NoncodingRNA.org www.non-coding.org

How to Incorporate miRNA Targets (High Throughput)

- ***In Silico* Predictions**
  - TargetScan & microT Targets can be downloaded for local use
  - TargetScan annotation is miRNA family-based
  - TargetScan can be downloaded to run locally
  - miRanda is an algorithm that can run locally without much effort
  - Prediction Thresholds are not universal and have to be tailored to the species/analysis type
- **Experimentally Supported Interactions**
  - TarBase can be downloaded following application (free)
  - miRTarBase can also be downloaded
  - Experiments are not equally informative or accurate
  - Sometimes validated interactions can be expanded with predictions
- **miRNA interactions exhibit high tissue and cell type specificity**

www.NoncodingRNA.org www.non-coding.org

www.NoncodingRNA.org www.non-coding.org

Revisiting AGO CLIP-Seq Analysis

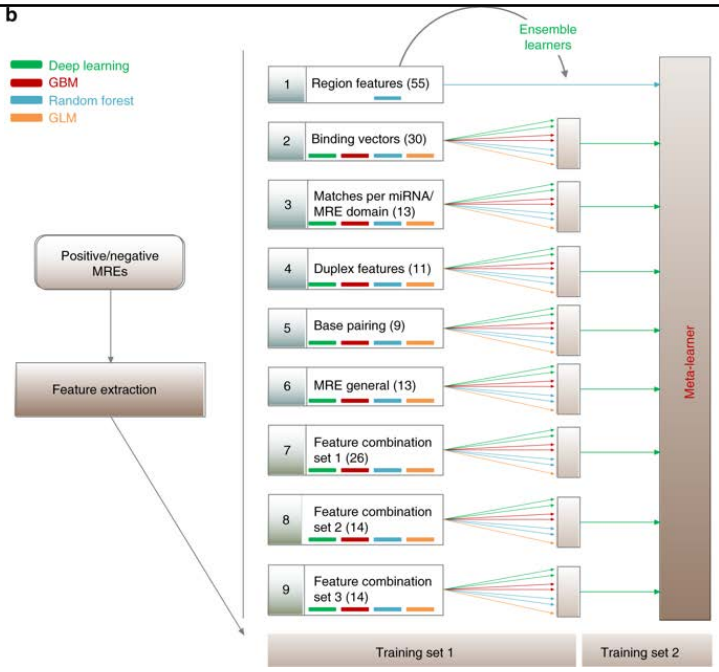
High Confidence Dataset (n=5,774)

Reporter Gene Assays
CLASH (crosslinking ligation and sequencing of hybrids)
CLEAR-CLIP (covalent ligation of endogenous Argonaute-bound RNAs)
Chimeric CLIP-Seq

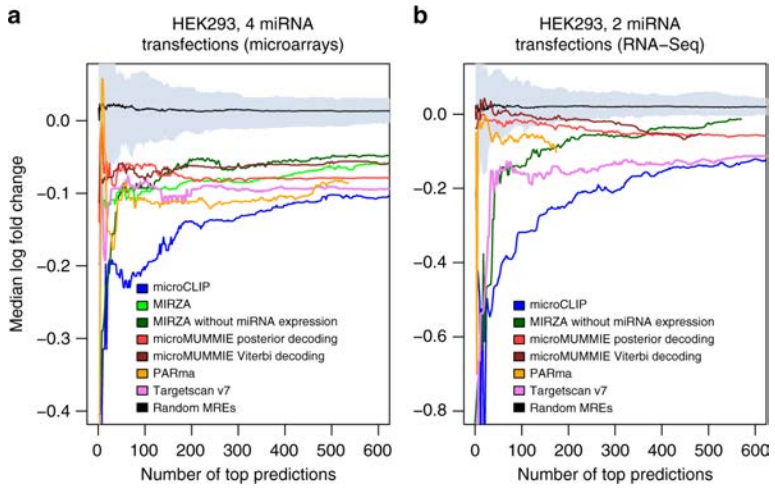
Functional Interactions Dataset (Perturbation Experiments n=12,400)

RNA: 12 RNA-Seq & 85 microarrays (n=3,900)
Translation: 5 ribosome profiling & 5 (pSILAC) (n=5,900)
Binding: AGO-IP & Biotin pull-down (n=2,600)

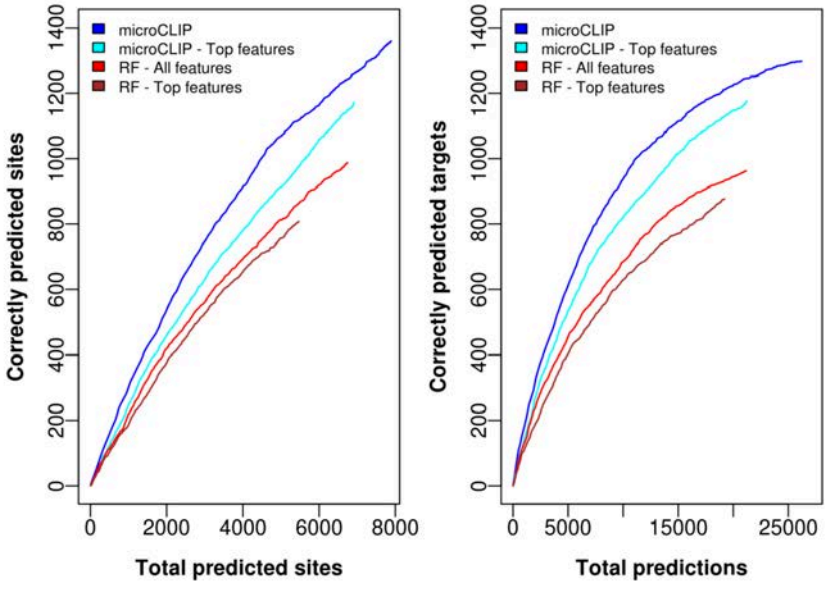
microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions. Nature Communications (2018).



microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions. Nature Communications (2018).



microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions. Nature Communications (2018).



microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions. Nature Communications (2018).

Using microCLIP

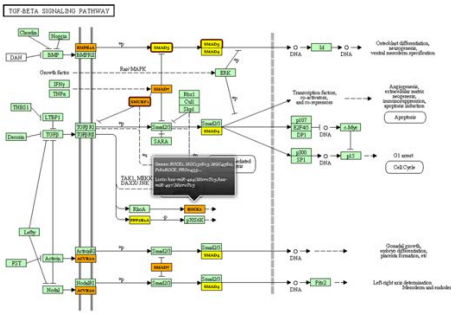
- Getting microCLIP: www.microrna.gr/microCLIP
- **Input:**
  - 1. FASTA file of the miRNA sequences.
  - 2. Alignment file in BAM/SAM format.
  - 3. PhastCons (from UCSC)
  - 4. Gene annotation file in one-based BED-like format (OPTIONAL)
- **Configuration**:
  - Set up the parameters in init.R source code.
- **Usage**:
  - Rscript main.R
- **Output**
  - MRE coordinates
  - miRNA
  - score: The MRE predicted score.
  - binding_type
  - binding_class (canonical/non-canonical).
  - cluster_type (TC/non-TC).
  - overlapping.reads

www.NoncodingRNA.org www.non-coding.org

microRNA Functional Investigation

Complex Relationships

microRNA:genes
Gene:microRNAs
Gene:pathway(s)
Pathway:microRNA(s)
microRNA:pathway(s)



Modelling Challenges

Groups of microRNAs
Biases in the annotation
Target distributions
Independence

Functional Enrichment

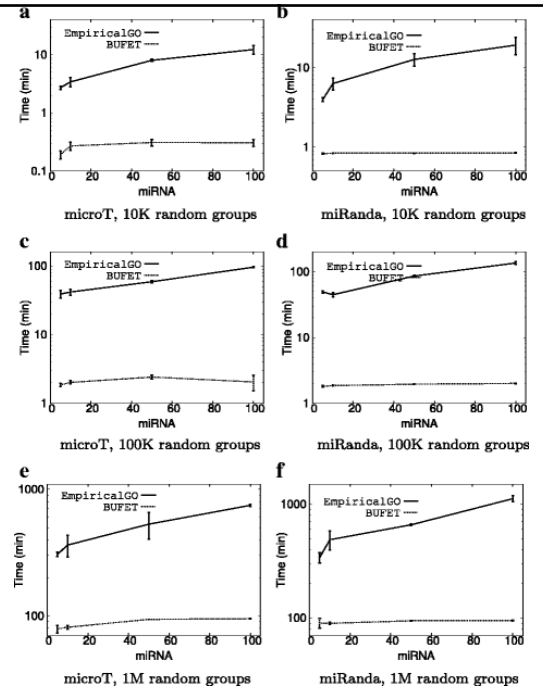
- **Hypergeometric Test** of miRNA targets against annotation sets (e.g. KEGG, GO)
  - Requires independence of events which does not occur for miRNA targets (structured data)
  - Combining targets of >4-5 miRNAs results in sets with thousands of genes
- **Empirical Tests**
  - Are able to capture the structure and return the enrichment at miRNA level
  - Permutations need time and computational resources
- **Combining miRNA Activity**
  - Keeping each target once (most common)
  - Weighted heuristics (ranking but not statistics)
  - Meta-Analysis statistics (return pathway-level significance)

www.NoncodingRNA.org www.non-coding.org

How To

- **miRNA Target sources**
  - *In Silico*
    - TargetScan
    - microT-CDS
  - Experimentally Supported
    - TarBase
    - miRTarBase
- **Gene Sets**
  - KEGG
  - Gene Ontology
  - MSigDB
- **Enrichment**
  - One-sided Fisher's exact test (>)
  - Empirical-GO (http://sgjlab.org/empirical-go/)
  - BUFET (https://github.com/ivlachos/BUFET)
  - Meta-Analysis tests (e.g. Fisher's combined probability test, Stouffer's Z, etc)
- **Online Implementations**
  - miRPath (www.mirorna.gr/miRPath)

www.NoncodingRNA.org www.non-coding.org



Zagganas et al, BUFET: boosting the unbiased miRNA functional enrichment analysis using bitsets. BMC Bioinformatics (2017)

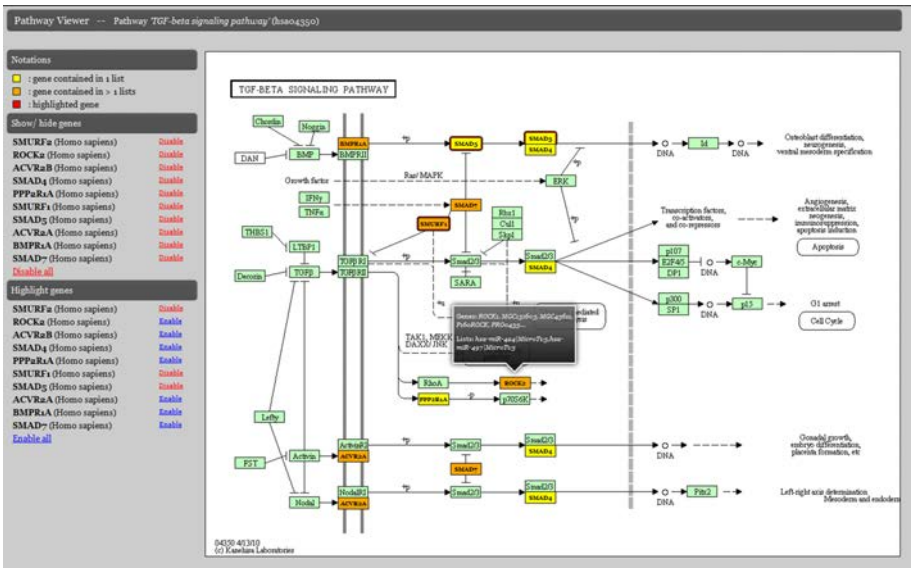
Online Resources: miRPath v3.0

- **Interactions**
  - Experimentally Supported (TarBase v7.0)
  - Predicted (microT-CDS, TargetScan)
- 7 **species** (Hsa, Mmu, Rno, Dme, Cel, Dre, Gga)
- **Statistics**
  - Fisher's exact test
  - Empirical Distributions
  - Meta-analysis statistics (Fisher's Combined Probability Method)
- Automated **plot** generation
- **Real time** functionality

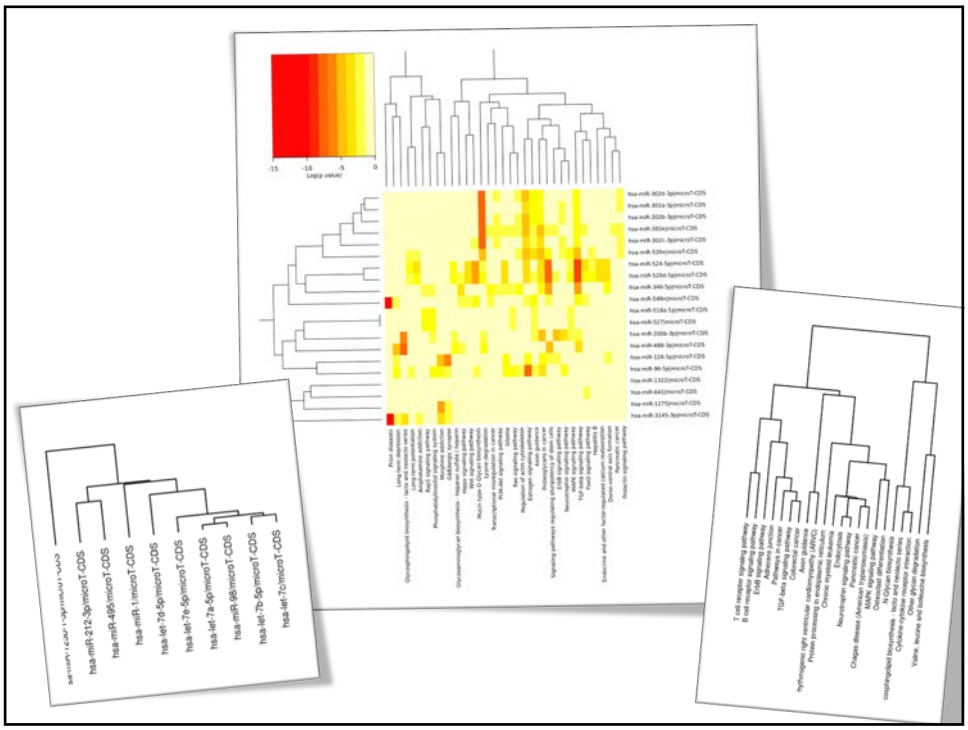
www.NoncodingRNA.org www.non-coding.org

Vlachos IS, et al, (2015), DIANA-miRPath v3.0: Deciphering microRNA function with experimental support, Nucleic Acids Res 43: W460-466.

A (friendly) microRNA Analysis Powerhouse



www.NoncodingRNA.org www.non-coding.org




sRNA Expression
1
lncRNA Expression
2
Diff Expression
3
miRNA Targets
4
AGO CLIP-Seq
5
6
miRNA Function
7
LncRNA Function
8
Promoters
9
Variation
10
Integration
www.NoncodingRNA.org
www.non-coding.org

LncRNA Function

- Not as straightforward
- LncRNAs function can be:
  - In *cis/trans*
  - By direct lncRNA activity (e.g. binding)
    - Chromatin
    - miRNAs (sponging)
    - mRNAs
    - Proteins
    - …
  - In the nucleus
  - In the cytoplasm
  - …
- There is no uniform way to investigate the function of a novel/unannotated (majority) lncRNA

LncRNA Annotation

- LNCIPedia (http://www.lncipedia.org/)
- NONCODE (http://www.noncode.org/)
- LncRBase (http://bicresources.jcbose.ac.in/zhumur/lncrbase/)
- lncRNAMap (http://lncrnamap.mbc.nctu.edu.tw/php/)
- LncRNAWiki (http://lncrna.big.ac.cn/)
- LncRNAome (genome.igib.res.in/lncRNome)

Dissecting an LncRNA

- Coding Potential
- Localization
- Coexpression
- miRNA sponging
- Circularization
- Chromatin binding

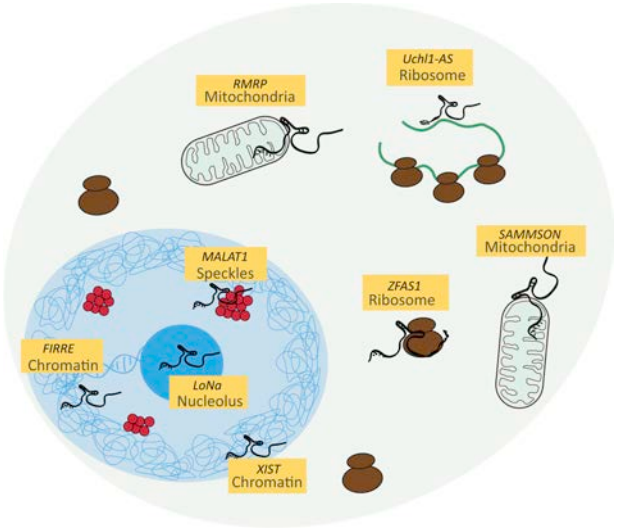
www.NoncodingRNA.org www.non-coding.org

Coding Potential

	Approach	ORF length	Protein homology	Conservation	Nucleotide composition	Substitution ratio (dN/dS)	Secondary structure	sORF detection	Coding/ noncoding prediction	P value
CONC	SVM	✓	✓	✓	✓		✓		✓	
CPC	SVM	✓	✓						✓	
PORTRAIT	SVM	✓			✓					
sORF finder	–				✓	✓		✓	✓	✓
PhyloCSF	EM			✓		✓		✓		
RNAcode	–		✓	✓		✓		✓		✓
CNCI	SVM	✓			✓			✓		
CPAT	Logistic regression	✓			✓					
iSeeRNA	SVM	✓	✓	✓	✓				✓	
PLEK	SVM				✓				✓	
Linc-SF	GA-SVM				✓		✓		?	?
LncRNA-ID	Balanced random forest	✓	✓						?	
lncRNA-MFDL	Deep stacking network	✓			✓				?	?
CPC2	SVM	✓			✓			✓	✓	
COME	Balanced RF			✓	✓		✓		✓	

Choi et al., The small peptide world in long noncoding RNAs, *Briefings in Bioinformatics, 2018*

Localization



Molecular Cell 2019 73, 869-883DOI: (10.1016/j.molcel.2019.02.008)

LncRNA Localization

- Prediction
  - **iLoc-lncRNA (**http://lin-group.cn/server/iLoc-LncRNA)
  - Sequence-based localization prediction
- **Databases**
  - **lncSLdb** (http://bioinformatics.xidian.edu.cn/lncSLdb)
    - 11K transcripts from 3 species
    - Literature-based
  - **lncATLAS** (http://lncatlas.crg.eu/)
    - 7K transcripts (human)
    - ENCODE data reanalysis
  - **RNALocate** (http://www.rna-society.org/rnalocate/)
    - ~2K (multiple species)
    - Literature-based

www.NoncodingRNA.org www.non-coding.org

LncRNA Interactions

- **miRNAs**
  - ***LncBase*** (www.microrna.gr/LncBase)
    - Predictions (expanded microT model)
    - Experimentally-supported interactions
  - ***Linc2GO*** (http://www.bioinfo.tsinghua.edu.cn/~liuke/Linc2GO/index.html)
- **RNAs**
  - ***LncRRIsearch*** (http://rtools.cbrc.jp/LncRRIsearch/) [complementarity-based predictions]
- **Diverse**
  - ***LncRNA2Target v2.0 (http://123.59.132.21/lncrna2target)***
    - lncRNA expression perturbations
    - Binding assays
    - ~150K lncRNA-target associations
  - ***LncTarD*** (http://biocc.hrbmu.edu.cn/LncTarD/)
    - **Manually curated interactions**
    - **475 lncRNAs - 1K targets, 2.8K interactions**
  - ~~***LncRNAtor*** (http://lncrnator.ewha.ac.kr)~~
  - ~~***LnRNAdb*** (http://lncrnadb.org)~~

www.NoncodingRNA.org www.non-coding.org

DIANA-LncBase

www.microrna.gr/lncBase

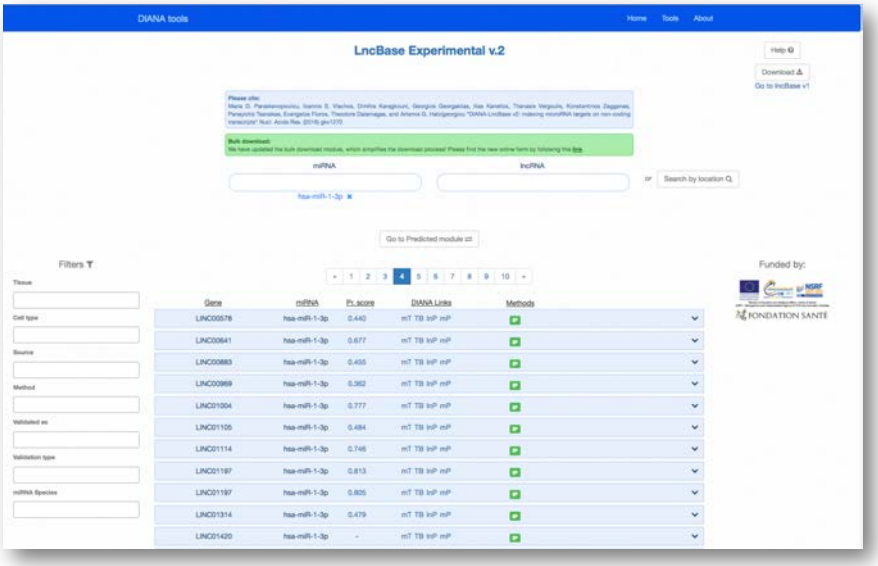
A comprehensive collection miRNA targets on lncRNAs

Accurate lncRNA expression by analyzing more than 6 billion RNA-Seq reads

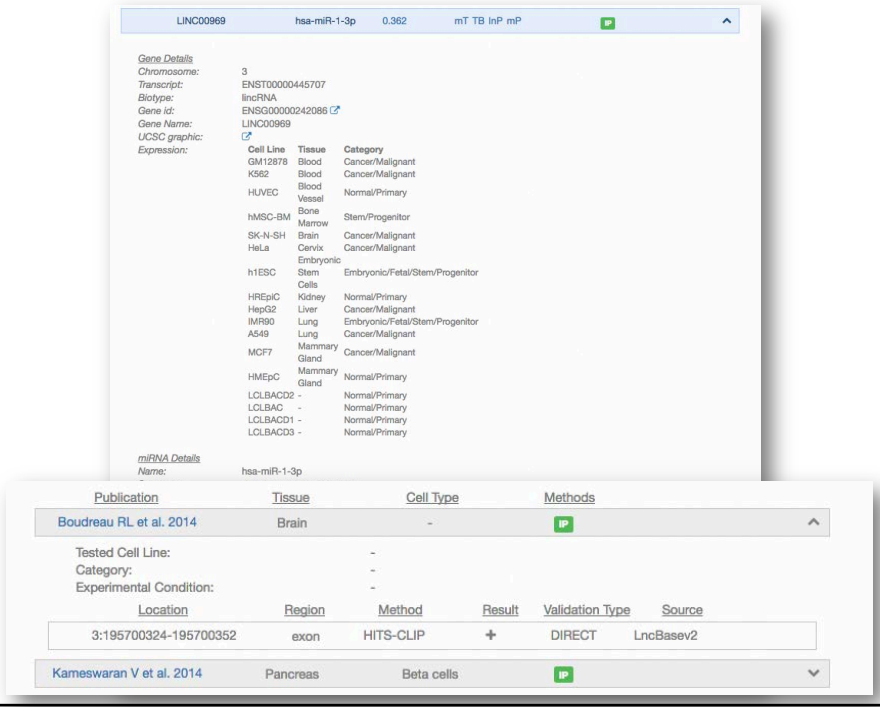
Updated *in silico* target predictions with modified microT algorithm

Experimentally verified miRNA:lncRNA targets with high-throughput or specific experiments

Paraskevopoulou MD, Vlachos IS, *et al*, Nucleic Acids Res (2015, 44, D231-238).



www.NoncodingRNA.org www.non-coding.org



Co-expression / Guilty by Association

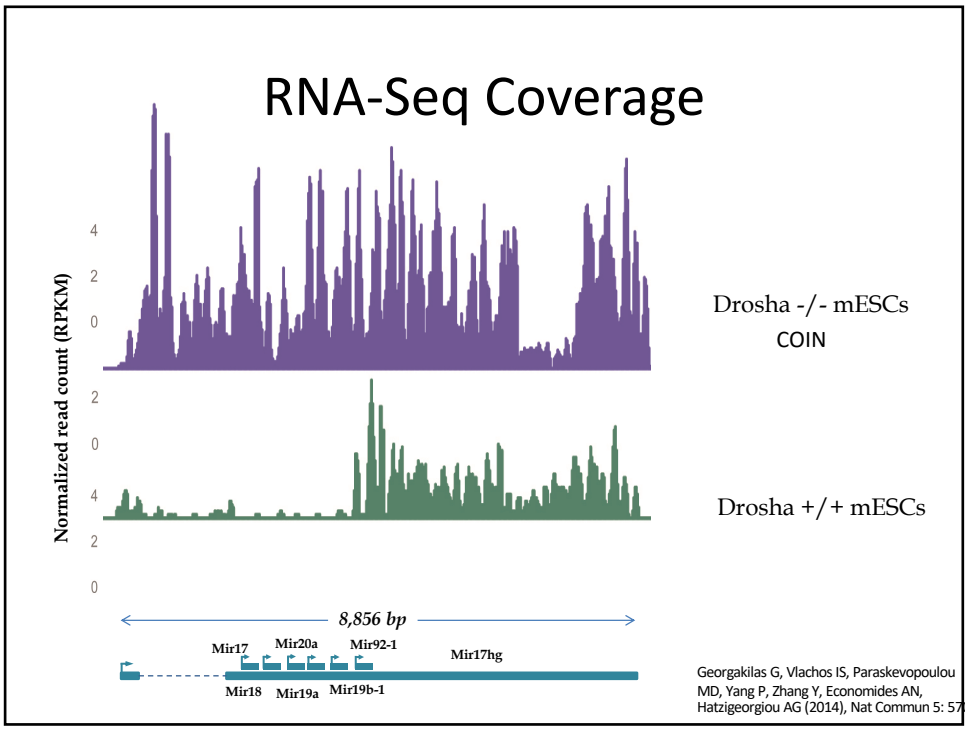
- **Statistical Metrics**
  - Spearman's Rho
  - Pearson CC
- **Geometrical Metrics**
  - Sobolev metric
  - Fisher information
- **Examples:**
  - **mRNAs**
    - Co-LncRNA (http://www.bio-bigdata.com/Co-LncRNA)
    - LncRNA2Pathways (https://cran.r-project.org/web/packages/LncPath/)
  - **TFs and TcoFs**
    - FARNA (http://cbrc.kaust.edu.sa/farna)

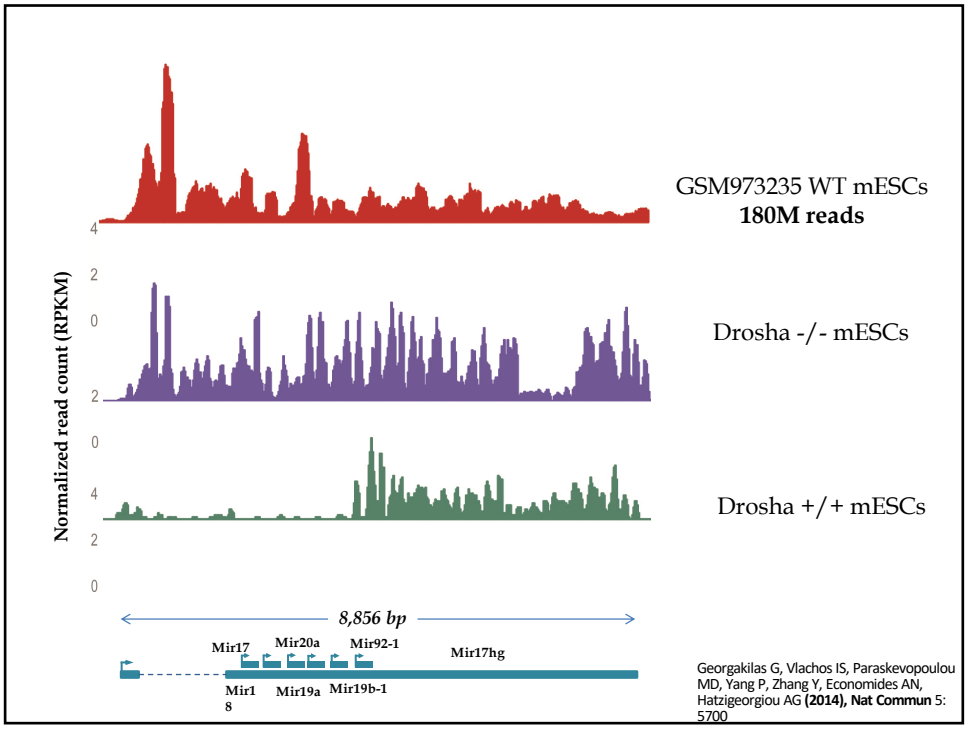
Phenotypes

- **LncRNADisease 2.0** (http://www.rnanut.net/lncrnadisease/)
  - 200K lncRNA-disease associations
- **Lnc2Catlas** (http://lnc2catlas.bioinfotech.org/)
  - lncRNA associations with cancer risk
  - SNPs & cancer protein co-expression
- **CRlncRNA** (http://crlnc.xtbg.ac.cn/)
  - Manually curated lncRNA-cancer associations
  - 1K entries
- **TANRIC** (http://bioinformatics.mdanderson.org/main/TANRIC
  - Visual exploration of lncNRAs in cancer

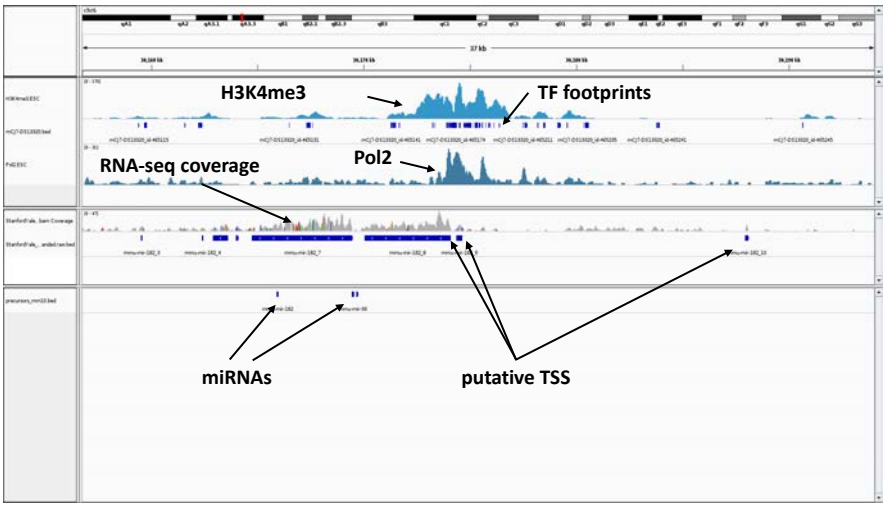
Finding miRNA/lncRNA Promoters

- LncRNA promoters share many characteristics of mRNA promoters
  - Histone Marks
  - DNase Footprints
  - Pol II Signal
- However, many are marked with H3K4me1 vs H3K4me3
- miRNAs are a different story
  - ~½ are derived from intronic regions (with or without TSS)
  - ~ ½ of miRNAs are transcribed by Pol II as Pri-miRs
    - Transcripts can be long (>10kb)
    - TSS can be far from the mature miRNA locus
    - Drosha cleaves the Pri-miR at transcription
- miRNA and lncRNA TSSs can be identified:
  - Drosha KO* (for miRNAs) + RNA-Seq
  - Ultra Deep RNA-Seq + Epigenetic Marks
  - Epigenetic Marks (are not usually sufficient)
  - CAGE (Capped Analysis of RNA Ends)
  - GRO-Seq (Global run-on sequencing)


RNA-Seq Coverage
Normalized read count (RPKM)
4
2
0
2
0
4
2
0
Drosha -/- mESCs
COIN
Drosha +/+ mESCs
8,856 bp
Mir17
Mir20a
Mir92-1
Mir17hg
Mir18
Mir19a
Mir19b-1
Georgakilas G, Vlachos IS, Paraskevopoulou MD, Yang P, Zhang Y, Economides AN, Hatzigeorgiou AG (2014), Nat Commun 5: 57

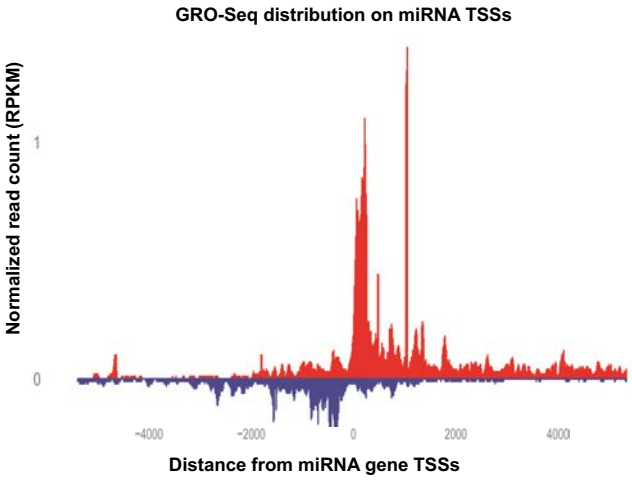

Normalized read count (RPKM)
4
2
0
2
0
4
2
0
GSM973235 WT mESCs
180M reads
Drosha -/- mESCs
Drosha +/+ mESCs
8,856 bp
Mir17
Mir20a
Mir92-1
Mir17hg
Mir18
Mir19a
Mir19b-1
Georgakilas G, Vlachos IS, Paraskevopoulou MD, Yang P, Zhang Y, Economides AN, Hatzigeorgiou AG (2014), Nat Commun 5: 5700

Combining Expression with ChIP-Seq data



Georgakilas G, Vlachos IS, Paraskevopoulou MD, Yang P, Zhang Y, Economides AN, Hatzigeorgiou AG **(2014), Nat Commun** 5: 5700

GRO-Seq



Georgakilas G, Vlachos IS, Paraskevopoulou MD, Yang P, Zhang Y, Economides AN, Hatzigeorgiou AG (2014), Nat Commun 5: 5700

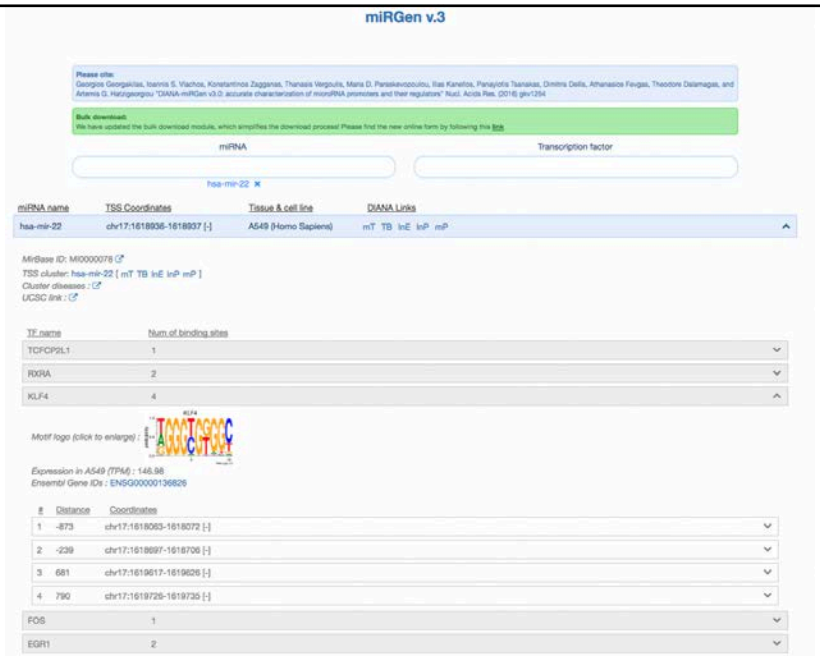
microTSS Validation

Precision and Sensitivity at 1kbp distance threshold from validated TSSs in mES, hES and IMR90 cells						
	mESCs [COIN] (N=47)		hESCs GRO-Seq (N=72)		IMR90 GRO-Seq (N=81)	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
Marson et al	54	64.5	15.2	40.7	18.5	29.4
PROmiRNA	78.7	25.4	83.3	41.7	85.1	33.3
S-Peaker	76.5	18.8	59.7	22.9	71.6	13.5
microTSS	**93.6**	**100**	**94.4**	**97.1**	**91.3**	**91.3**
miRStart	0	0	5.5	13.7	4.9	10.8

Transcription Factors Controlling miRNA Expression

- **miRGen v3.0** (www.microrna.gr/miRGen)
  - Application of microTSS on diverse human/mouse datasets
  - >19M TF binding sites on miRNA promoters
  - Data for 428 human miRNA precursors
- **TransmiR v2.0** (http://www.cuilab.cn/transmir)
  - 3,730 TF:miRNA interactions from the literature
  - 19 species
  - 1.7M TF:miRNA regulations

www.NoncodingRNA.org www.non-coding.org



Georgakilas G*, Vlachos IS*, Zagganas K, Vergoulis T, Paraskevopoulou MD, Kanellos I, Tsanakas P, Dellis D, Fevgas A, Dalamagas T, Hatizgeorgiou AG (2015), DIANA-miRGen v3.0: accurate characterization of microRNA promoters and their regulators, **Nucleic Acids Res** (44, D190-195).

www.NoncodingRNA.org www.non-coding.org

Non-coding RNA Variation

- **Transcription Regulation**
  - Promoters
  - Enhancers
- **Interacting domains**
  - miRNA seed regions
  - 3'UTRs
  - lncRNAs?
- **Other**
  - Secondary structure
  - Localization
  - Stability
  - …
- **CNVs**

Methods - miRNAs

- Most of these analyses are preferably performed in-house
  - QTLs
  - Binding prediction
  - Mediation analysis
  - Co-localization with regulatory SNPs / GWAS hits
  - Network biology
- There are published resources:
  - **PolymiRTS** [SNP effect - predictions] (http://bioinfo.bjmu.edu.cn/mirsnp/search/)
  - **ImiRP** [illegitimate site creation] (https://imirp.org)
  - **MSDD** [SNP effect - literature] (http://bioinfo.bjmu.edu.cn/mirsnp/search/)
    - 197 SNPs
  - **miRdSNP** [SNP effect - literature] (http://mirdsnp.ccr.buffalo.edu/)
    - 630 SNPs
  - **ADMIRE** [miRNA variation] (https://github.com/nroak/ADmiRE)
  - miRNASNP (https://www.bioguo.org/miRNASNP/)

Resources - lncRNAs

- **GTEx**
  - GENCODE transcript eQTLs
  - https://www.gtexportal.org
- **ncRNA-eQTL**
  - ncRNA reg eQTLs
  - http://ibi.hzau.edu.cn/ncRNA-eQTL/
- **LncRNASNP2**
  - lncRNA – SNPs
  - http://bioinfo.life.hust.edu.cn/lncRNASNP
- **LnCeVar**
  - CeRNA – SNP
  - http://www.bio-bigdata.net/LnCeVar/
- **LincSNP**
  - ceRNA/lncRNA/TF – SNPs
  - http://bio-bigdata.hrbmu.edu.cn/lincsnp/

www.NoncodingRNA.org www.non-coding.org

www.NoncodingRNA.org www.non-coding.org

miRNAs: Integration of NGS datasets

- Integrating RNA & miRNA expression can:
  - Identify functional interactions
  - Increase signal/noise ratio
  - Permit meaningful regulatory analyses and network biology
  - Reduce the search space
  - Prioritize central regulators
- Can be performed by:
  - Enrichment (GSEA, Fisher's, Empirical)
  - Network Statistics

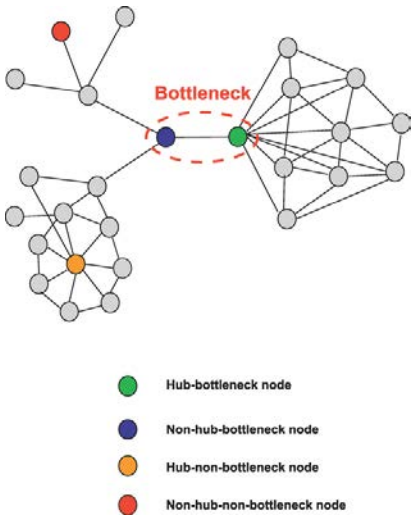
How To:

- **Concordant**
  - RNA-Seq
  - Small RNA-Seq (or microRNA expression assay)
- **miRNA Targets**
  - *In Silico*
    - TargetScan
    - microT-CDS
  - Experimentally Supported
    - TarBase
    - miRTarBase
- **TF:RNA Regulation**
  - FANTOM 5 for TSSs
  - Epigenetic Roadmap
    - ChIP-Seq
    - DNase Footprints
  - ChIP-Seq databases (examine pipeline prior to use)
- **TF:miRNA Regulation**
  - miRGen
  - TransmiR
- **Data Curation (These sources are not ready-to use)**
- **Enrichment**
  - Fisher's exact test, Empirical, GSEA
- **Network Statistics**

miRNA:mRNA:TF Regulatory Networks

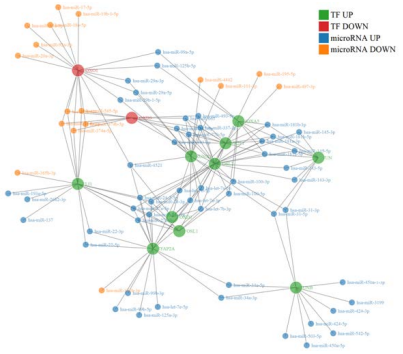
- **Nodes**
  - miRNAs
  - mRNAs
  - TFs
- **Edges**
  - Co-expression / Correlation
    - Pearson's CC
    - Spearman's Rho
  - De-Correlation / Deregulation
    - Loss of Significant Correlations
    - Significant (Z-score) change of association + filtering
  - Annotated Interactions (miRNA:mRNA, TF:miR, TF:mRNA)
- **Environment**
  - iGraph (R)
  - tidyGraph (R)
  - Cytoscape
- **Statistics**
  - Node Degree
  - Node Centrality
  - Hubs
  - H-index
  - Authorities
  - Bottlenecks
  - Module detection
  - Module detection + minimum cut

Figure 1. Schematic Showing a Bottleneck and the Four Categories of Nodes in a Network



Yu H, *et al*, PLOS Computational Biology 3(4): e59.

mirExTra 2

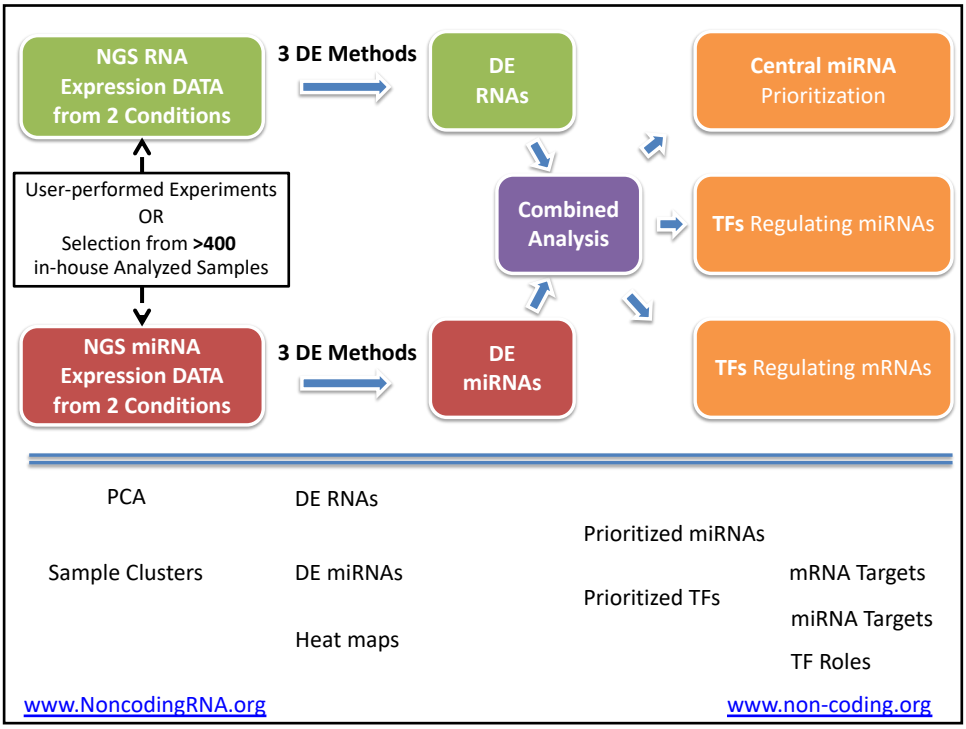


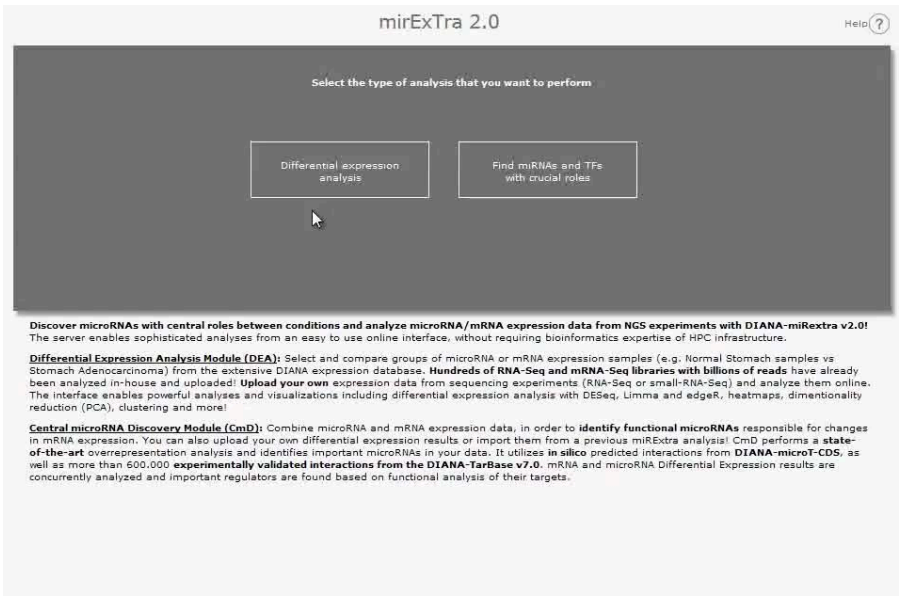
www.microrna.gr/mirextra

Integrated Data:

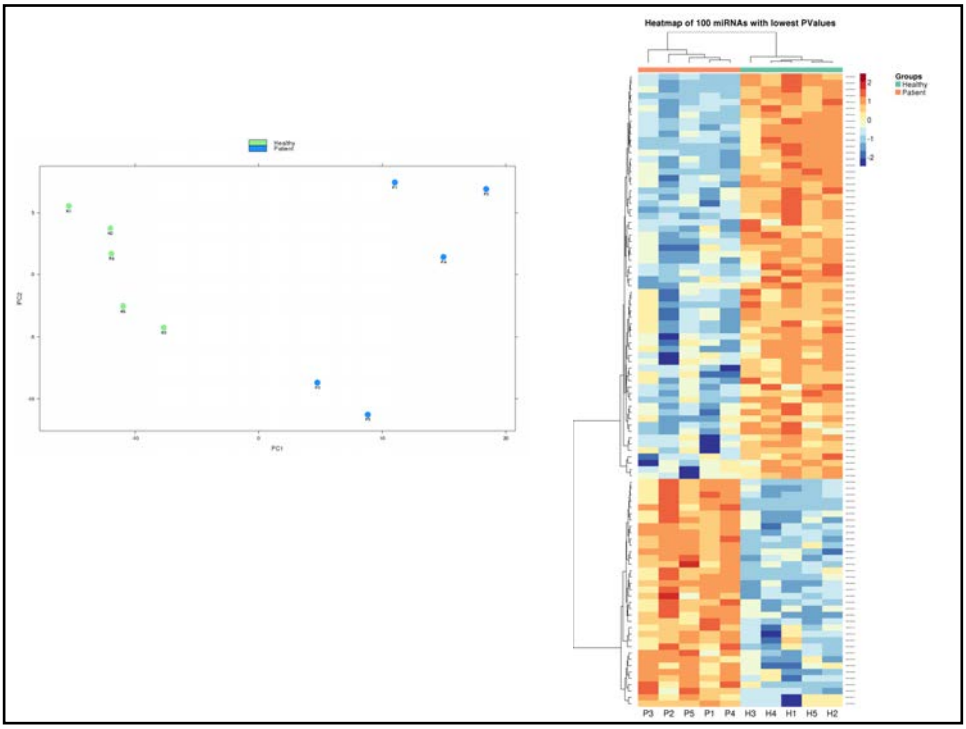
- miRNA:mRNA Interactions (TarBase v7.0, microT)
- miRNA TSSs (microTSS)
- TF:miRNA / TF:mRNA (ChIP-Seq, Oreganno 3, DNase-Seq)
- 350 small RNA-Seq / 65 RNA-Seq libraries
  - 70 tissues
  - 100 cell types
  - 90 conditions

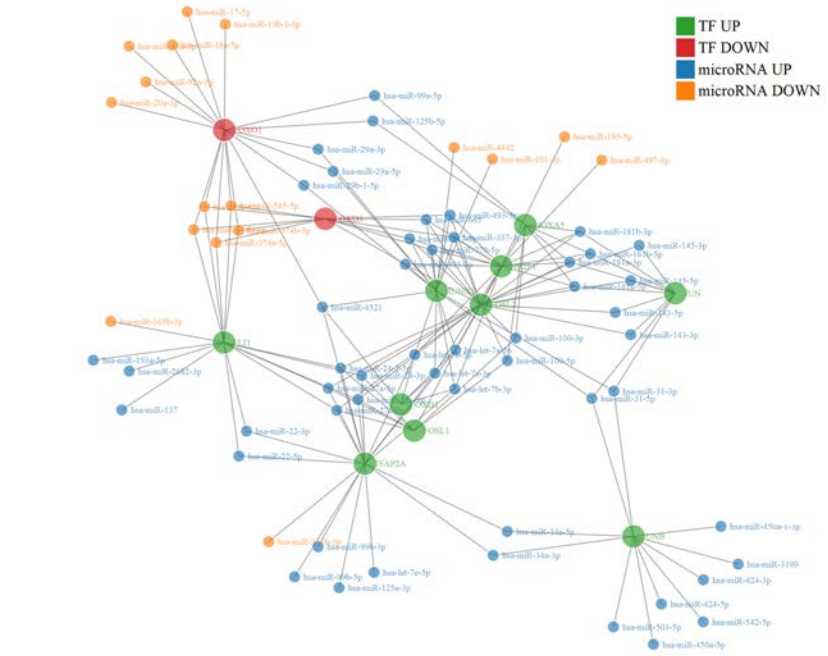
Vlachos IS, *et al*. (2016), DIANA-mirExTra v2.0: Uncovering microRNAs and transcription factors with crucial roles in NGS expression data. Nucleic Acids Research 44: W128-34.





www.NoncodingRNA.org www.non-coding.org

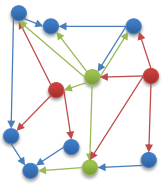




www.NoncodingRNA.org www.non-coding.org

LncRNA Networks

- Transcription Factors
- microRNAs
- Protein Coding Genes
- lncRNAs ??



- LncRNAs (still) can be added as part of ceRNA networks due to lack of direct target annotation
- LncRNAs can be incorporated in networks for lncRNA functional investigation

www.NoncodingRNA.org www.non-coding.org

Open Discussion

www.NoncodingRNA.org www.non-coding.org

www.NonCodingRNA.org

Clear RNA Focus & Expertise

Cutting-edge Pre-Release Methods

A-Z Project Support

Production – Grade Methods

Small to Population Size Studies

Computational Infrastructure