Small RNAs and how to analyze them using sequencing

Johan Reimegård

Enabler for Life Sciences

Karolinska Institutet
KTH Royal Institute of Technology
Stockholm University
Uppsala Universitet
SciLifeLab

Small RNAs

- Small RNAs are species of short non-coding RNAs, typically <100 nucleotides
  - micro RNAs (miRNAs)
  - short interfering RNAs (siRNAs)
  - piwi associated RNAs (piRNAs)
  - mirtrons, cis-natRNAs, TSS-miRNAs and other strange things
  - sRNAs

1. Background on regulatory small RNAs in eukaryotes

1993: The first microRNA is discovered in the worm genome

Cell, Vol. 75, 843–854, December 3, 1993, Copyright ©1993 by Cell Press

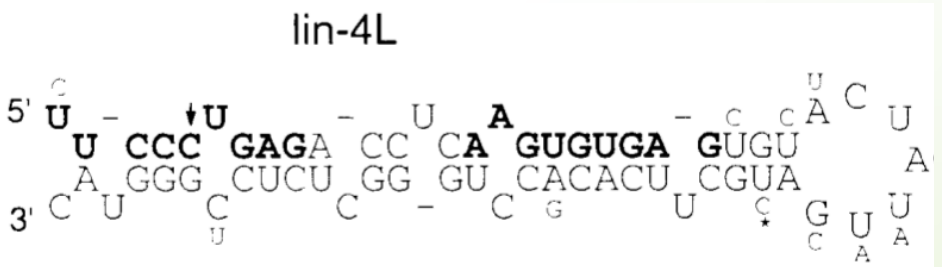
The C. elegans Heterochronic Gene *lin-4* Encodes Small RNAs with Antisense Complementarity to *lin-14*

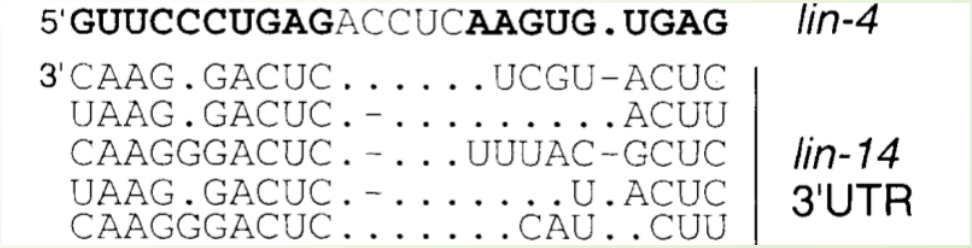
Rosalind C. Lee,*† Rhonda L. Feinbaum,*‡ and Victor Ambros†

1. A mutation in the lin-4 locus disrupts worm development.

2. The lin-4 locus encodes a non-coding RNA that forms a hairpin structure and produces two small transcripts, 61 and 22 nt.

3. Part of this RNA is complementary to the 3'UTR of a developmental gene, lin-14





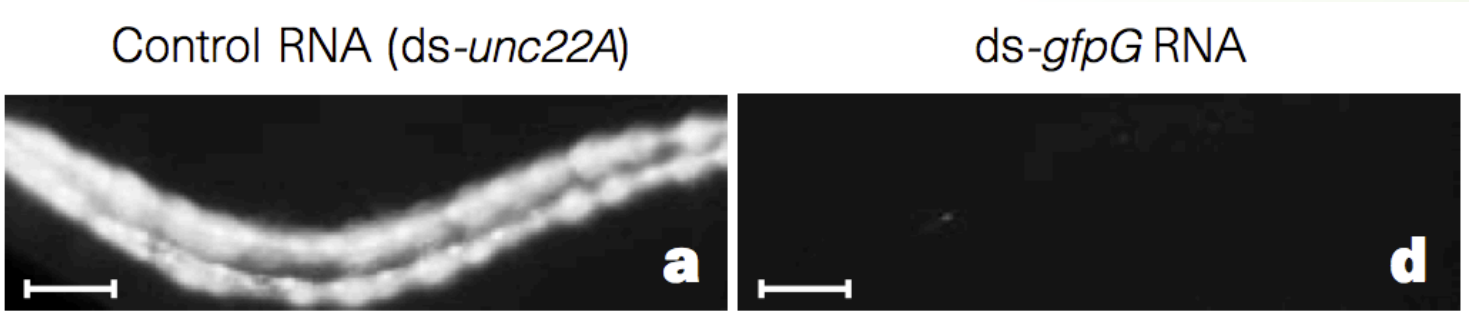
Karolinska Institutet · KTH Royal Institute of Technology · Stockholm University · Uppsala Universitet · SciLifeLab

1998: double stranded RNA can efficiently repress gene expression

Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*

Andrew Fire*, SiQun Xu*, Mary K. Montgomery*, Steven A. Kostas*†, Samuel E. Driver‡ & Craig C. Mello‡

"To our surprise, we found that double-stranded RNA was substantially more effective at producing interference than was either strand individually."

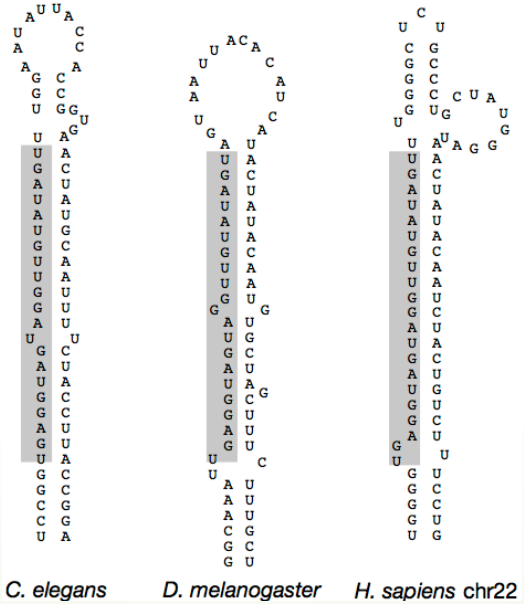


RNAi = RNA interference

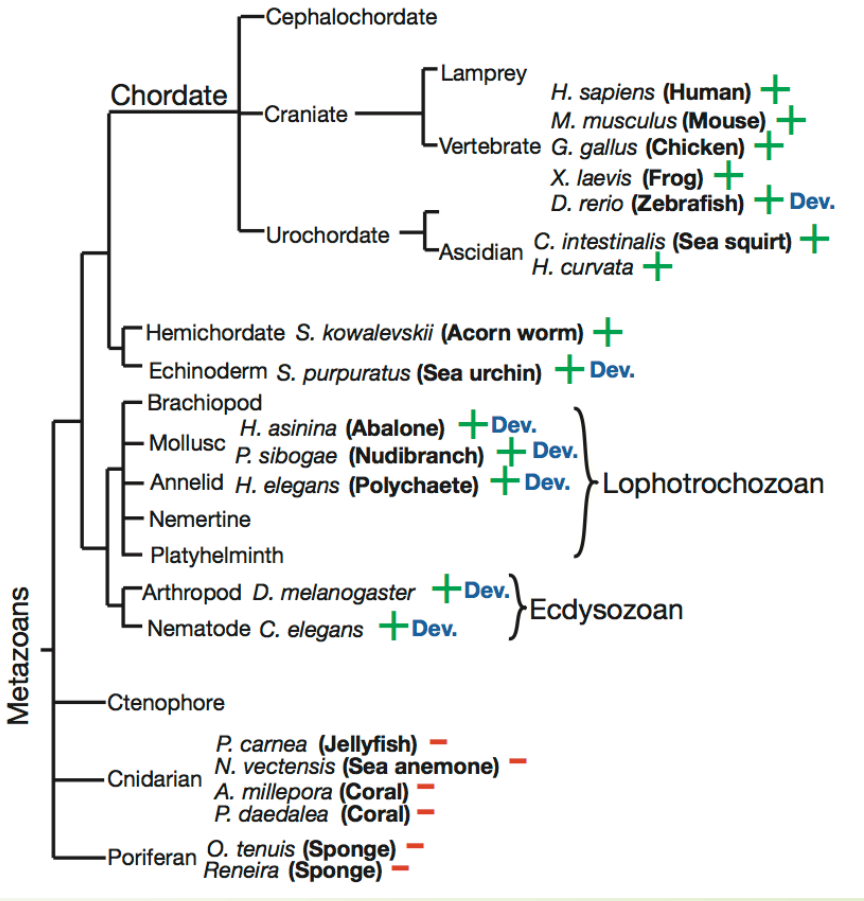
2000: a second, conserved, microRNA is found

Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA

Amy E. Pasquinelli*†, Brenda J. Reinhart*†, Frank Slack‡, Mark Q. Martindale§, Mitzi I. Kuroda‖, Betsy Maller‡, David C. Hayward¶, Eldon E. Ball¶, Bernard Degnan#, Peter Müller☆, Jürg Spring☆, Ashok Srinivasan, Mark Fishman**, John Finnerty††, Joseph Corbo‡‡, Michael Levine‡‡, Patrick Leahy§§, Eric Davidson§§ & Gary Ruvkun***



C. elegans *D. melanogaster* *H. sapiens* chr22



2001: many microRNAs are found in various animals

An Extensive Class of Small RNAs in *Caenorhabditis elegans*

Rosalind C. Lee and Victor Ambros*

An Abundant Class of Tiny RNAs with Probable Regulatory Roles in *Caenorhabditis elegans*

Nelson C. Lau, Lee P. Lim, Earl G. Weinstein, David P. Bartel*

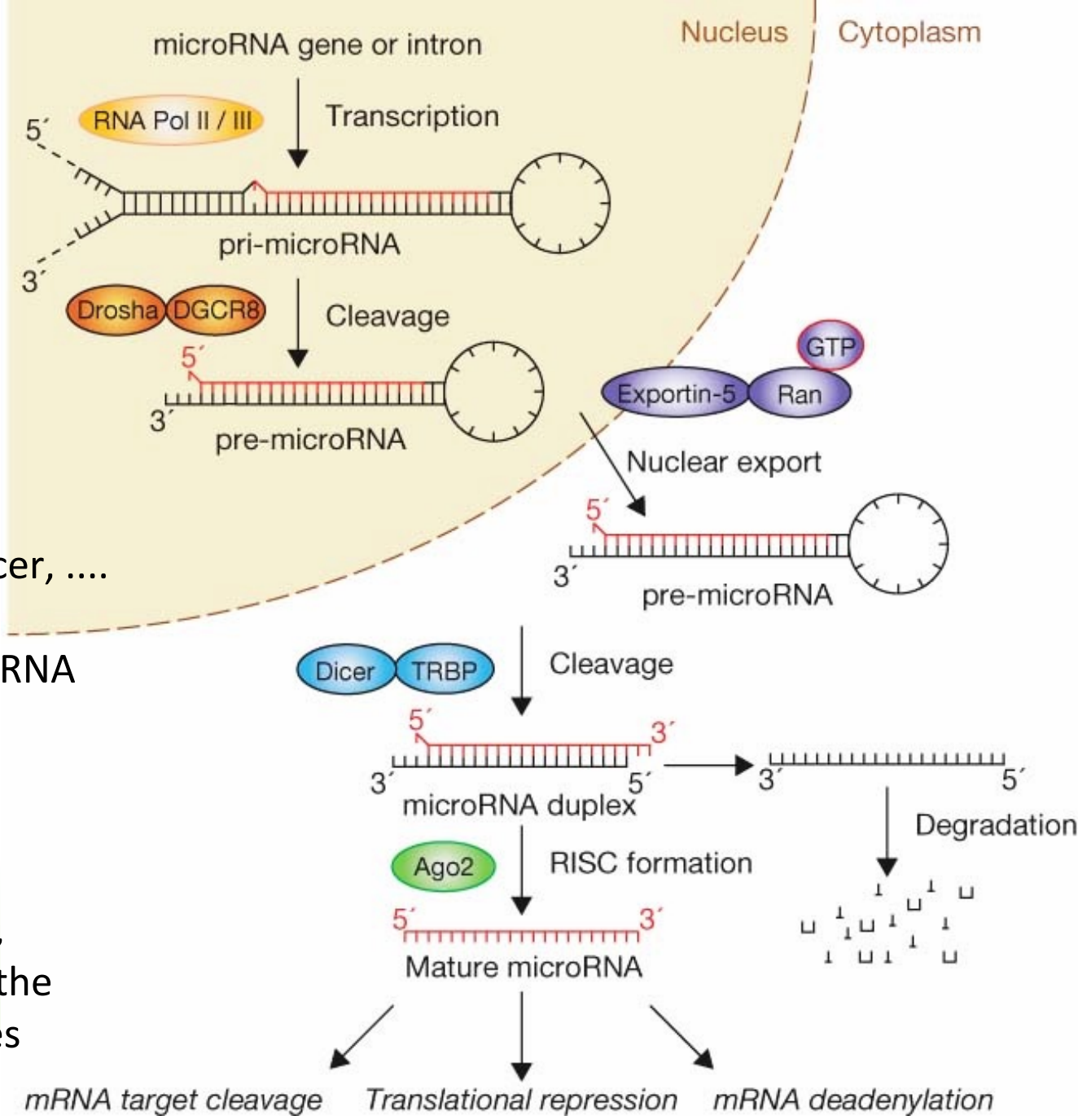
Using:

- (low throughput) sequencing
- RNA structure prediction
- Comparative genomics

Identification of Novel Genes Coding for Small Expressed RNAs

Mariana Lagos-Quintana, Reinhard Rauhut, Winfried Lendeckel, Thomas Tuschl*

microRNA biogenesis



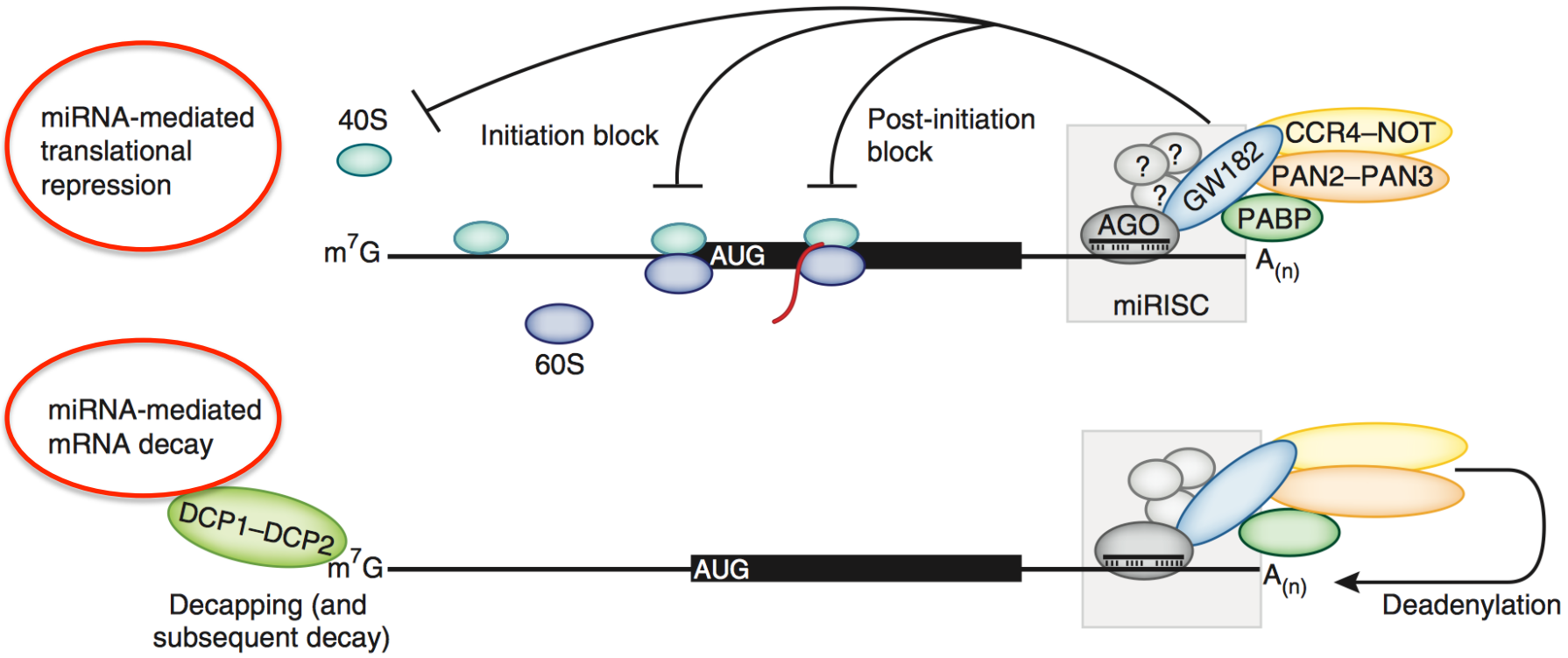
- Many enzymes etc. are involved: Drosha, Exp5, Dicer, ...

- The end result is a ~22nt RNA loaded into an Argonaute complex.

- The microRNA directs Argonaute to target genes, through base pairing with the 3'UTR (pos 2-8). This causes repression.

(Winter et. Nature Cell Biol, 2009)

Target repression by microRNAs

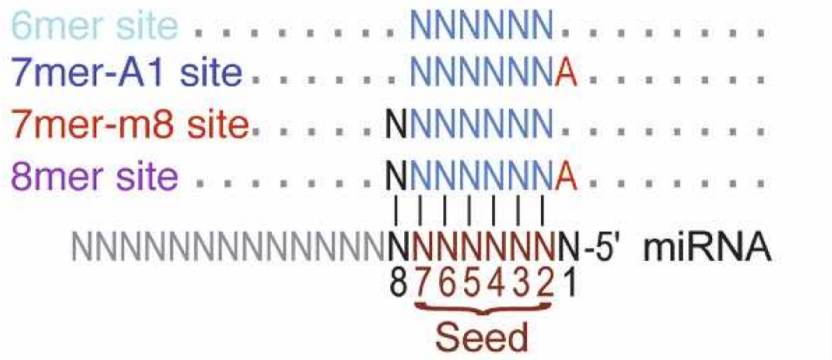


(This is in animals. microRNAs in plants work differently.)

(Fabian, NSMB, 2012)

How do microRNAs find their targets?

- In animals, microRNAs find their targets through pairing between the microRNA seed region (nucleotides 2-8) and the target transcript



(Friedman et al. Genome Research, 2009)

- Such short matches are common → a microRNA can have hundreds of targets.
- It is estimated that over half of all genes are targeted by microRNAs.

MicroRNA target prediction

- Besides seed pairing, other features are used in the target predictions:
  - Conservation (conserved target sites are more likely to be functional)
  - mRNA structure (it's hard for a microRNA to interact with a highly structured target mRNA)
  - Sequences around the target site (AU rich sequences around targets?)

- Many programs exist for microRNA target prediction (targetScan, mirSVR, PicTar, ..)

- These are not perfect. Target prediction is hard, and a lot of details about the mechanism are still not known.

MicroRNAs in animal genomes

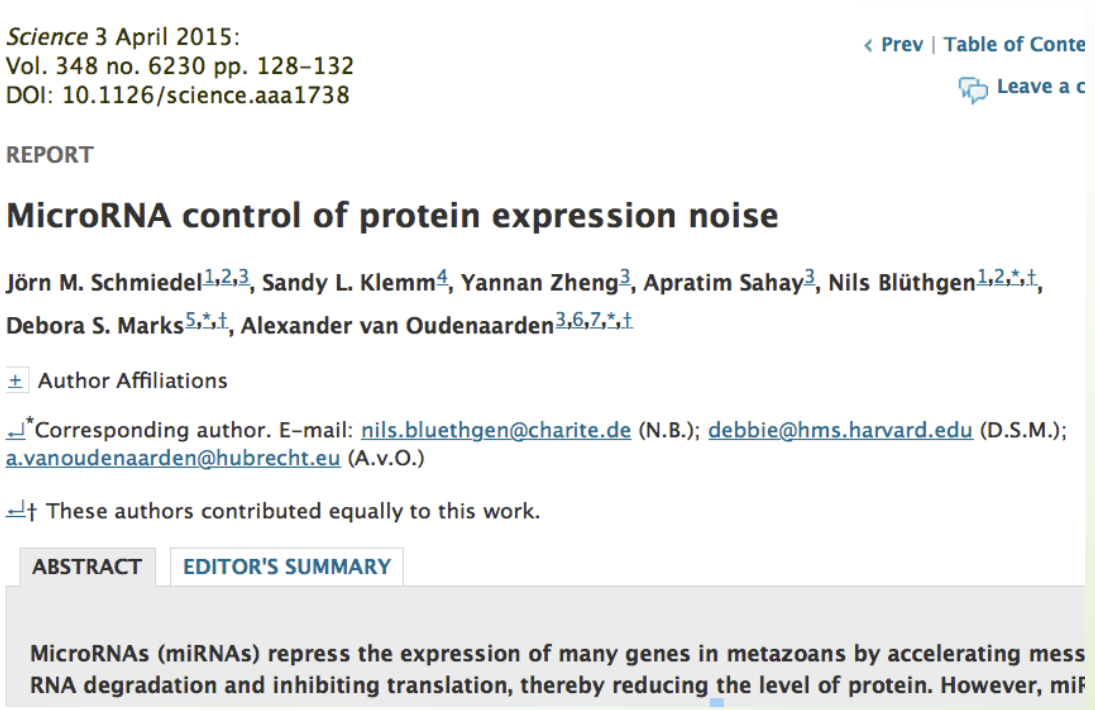
- There are typically hundreds or thousands microRNAs in animal genomes:
  - Fly: ~300 microRNA loci
  - Mouse: ~1200 microRNA loci
  - Human: ~1900 microRNA loci

- A single locus can produce multiple microRNA forms (called isomirs).

- In a given tissue, their expression can range over 6 orders of magnitude (a few to a few million reads)

microRNAs regulate many biological processes and are involved in disease

- Development
- Stress response
- Cancer
- Cardiovascular disease
- Inflammatory disease
- Autoimmune disease



Science 3 April 2015:
Vol. 348 no. 6230 pp. 128–132
DOI: 10.1126/science.aaa1738

REPORT

MicroRNA control of protein expression noise

Jörn M. Schmiedel[1,2,3], Sandy L. Klemm[4], Yannan Zheng[3], Apratim Sahay[3], Nils Blüthgen[1,2,*,†], Debora S. Marks[5,*,†], Alexander van Oudenaarden[3,6,7,*,†]

± Author Affiliations

*Corresponding author. E-mail: nils.bluethgen@charite.de (N.B.); debbie@hms.harvard.edu (D.S.M.); a.vanoudenaarden@hubrecht.eu (A.v.O.)

† These authors contributed equally to this work.

ABSTRACT EDITOR'S SUMMARY

MicroRNAs (miRNAs) repress the expression of many genes in metazoans by accelerating mess RNA degradation and inhibiting translation, thereby reducing the level of protein. However, miR

Karolinska Institutet · KTH Royal Institute of Technology · Stockholm University · Uppsala Universitet · SciLifeLab

2. Small RNA sequencing

Sequencing

- Small RNA sequencing is similar to mRNA sequencing, but:
  - There is no poly-A selection. Instead RNA fragments are size selected (typically less than 35 nucleotides, to avoid contamination by ribosomal RNA).
  - Low complexity libraries → more sequencing problems
  - FastQC results will look strange:
    - Length
    - Nucleotide content
    - Sequence duplication

Pre-processing of small RNA data I

- Since we are sequencing short RNA fragments, adaptor sequences end up in the reads too.
- Many programs available to remove adaptor sequences (cutadapt, fastx_clipper, Btrim..)
- We only want to keep the reads that had adaptors in them.

GTTTCTGCATTT**TCGTATGCCGTCTTCTGCTTGAA**
GTGGGTAGAACTTTGATTAAT**TCGTATGCCGTCTT**
GTTTGTAAATTCTGA**TCGTATGCCGTCTTCTGCTT**
GAATATATATAGATATATACATACATACTTATCGT
GCTGACTTAGCTTGAAGCATAAATGG**TCGTATGCC**
GACGATCTAGACGGTTTTCGCAGAATTCTGTTTAT

Adapter missing

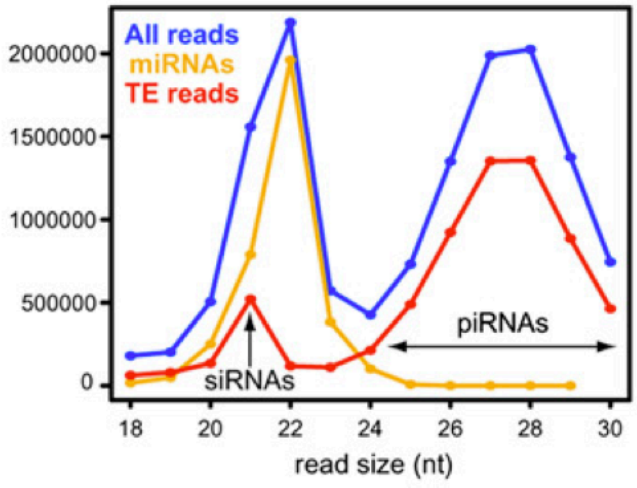
Pre-processing of small RNA data II

- microRNAs are expected to be 20-25 nt.
  - Short reads are probably not microRNAs, and are hard to map uniquely

GTTTCTGCATTT**TCGTATGCCGTCTTCTGCTTGAA**
GTGGGTAGAACTTTGATTAAT**TCGTATGCCGTCTT**
GTTTGTAAATTCTGA**TCGTATGCCGTCTTCTGCTT**
GCTGACTTAGCTTGAAGCATAAATGG**TCGTATGCC**

To short

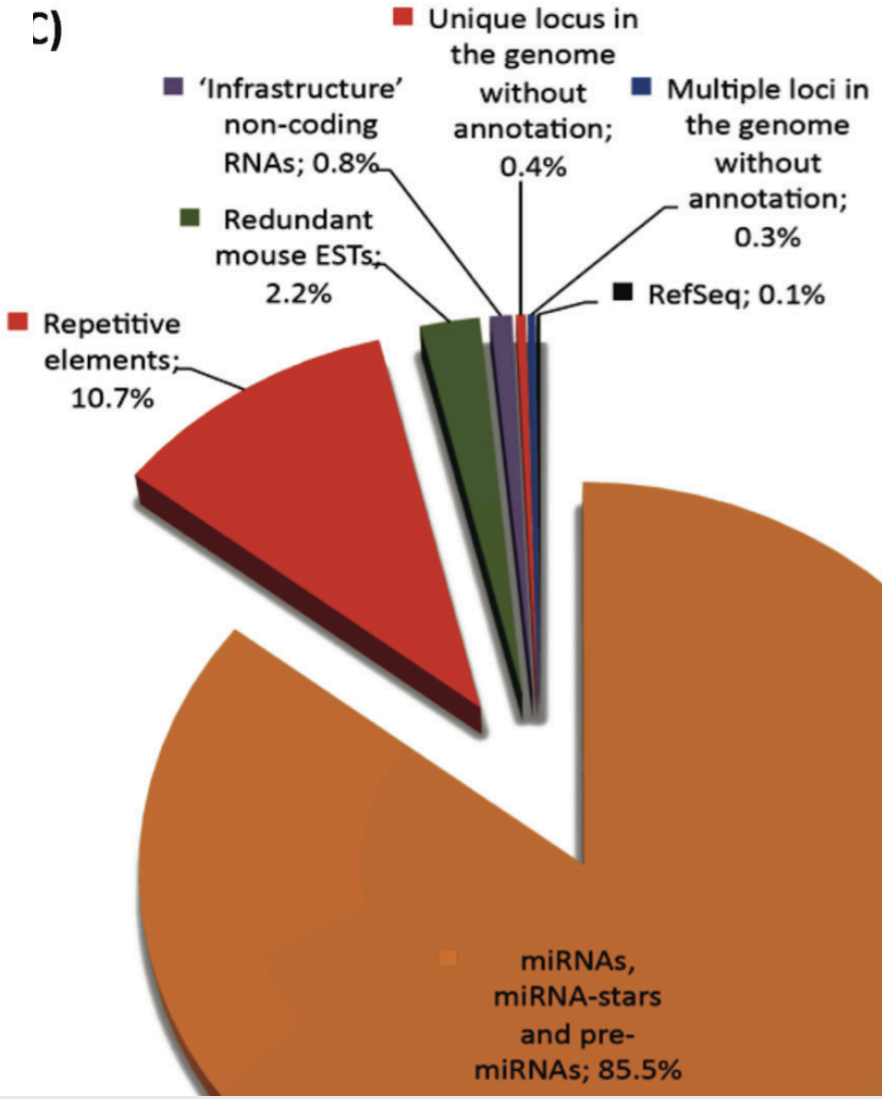
  - Long reads are probably not microRNAs



(Lau et al. Genome Research, 2010)

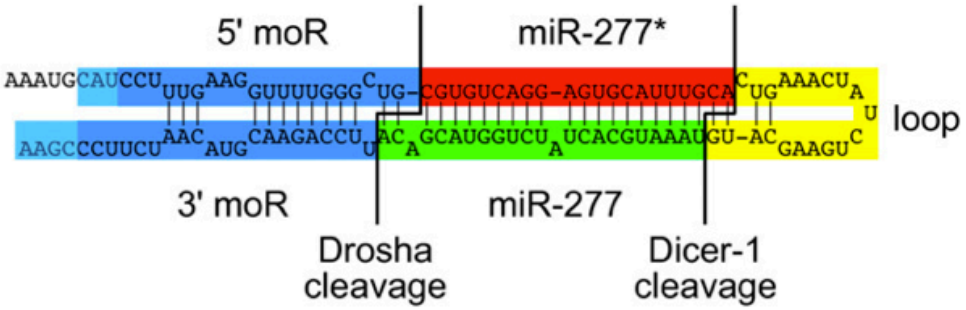
Pre-processing of small RNA data III

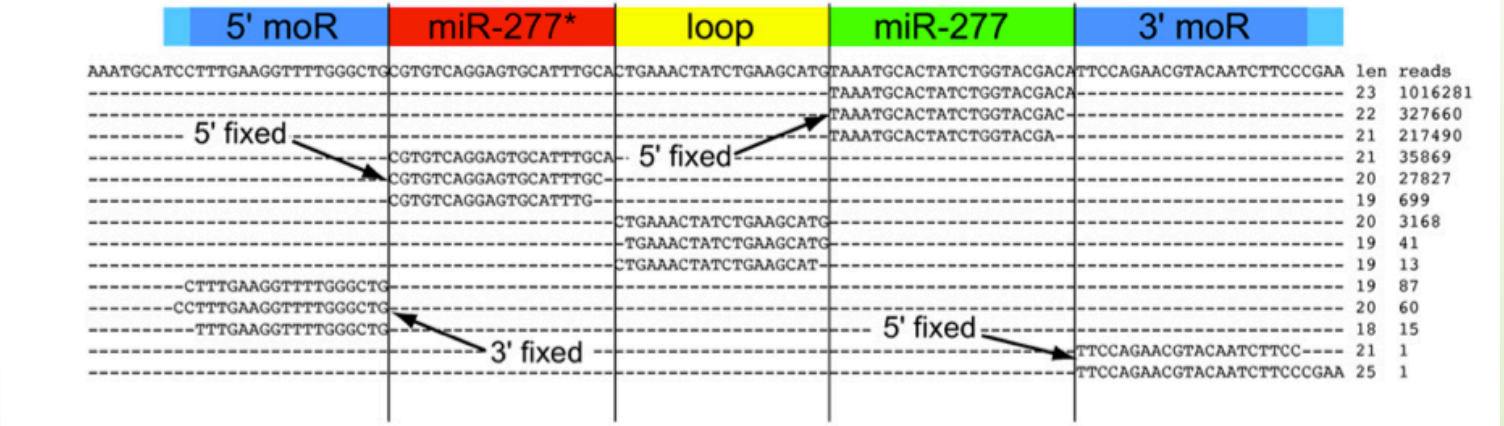
Another useful QC step is to check which loci the reads map to:



(Ling, BMC Genomics, 2011)

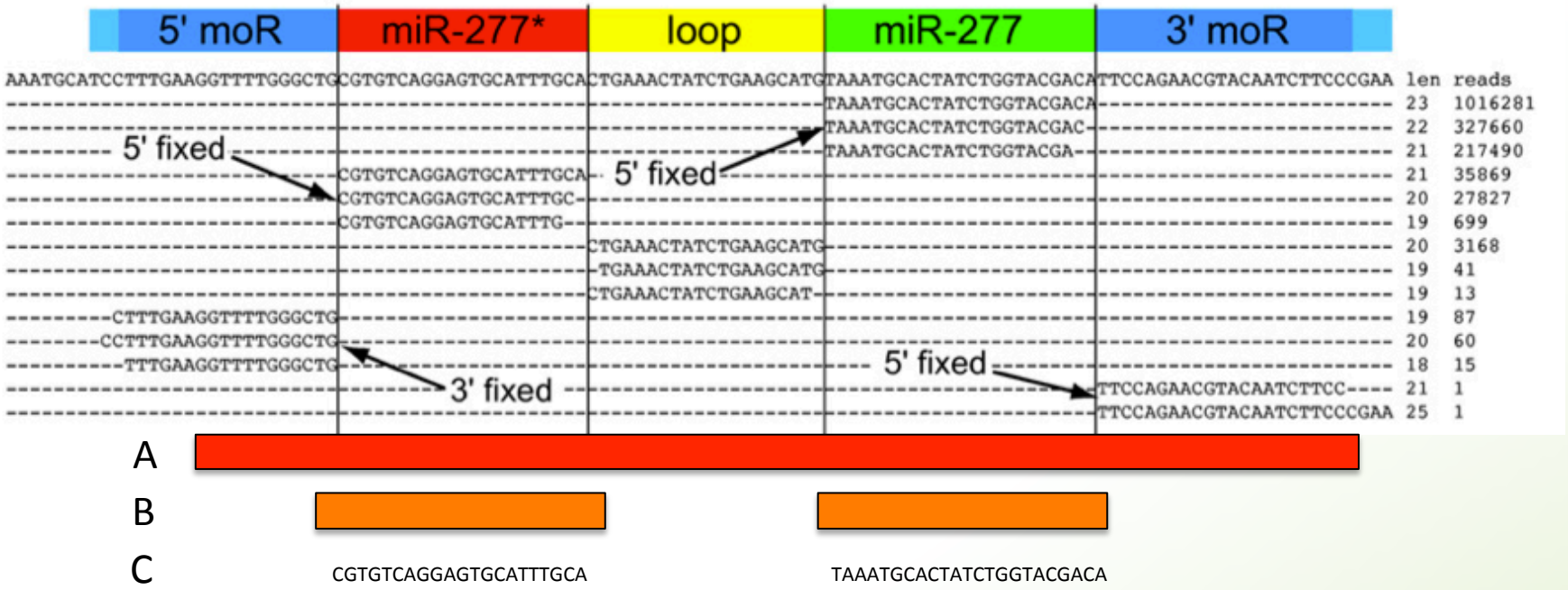
Example of reads mapping to a microRNA locus





(Berezikov et al. Genome Research, 2011.)

Quantifying small RNA expression I



A. Count all reads mapping to a locus? - The simplest option, usually good for profiling.

B. Count reads from each hairpin arm? - Useful if we want to correlate this with expression of target genes, or do more careful profiling.

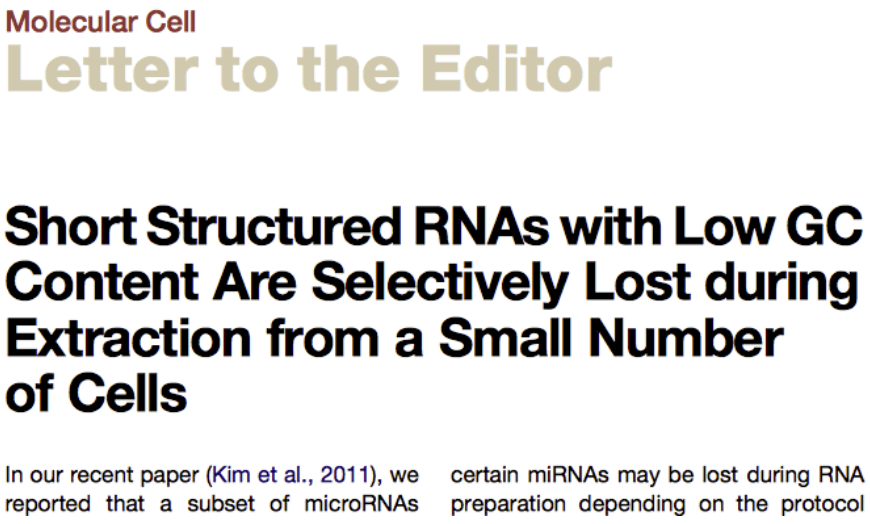
C. Only count reads that exactly match the mature microRNA? - Not a good idea, because we will miss variants

Quantifying small RNA expression II

- microRNAs from the same family can be very similar (or identical)
  - How treat this:
    - Keep in mind that some microRNAs are hard to separate.
    - If a read maps to several N loci, count 1/N read at each locus.
    - …

Error sources

- Different chemistries and protocols can have different effects on expression measurements.
  - We only get a few different sequences from each microRNA, so any biases can have big effects. (In normal RNA-seq each gene generates many different reads, so this is not a big problem)
  - Normalization doesn't fix these problems → it's hard to compare data from different platforms etc.
- The amount of starting material can influence the results:



Molecular Cell

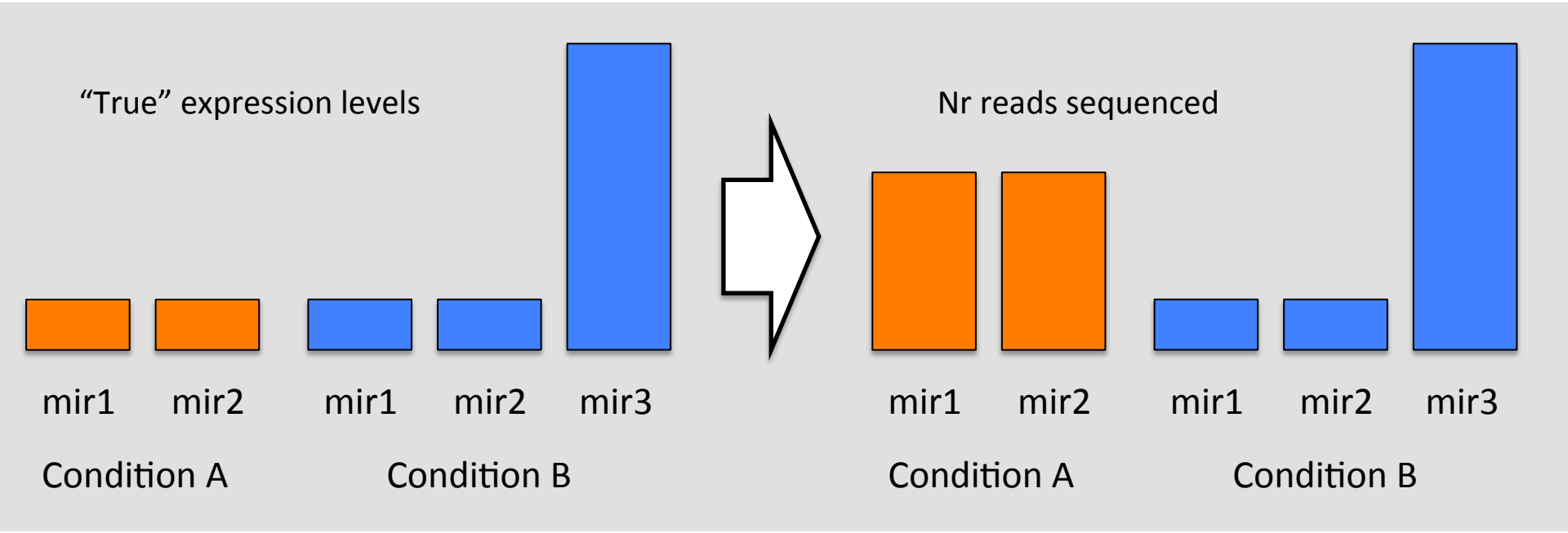
Letter to the Editor

Short Structured RNAs with Low GC Content Are Selectively Lost during Extraction from a Small Number of Cells

In our recent paper (Kim et al., 2011), we reported that a subset of microRNAs certain miRNAs may be lost during RNA preparation depending on the protocol

Normalizing small RNA expression levels I

- Only a few loci, and huge differences in expression levels → a few miRNAs can account for the majority of all reads, and skew expression levels of all microRNAs.



- Since many reads are used to sequence mir3 in condition B, fewer are available for mir1 and mir2.
- Normalization needs to deal with this situation. Simply scaling read counts by the total number mapped reads will not solve this problem.
- (Spike-in are always useful for normalization.)

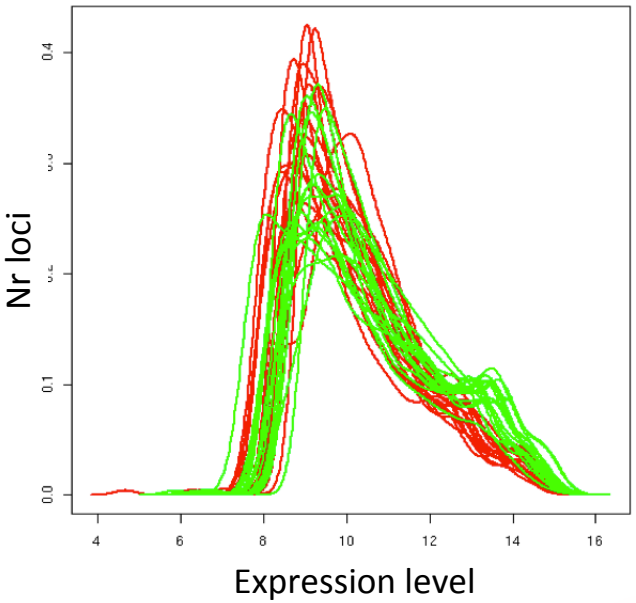
Normalizing small RNA expression levels II

- Different methods for normalization
  - TMM ("trimmed mean of M-values") normalization (Robinson et al. 2010, Genome Biology, McCormick et al. 2011, Silence)
  - In short, TMM normalization works like this:
    - Compute log ratios of all microRNA ("M-values")
    - Remove ("trim") the highest and the lowest log-ratios, and the highest and lowest expressed microRNAs.
    - Use a mean of the remaining log-ratios to compute the scaling factors
  - The underlying assumption is that most microRNAs have similar expression levels in the different samples, and should have similar expression levels after normalization.

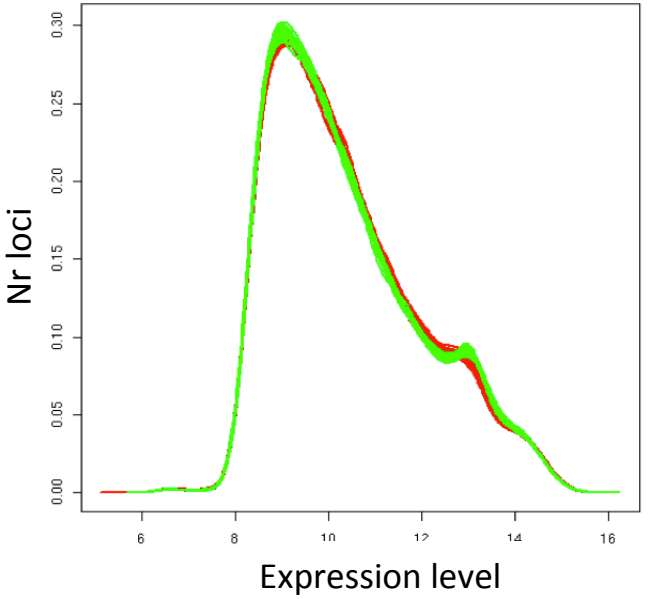
Normalizing small RNA expression levels III

- Quantile normalization (Garmire et al. RNA, 2012)
  - The underlying assumption is that the overall expression distribution is the same in all samples.

Before normalization



After normalization



- Many other methods exist, developed for RNA-seq or microarrays.
- No consensus about which method to use → always good to try a few different methods.

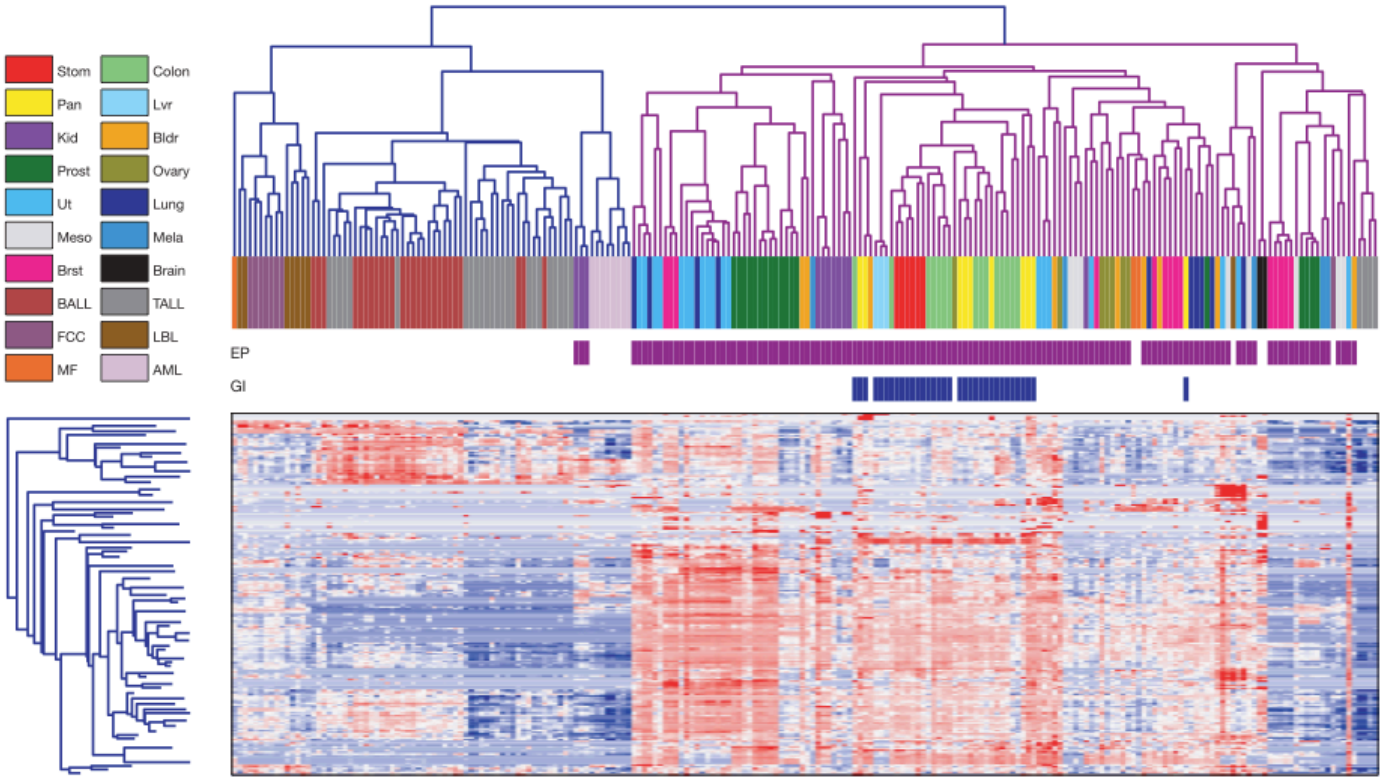
Differential expression

- After the expression levels have been properly normalized, methods for RNA-seq differential expression can be used.
  - ANOVA, t-tests
  - DeSeq, edgeR, voom, limma, etc..
  - No consensus on which method is best.

- Keep in mind: Since microRNA quantification is less reliable than normal RNA-seq:
  - More replicates are needed.
  - More validation experiments are required (Northern blots, in-situ hybridization, etc.).
  - Use caution when interpreting results!

3. What can we learn from microRNA expression analysis?

MicroRNA expression profiles classify human cancers



microRNA expression profiles cluster according to cancer type.

(Lu et al. Nature 2005)

microRNA profiles can be used to distinguish cancer subtypes

Table 1. Cancer subtypes that can be distinguished by microRNA or miRNA profiles

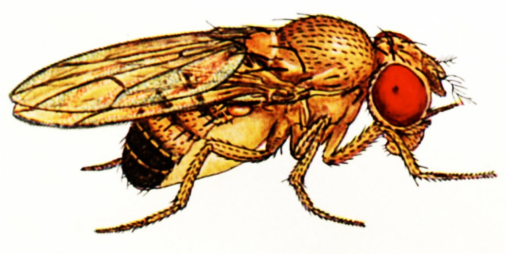
Cancer type	miRNAs[a]	Ref.
Breast		
ER status	miR-26a/b, miR-30 family, miR-29b, miR-155, miR-342, miR-206, miR-191	[38–40,42]
PR status	let-7c, miR-29b, miR-26a, miR-30 family, miR-520g	[41,42]
HER2/*neu* status	miR-520d, miR-181c, miR-302c, miR-376b, miR-30e	[38,41]
Lung		
Squamous vs non-squamous cell	miR-205	[33]
Small cell vs non-small cell	miR-17-5p, miR-22, miR-24, miR-31	[32]
Gastric		
Diffuse vs intestinal	miR-29b/c, miR-30 family, miR-135a/b	[35]
Endometrial		
Endometrioid vs uterine papillary	miR-19a/b, miR-30e-5p, miR-101, miR-452, miR-382, miR-15a, miR-29c	[37]
Renal		
Clear cell vs papillary	miR-424, miR-203, miR-31, miR-126	[34,36]
Oncocytoma vs chromophobe	miR-200c, miR-139-5p	[36]
Myeloma		
with t(14;16)	miR-1, miR-133a	[60]
with t(4;14)	miR-203, miR-155, miR-375	[60]
with t(11;14)	miR-125a, miR-650, miR-184	[60]
Acute myeloid leukemia		
with t(15;17)	miR-382, miR-134, miR-376a, miR-127, miR-299-5p, miR-323	[52]
with t(8;21) or inv(16)	let-7b/c, miR-127	[52]
with *NPM1*[b] mutations	miR-10a/b, let-7, miR-29, miR-204, miR-128a, miR-196a/b	[51,52]
with *FLT3* ITD	miR-155	[51,52,54]
Chronic lymphocytic leukemia		
ZAP-70 levels and IgVH status	miR-15a, miR-195, miR-221, miR-155, miR-23b	[50]
Melanoma		
with BRAF V600E	miR-193a, miR-338, miR-565	[56]

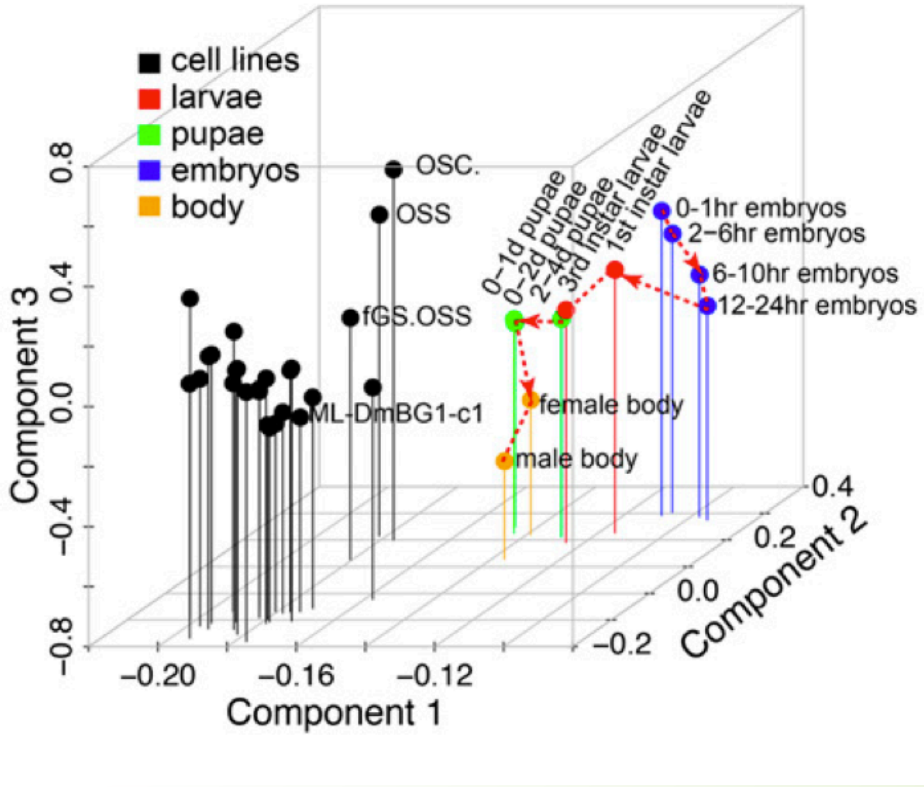
[a]Not all distinguishing miRNAs are represented in this table.

[b]nucleophosmin 1.

(Chan et al. Trends in Molecular Medicine, 2010)

microRNA profiles in cell lines vs tissues





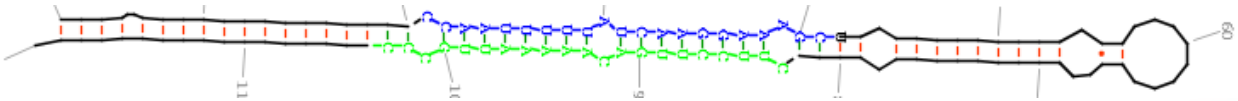
PCA plot showing that microRNA profiles in most cell lines are more similar to each other to normal tissues.

(Wen et al. Genome Research 2014)

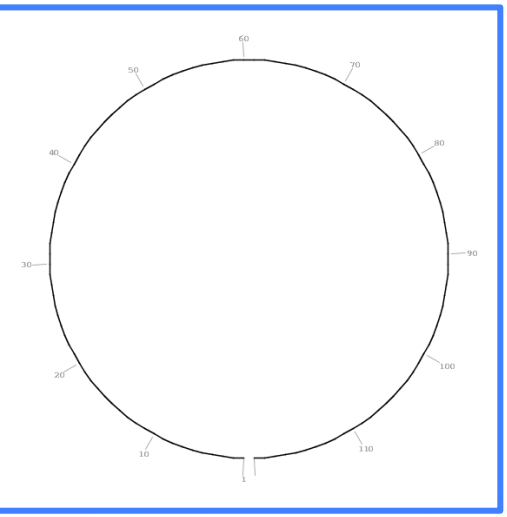
Discovering new small RNA loci

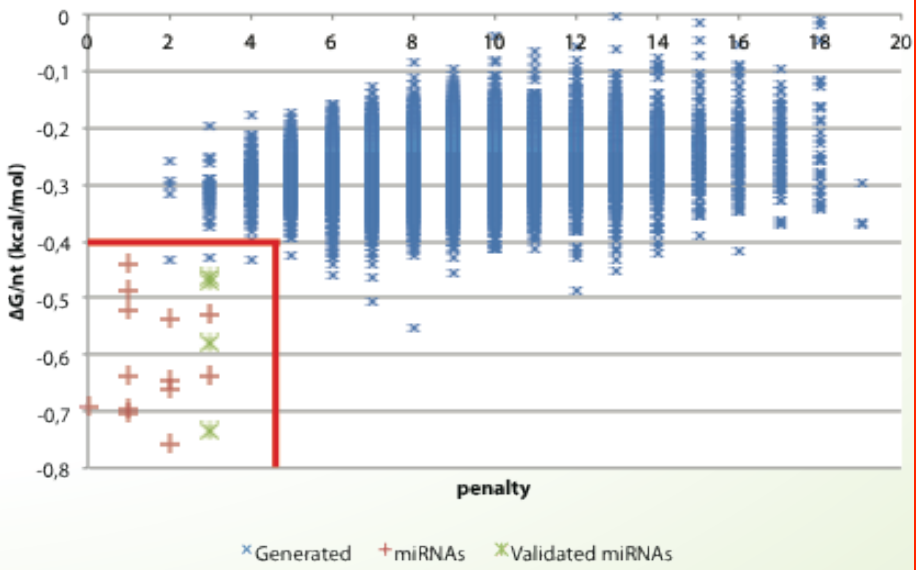
- microRNAs have very specific patterns, when it comes to
  - Read size and mapping
  - RNA structure
  - Conservation
- This makes it possible to find microRNAs using small RNA sequencing data.

miRNA prediction using known miRNA features

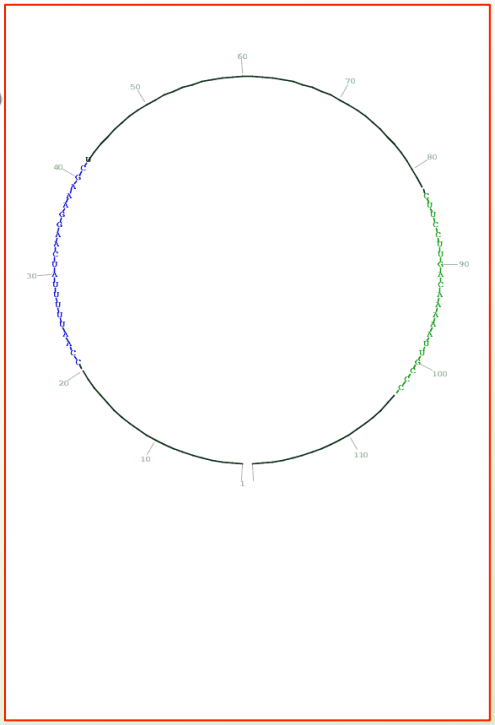


Generated
Scrambled



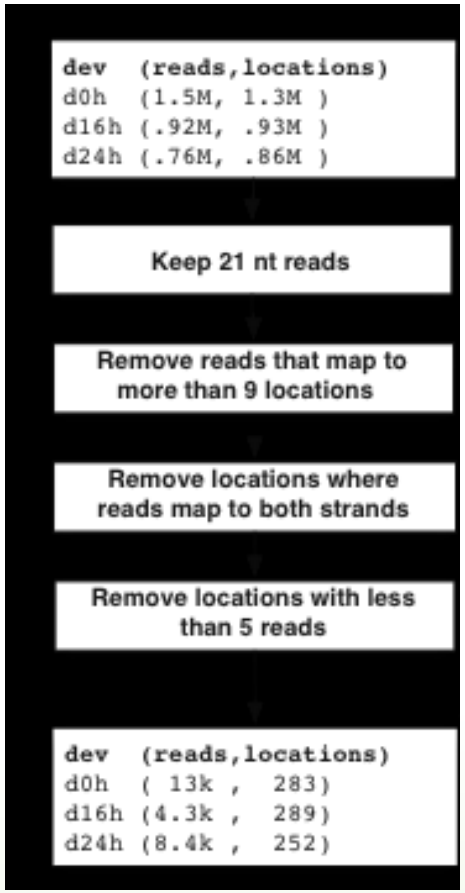


True



miRNAs in Dicty were found using RNAseq and miRNA pre

Expression criteria



pre-miRNA structure criteria

A

	d0h	d16h	d24h	total
Locations	283	289	252	567

Identify pre-miRNA structures using MiRfold

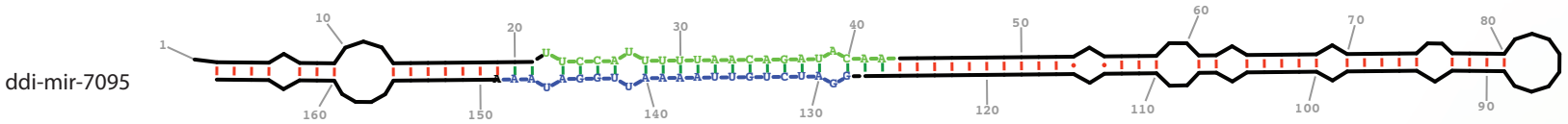
ΔG/nt < -0.4 kCal/mol
penalty < 5

Identify human-like pre-miRNA structures using miRabela

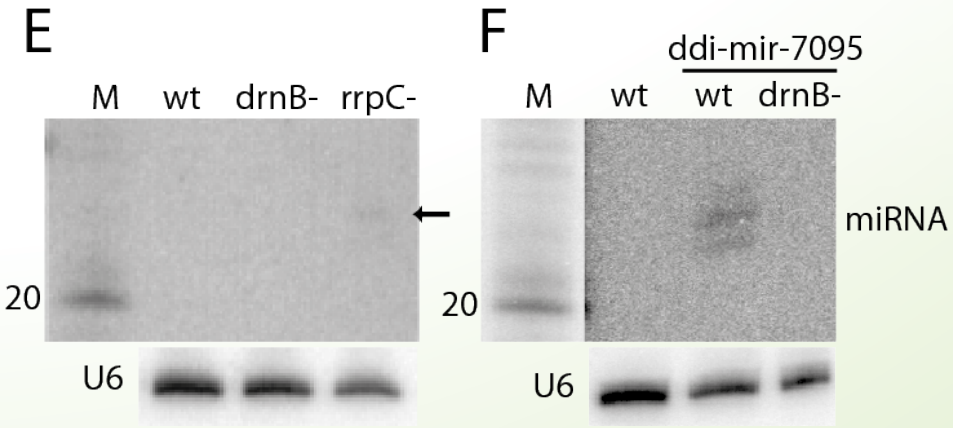
score > 0

	d0h	d16h	d24h	total
miRNAs	9	9	13	17
pre-miRNAs	10	15	20	45

ddi-mir-7095




```
>ddi-mir-7095 miRNAstart=128 miRNAstarStart=22
AAACUCAAAUAUUCAUUUUUUUUUCCAUUUUUAACAGAUACAAAACAAAAUAAUAGUAGAGGUUUAAUUAAACCUAAAAUUCAGAAUAAAAUUCAGGUUCAAUUAUACACCUAUAGUUAUUUUGUUUUGGAUCUGUUAAAAUUGGAUAAAAAAUGAUUUUUUUUAGUU
.....................UUUCCAUUUUUAACAGAUACA.....................................................................................................................  2
.....................UUCCAUUUUUAACAGA.........................................................................................................................  1
.....................UUCCAUUUUUAACAGAU........................................................................................................................  1
.....................UUCCAUUUUUAACAGAUA.......................................................................................................................  27
.....................UUCCAUUUUUAACAGAUAC......................................................................................................................  7
.....................UUCCAUUUUUAACAGAUACA.....................................................................................................................  119
.....................UUCCAUUUUUAACAGAUACAA....................................................................................................................  17
.....................UUCCAUUUUUAACAGAUACAAAA..................................................................................................................  1
.....................UCCAUUUUUAACAGAUACA.......................................................................................................................  2
.....................UCCAUUUUUAACAGAUACAAA.....................................................................................................................  1
.....................................................................................GGAUCUGUUAAAAUUGGAU.....................................  7
.....................................................................................GGAUCUGUUAAAAUUGGAUA....................................  67
.....................................................................................GGAUCUGUUAAAAUUGGAUAA...................................  247
.....................................................................................GGAUCUGUUAAAAUUGGAUAAA..................................  5
.....................................................................................GAUCUGUUAAAAUUGGAUAA....................................  2
```



Karolinska Institutet · KTH Royal Institute of Technology · Stockholm University · Uppsala Universitet · SciLifeLab

mirDeep(2)

- The most used program for finding new microRNAs. Takes as input:
  - A genome sequence
  - Small RNA sequencing data
  - A set of loci to exclude (optional)

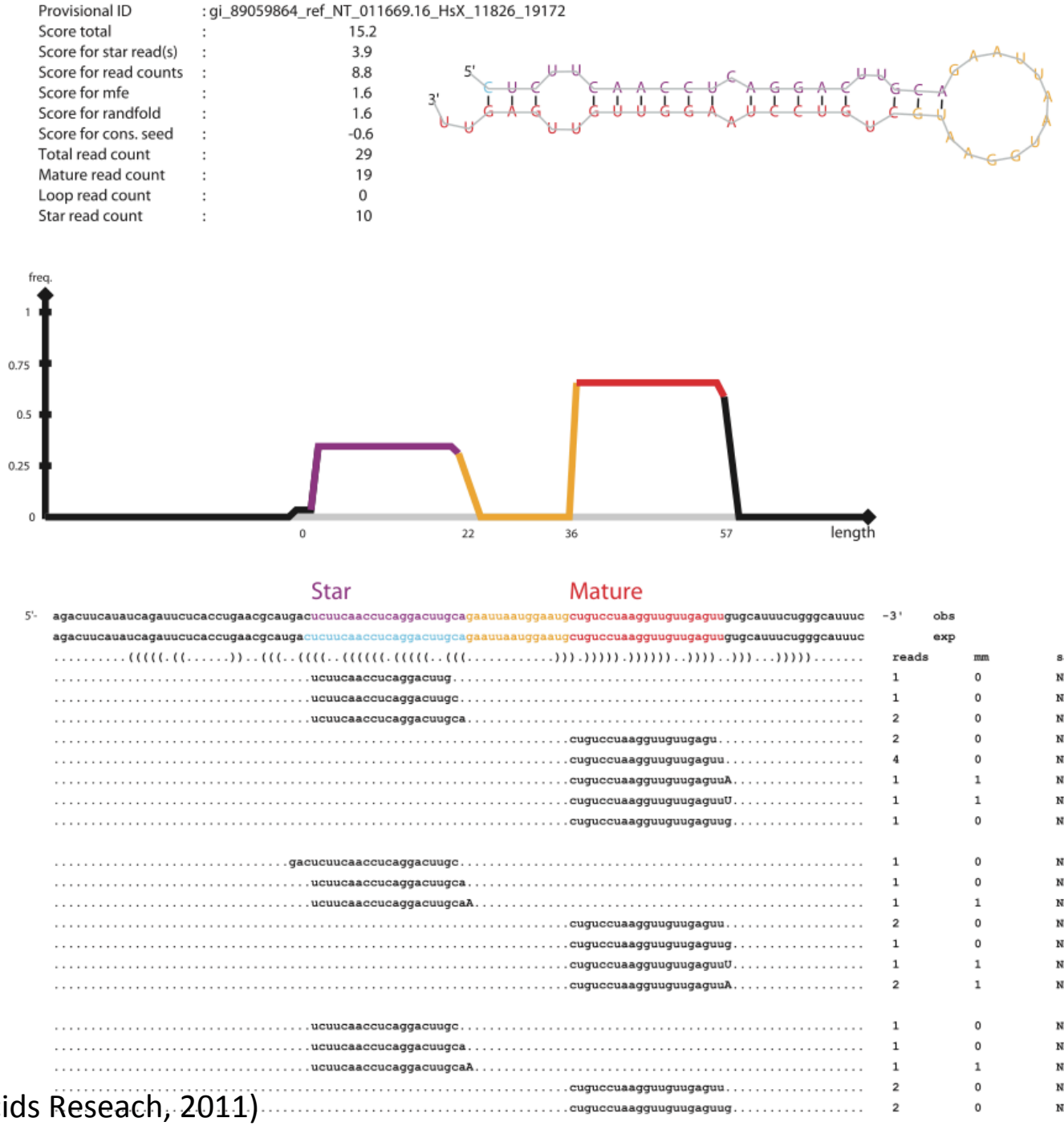
- The basic idea is:
  - Look at sequence data to find a (large) set of possible loci.
  - Look at RNA folding and read mapping patterns to give a score to each candidate
    - Nr reads from both arms
    - Fixed read ends
    - Free energy, base pairing in the hairpin structure, ..

- Output is a list of microRNA candidates, with scores, and a plot for each candidate:

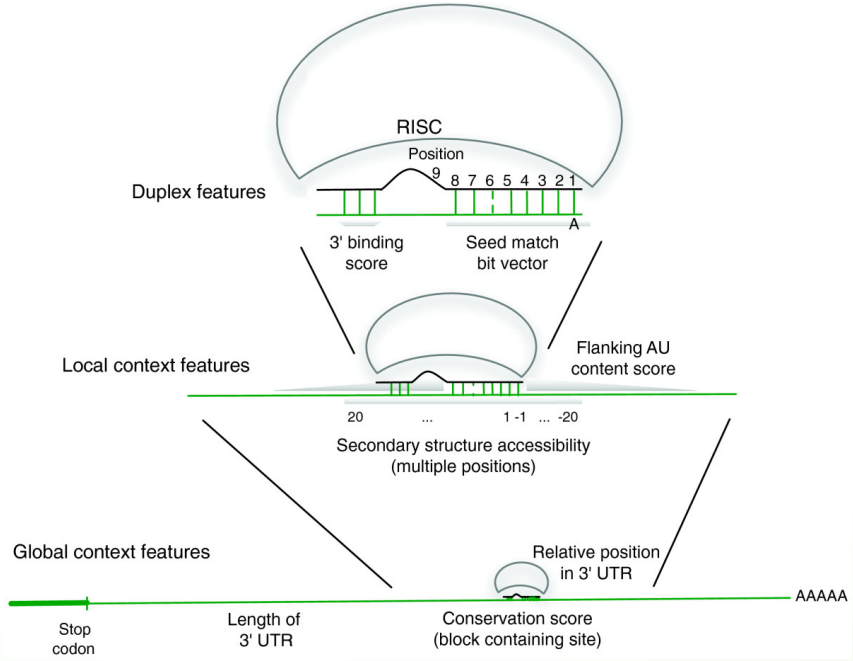
- mirDeep2 is installed on UPPMAX.

- There are also other programs, e.g. shortStack (Axtell, RNA, 2013) which also finds other small RNAs.

(Friedländer et. al. Nucleic Acids Reseach, 2011)

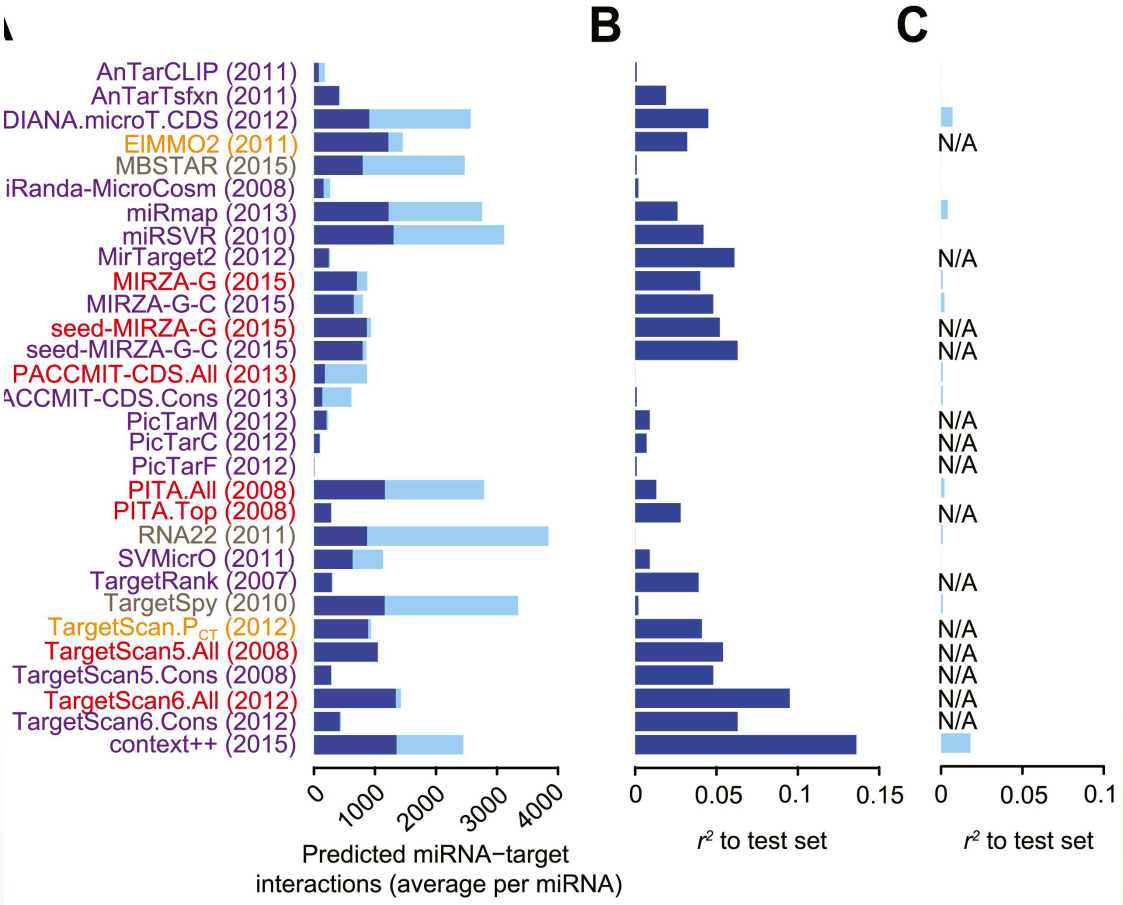


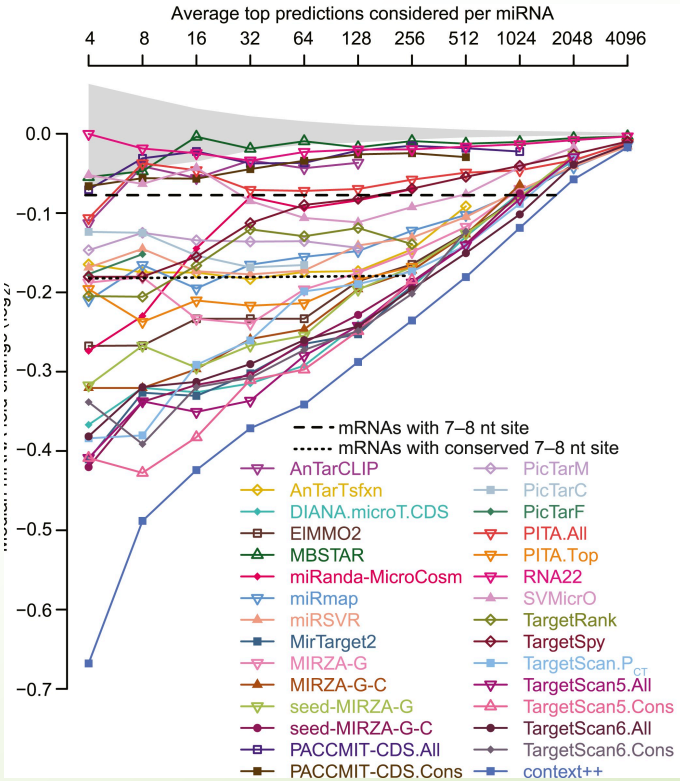
miRNA target search using bioinformatic tools



- Different factors that determines the regulation of a miRNA mRNA interaction
  - Duplex features
  - Local Context features
  - Global context features
- Target prediction programs
  - miRanda-mirSVR
  - TargetScan

Still hard to know which targets are real





Predicting effective microRNA target sites in mammalian mRNAs *Agarwal et al*. 2015

Relation between microRNAs and their predicted targets

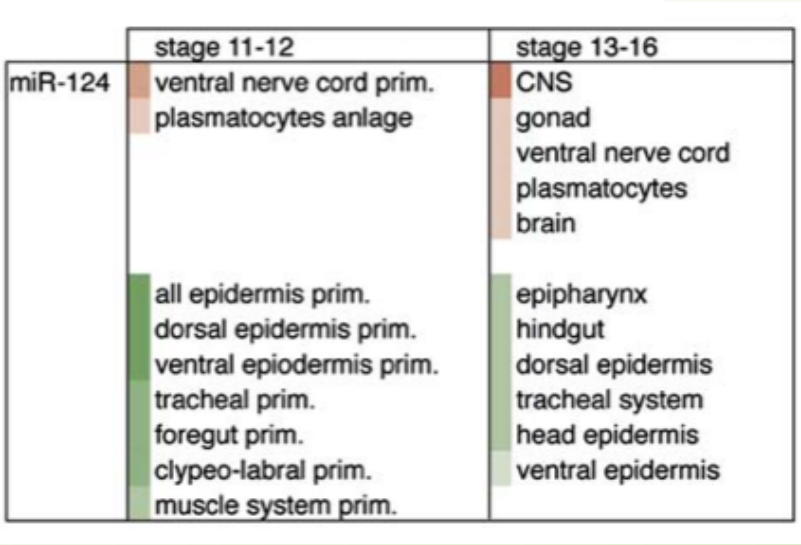
It is possible to find statistical correlations between expression of microRNAs and of their predicted target genes.

Example: mir-124 is expressed in the nervous system.

Neural genes are depleted for mir-124 target sites in the 3' UTRs.

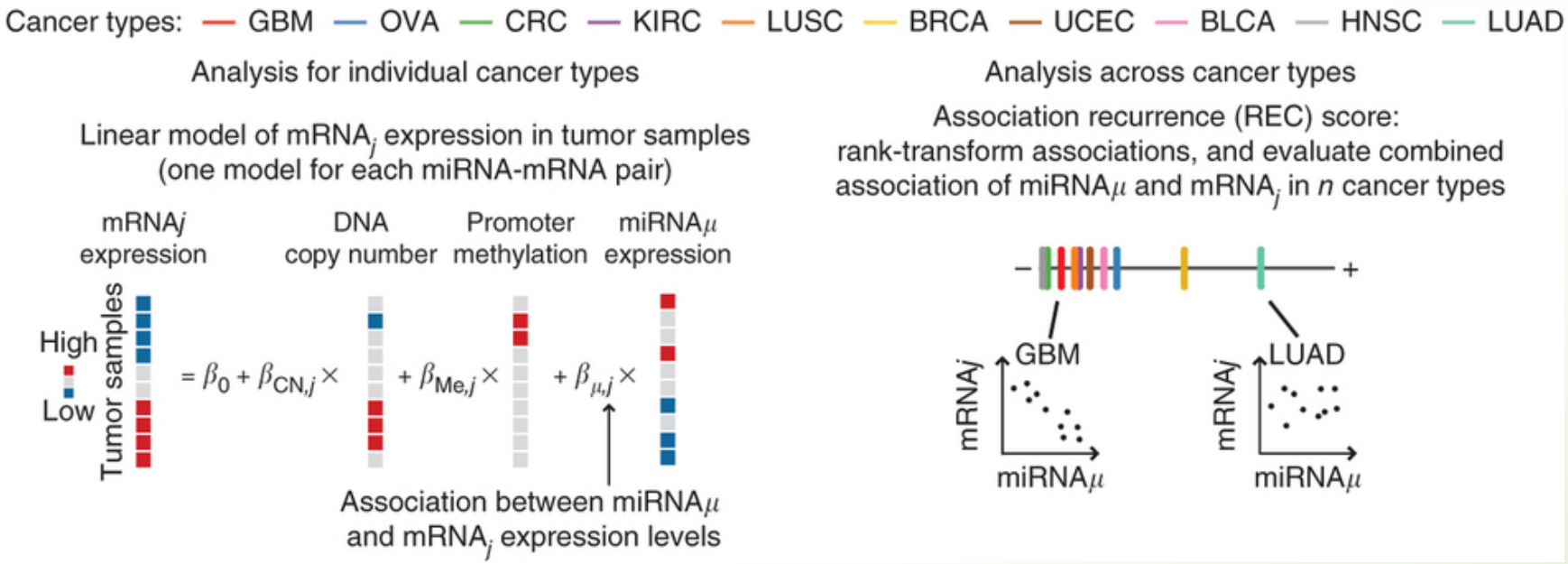
Genes expressed in epidermis, muscle, gut etc. are enriched for mir-124 sites.

But we cannot be sure that this is true, since we are only looking at predicted targets!



(Stark et al. Cell, 2005)

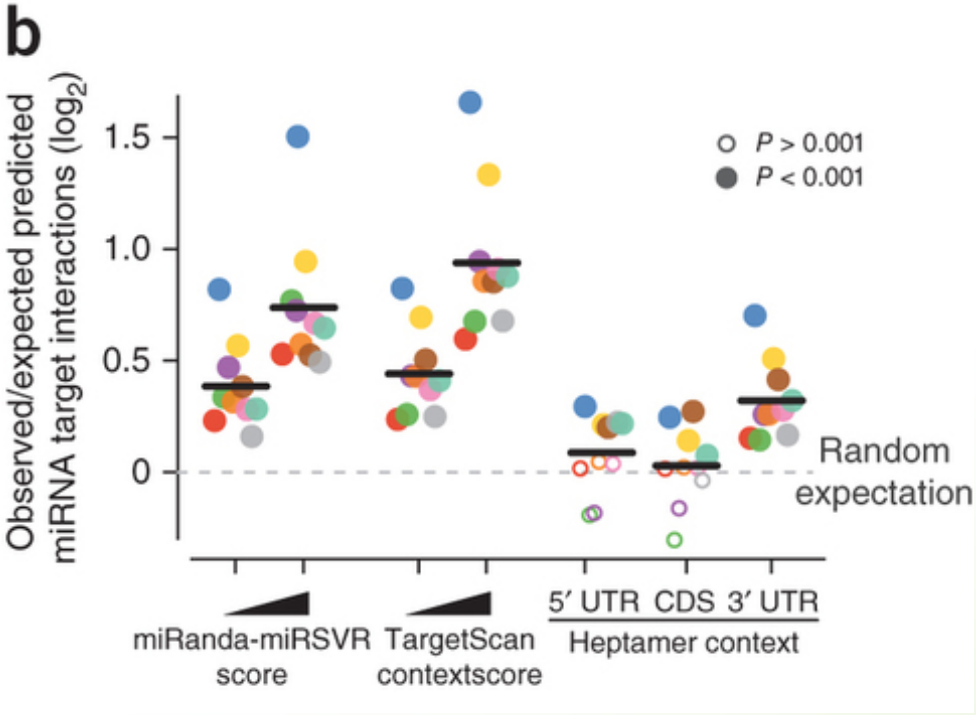
Using expression levels of mRNAs and miRNAs to identify targets



Analysis of microRNA-target interactions across diverse cancer types
Anders Jacobsen et al.
Nature Structural & Molecular Biology 20, 1325–1332 (2013)

Karolinska Institutet · KTH Royal Institute of Technology · Stockholm University · Uppsala Universitet · SciLifeLab

Positive correlation between prediction programs and expression correlation



Analysis of microRNA-target interactions across diverse cancer types
Anders Jacobsen et al.
Nature Structural & Molecular Biology 20, 1325–1332 (2013)

Other strange small RNAs that show up in sequencing data

mirtrons

piRNAs

tRNA fragments

hp-RNAs

TSS-microRNAs

cis-natRNAs

- Some of these are functional
- Some are by products of RNA processing, and can be informative (e.g. microRNA loop sequences).
- Some are probably just “noise”.

THE END