Small RNAs and how to analyze them using sequencing

RNA-seq Course

November 8th 2017

Marc Friedländer

Computational RNA Biology Group

SciLifeLab / Stockholm University

Special thanks to Jakub Westholm for sharing slides!

Small RNAs

- Small RNAs are species of short non-coding RNAs, by definition <200 nucleotides
  - microRNAs (miRNAs)
  - short interfering RNAs (siRNAs)
  - piwi associated RNAs (piRNAs)
  - clustered regularly interspaced short palindromic repeats (CRISPRs)
  - mirtrons, cis-natRNAs, tasi-RNAs, enhancer RNAs and other strange things

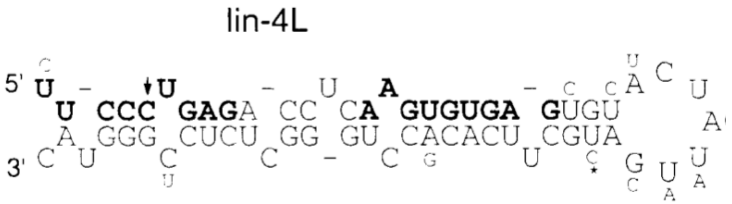
1. Background on regulatory small RNAs

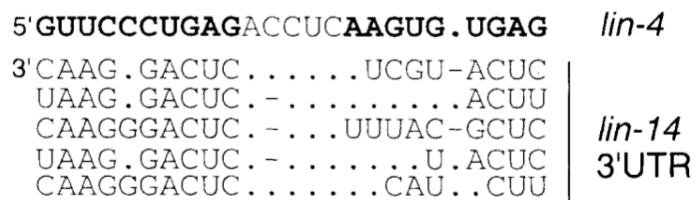
1993: Discovery of first miRNA

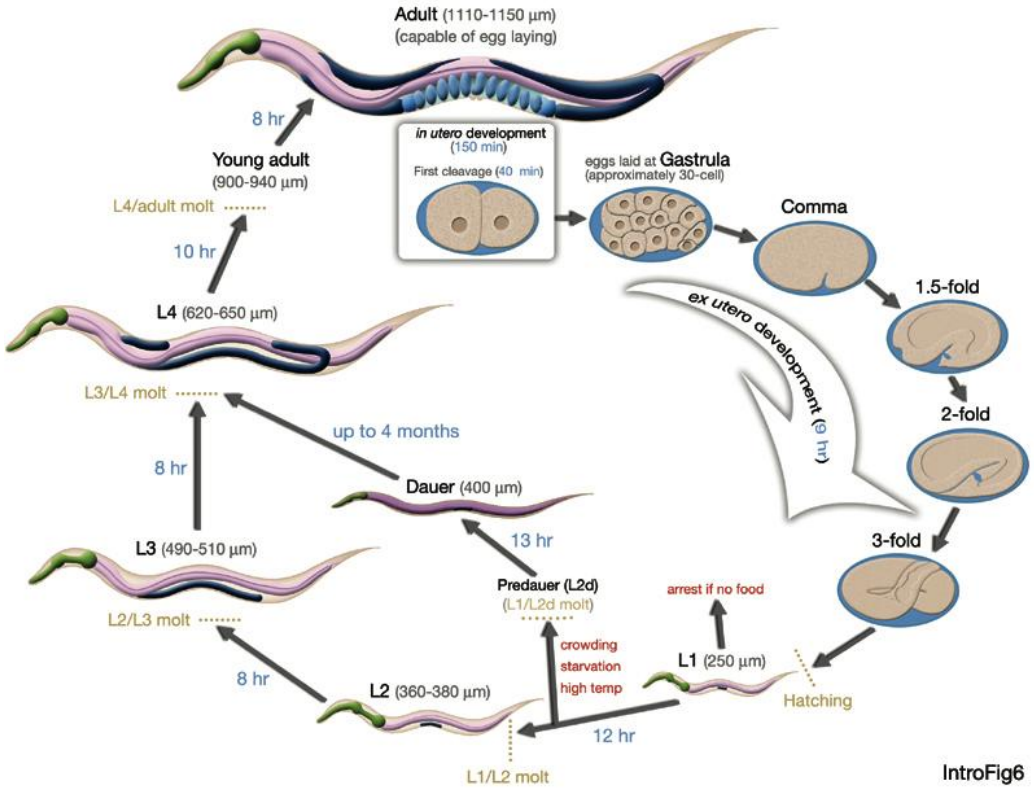
- *lin-4* necessary for progression to L2

- it was found that *lin-4* is an RNA!

- part of this RNA is complementary to the 3'UTR of the gene *lin-14*



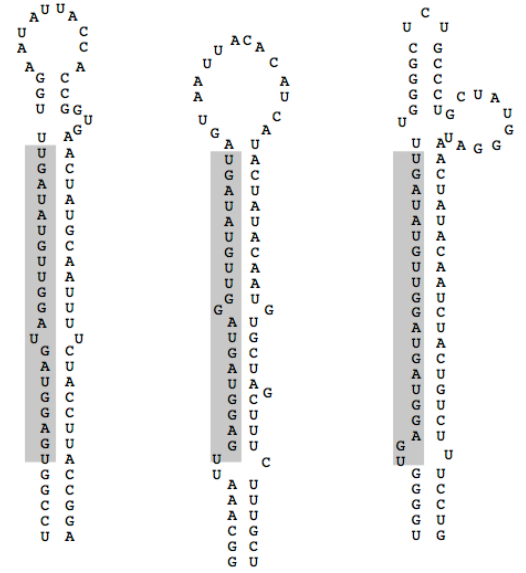




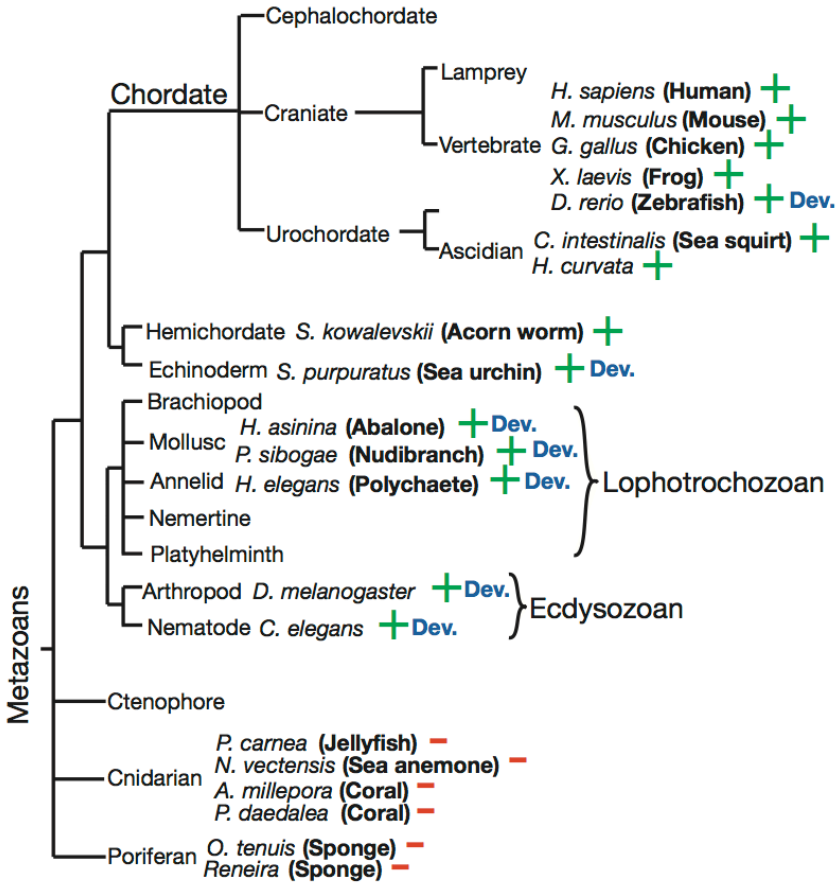
2000: a second, conserved, microRNA is found

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*

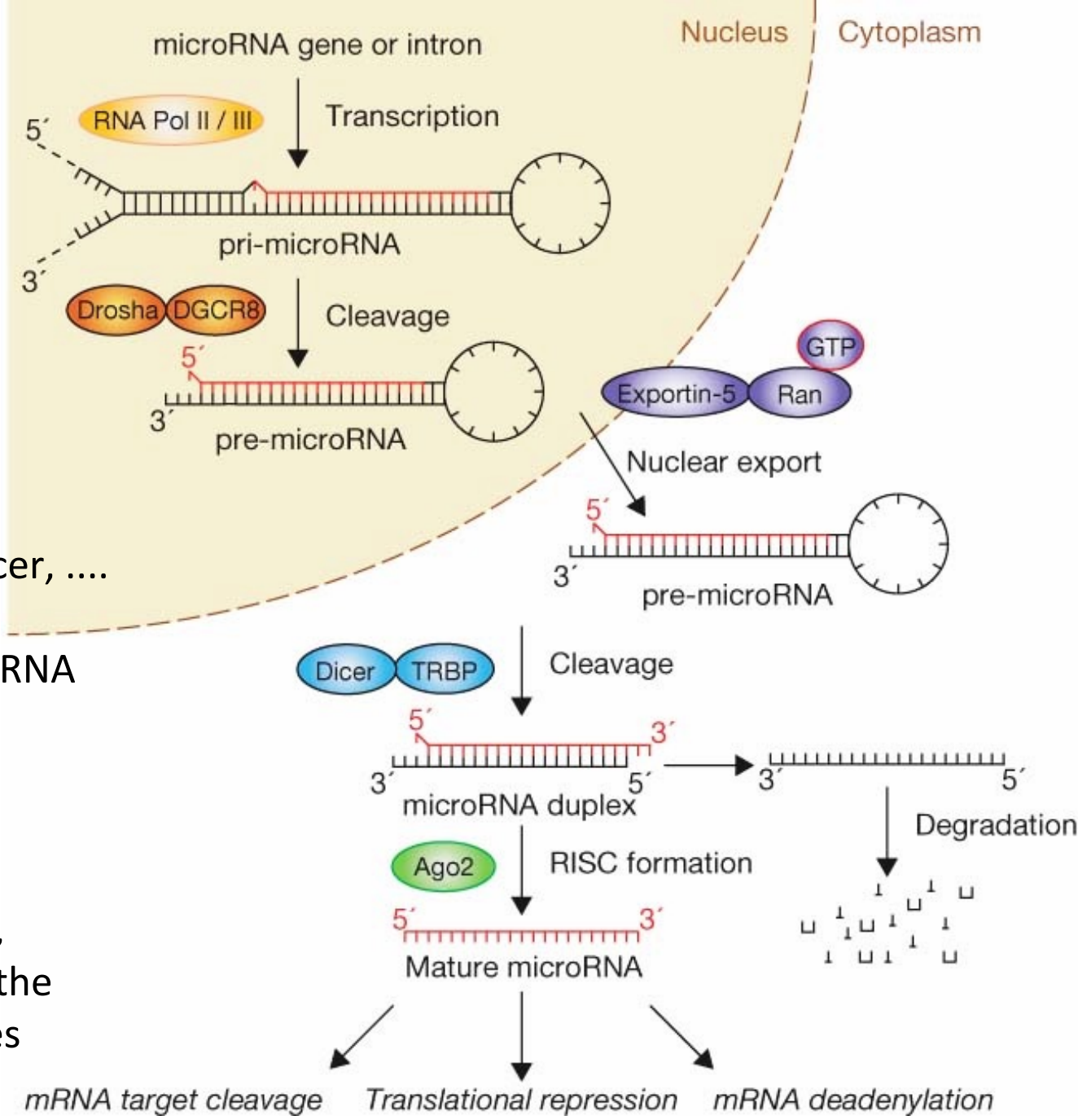
Using:

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

Identification of Novel Genes Coding for Small Expressed RNAs

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

microRNA biogenesis



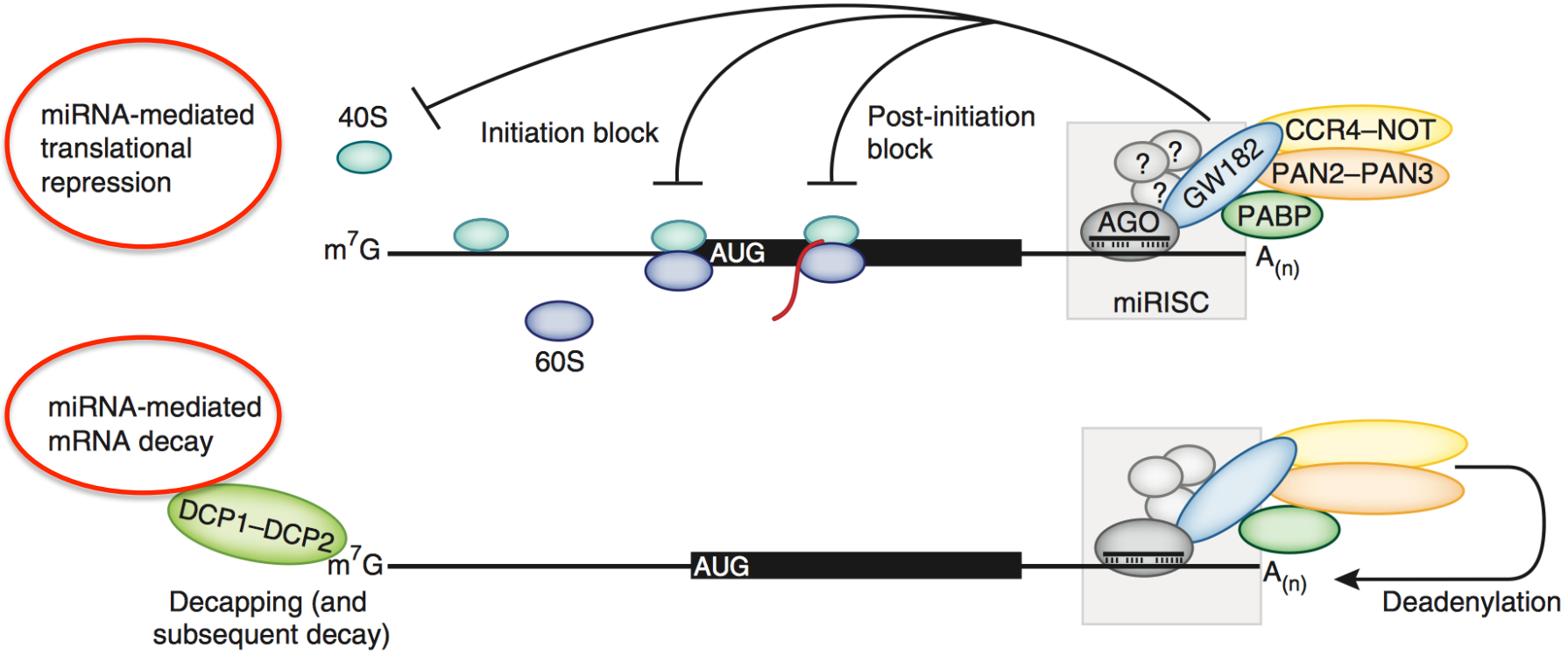
- Many enzymes etc. are involved: Drosha, Exp5, Dicer, ...

- 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

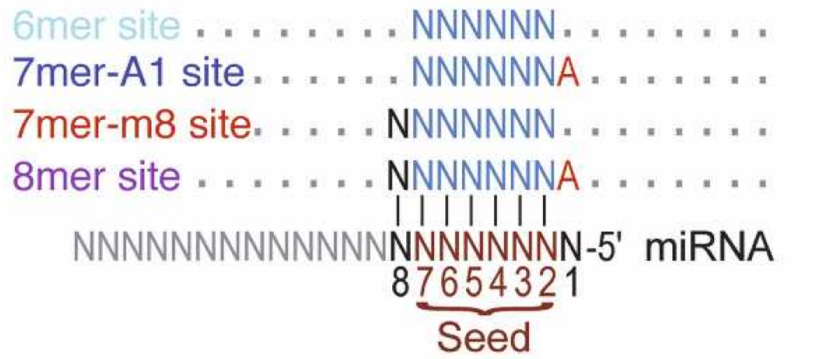


(This is in animals. microRNAs in plants work differently.)

(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, PicTar, ..)

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

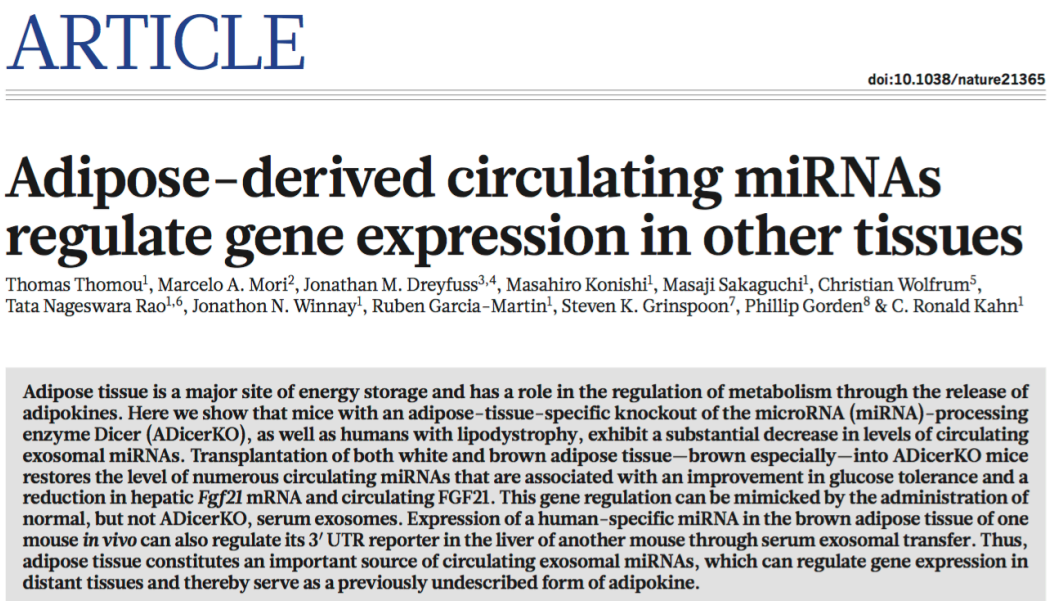
MicroRNAs in animal genomes

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

- In a given tissue, their expression can range over more than 5 orders of magnitude (a few to > 100,000 molecules per cell)

microRNAs regulate many biological processes and are involved in disease

- Development
- Differentiation
- Formation of cell identity
- Stress response
- Cancer
- Cardiovascular disease
- Inflammatory disease
- Autoimmune disease



ARTICLE

doi:10.1038/nature21365

Adipose-derived circulating miRNAs regulate gene expression in other tissues

Thomas Thomou[1], Marcelo A. Mori[2], Jonathan M. Dreyfuss[3,4], Masahiro Konishi[1], Masaji Sakaguchi[1], Christian Wolfrum[5], Tata Nageswara Rao[1,6], Jonathon N. Winnay[1], Ruben Garcia-Martin[1], Steven K. Grinspoon[7], Phillip Gorden[8] & C. Ronald Kahn[1]

Adipose tissue is a major site of energy storage and has a role in the regulation of metabolism through the release of adipokines. Here we show that mice with an adipose-tissue-specific knockout of the microRNA (miRNA)-processing enzyme Dicer (ADicerKO), as well as humans with lipodystrophy, exhibit a substantial decrease in levels of circulating exosomal miRNAs. Transplantation of both white and brown adipose tissue—brown especially—into ADicerKO mice restores the level of numerous circulating miRNAs that are associated with an improvement in glucose tolerance and a reduction in hepatic *Fgf21* mRNA and circulating FGF21. This gene regulation can be mimicked by the administration of normal, but not ADicerKO, serum exosomes. Expression of a human-specific miRNA in the brown adipose tissue of one mouse *in vivo* can also regulate its 3′ UTR reporter in the liver of another mouse through serum exosomal transfer. Thus, adipose tissue constitutes an important source of circulating exosomal miRNAs, which can regulate gene expression in distant tissues and thereby serve as a previously undescribed form of adipokine.

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 15-30 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.

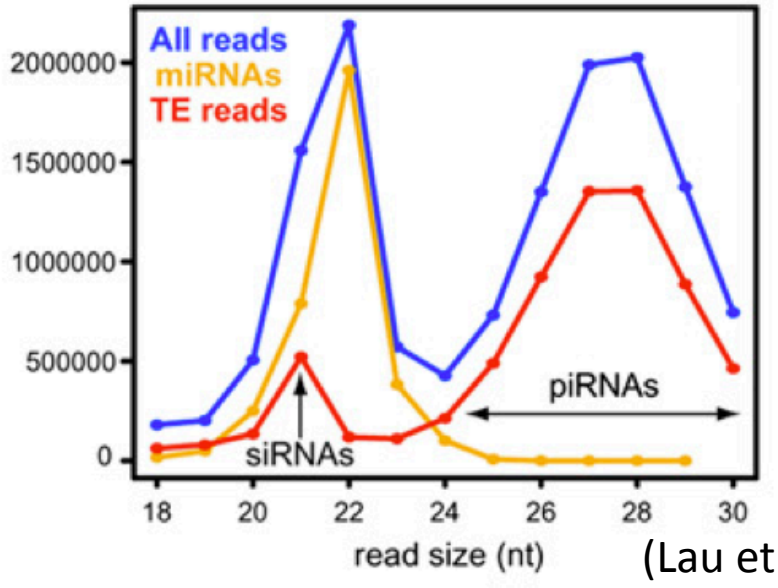
Pre-processing of small RNA data II

- 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

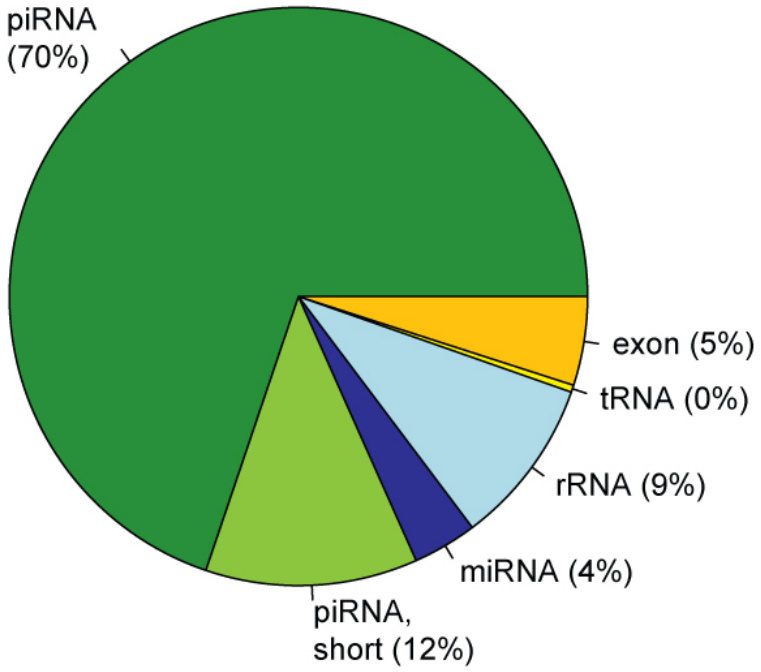
  - Long reads are probably not microRNAs



(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:



(Figure from Friedländer *et al.*, PNAS, 2009)

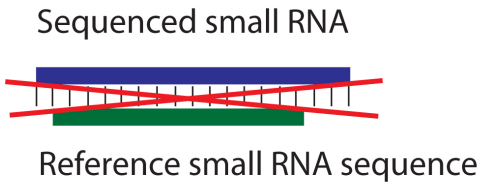
Small RNA expression profiling

-The number of times a small RNA is sequenced is a function of its expression

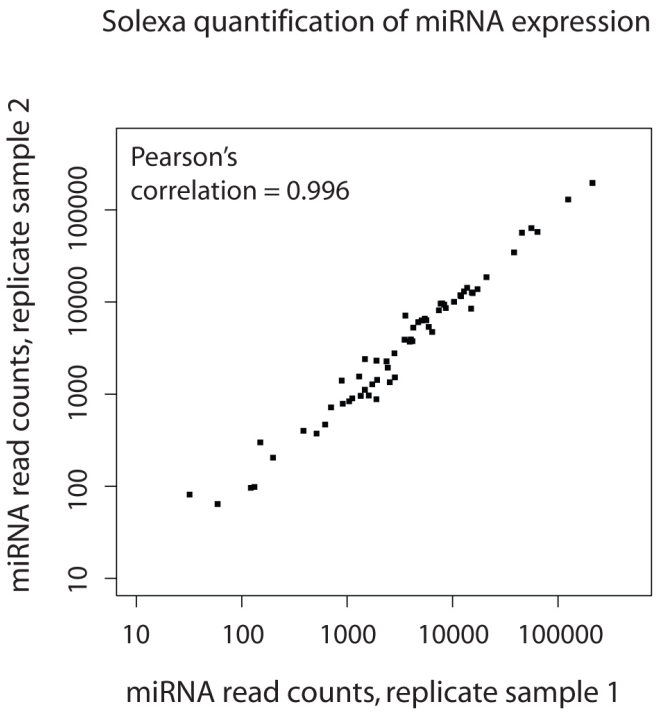
-to count this number, the sequenced small RNAs must first be compared to reference sequences

-however some reference small RNA sequences are truncated, making mapping against them difficult

-It is more robust to map the sequenced RNAs against the genome/precursors



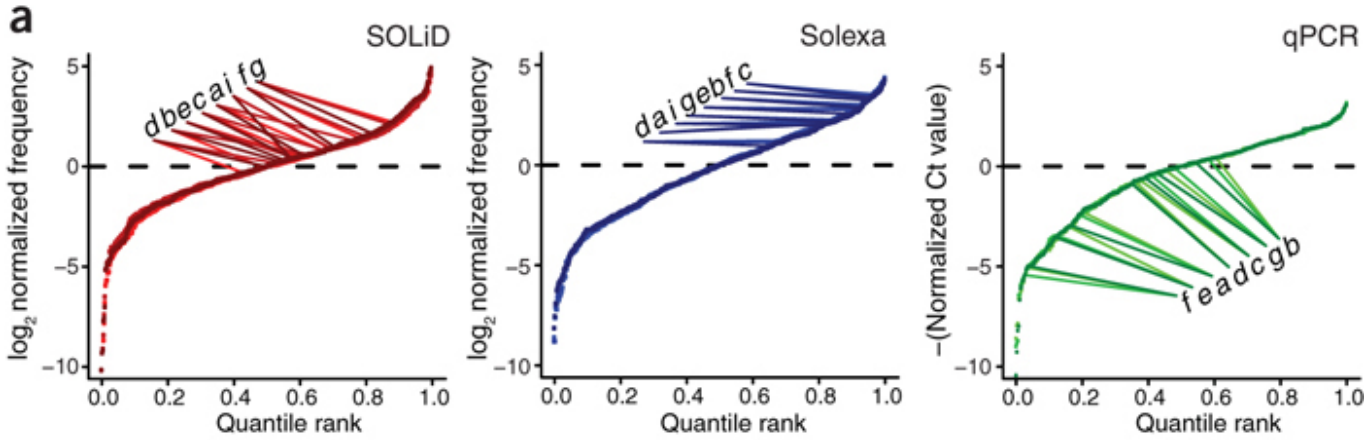
Small RNA-seq is reproducible



Sequencing frequency of microRNAs in planarian biological replicates

(Figure from Friedländer *et al.,* PNAS. 2009)

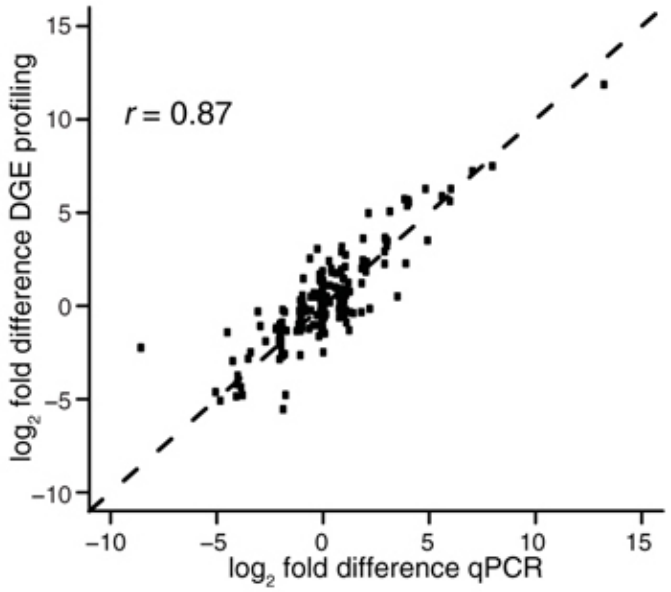
Small RNA-seq cannot measure absolute abundances



Sequencing frequency of 473 artificial microRNAs in equal abundance

(Figure from Linsen *et al.,* Nature Methods. 2009)

Small RNA-seq can measure relative abundances (fold-changes)



Fold-changes: deep sequencing vs. qPCR

(Figure from Linsen *et al.,* Nature Methods. 2009)

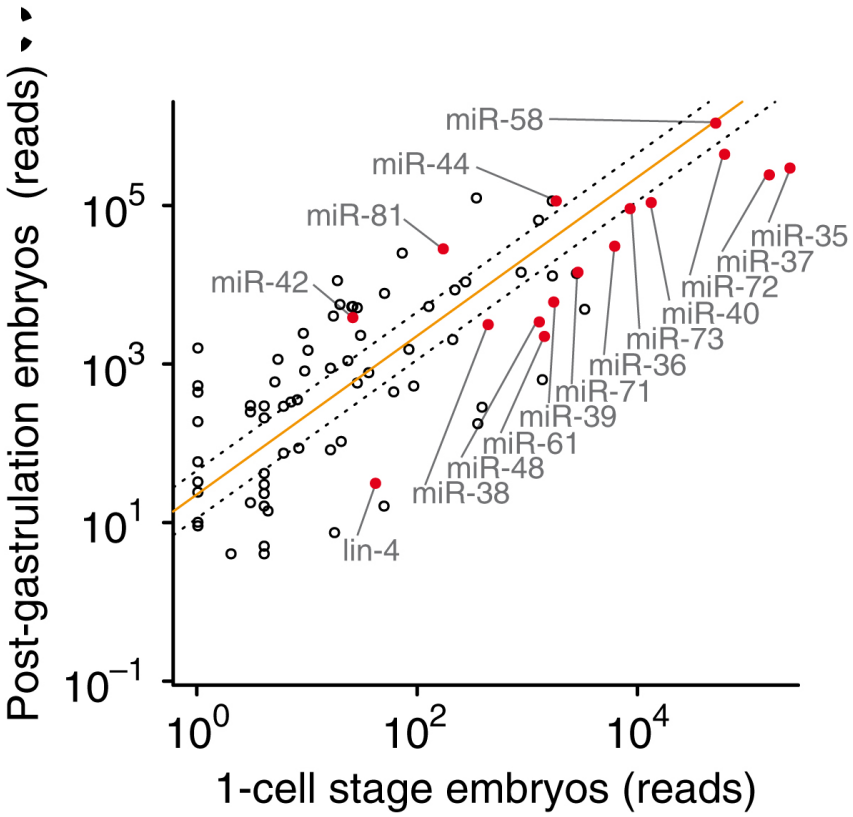
Identifying differentially expressed small RNAs

-Once the sequence data is transformed to counts, they are in essence not different from ordinary RNA-seq data

-microRNA counts should be normalized to the total miRNA counts in the sample (RPM) or to 'trimmed mean of M-values' (TMM)

-for comparisons between two datasets, an initial eyeballing works as sanity check

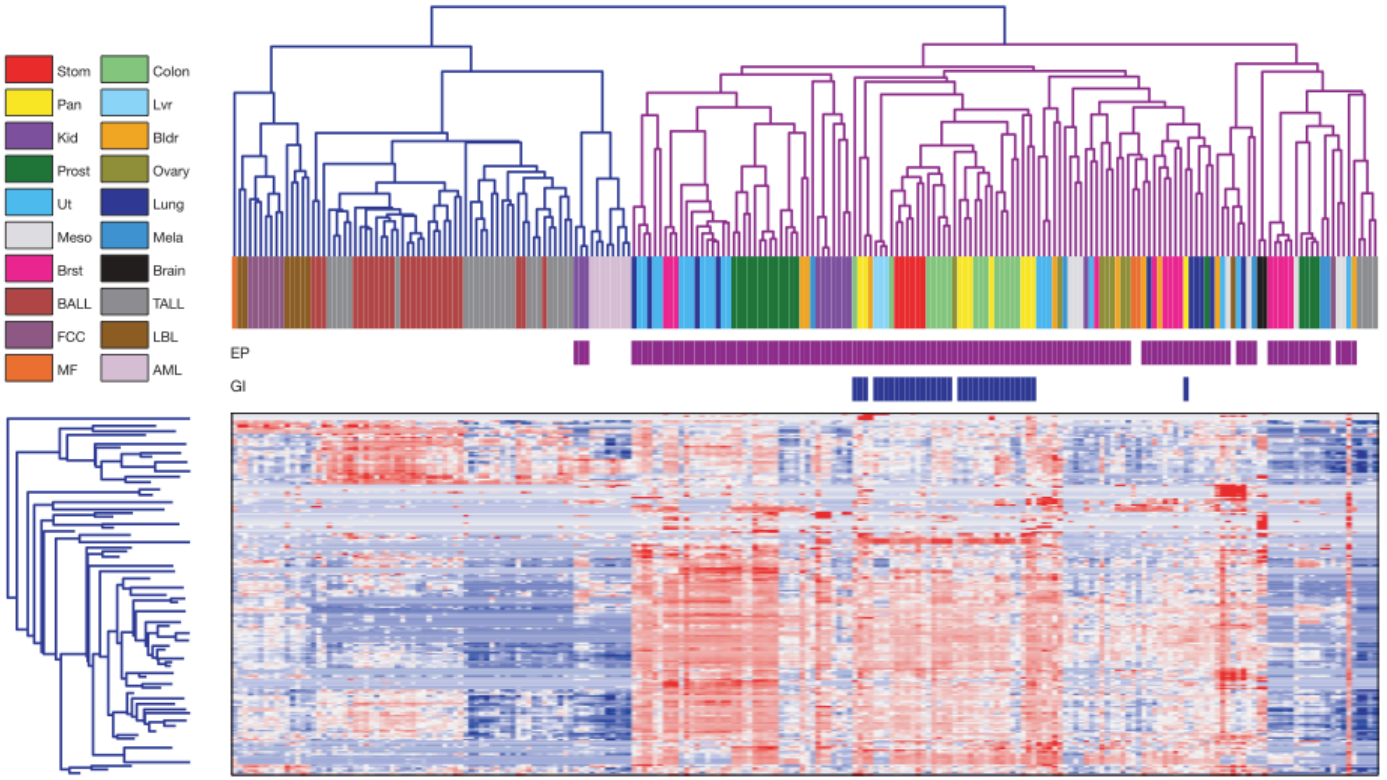
Dedicated tools:
- DEseq2
- edgeR
- NOISEQ



(Figure from Stoeckius *et al.*, Nature Methods, 2009)

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

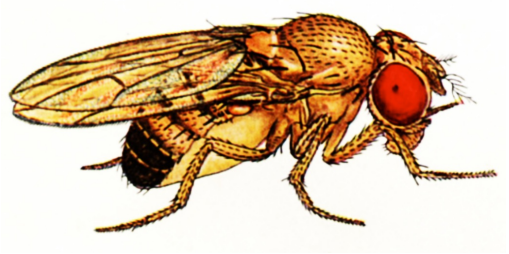
Cancer type	miRNAs[a]	Ref.
Breast		
ER status	miR-26a/b, miR-30 family, miR-29b, miR-155, miR-342, miR-206, miR-191	[38–40,42]
PR status	let-7c, miR-29b, miR-26a, miR-30 family, miR-520g	[41,42]
HER2/*neu* status	miR-520d, miR-181c, miR-302c, miR-376b, miR-30e	[38,41]
Lung		
Squamous vs non-squamous 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]

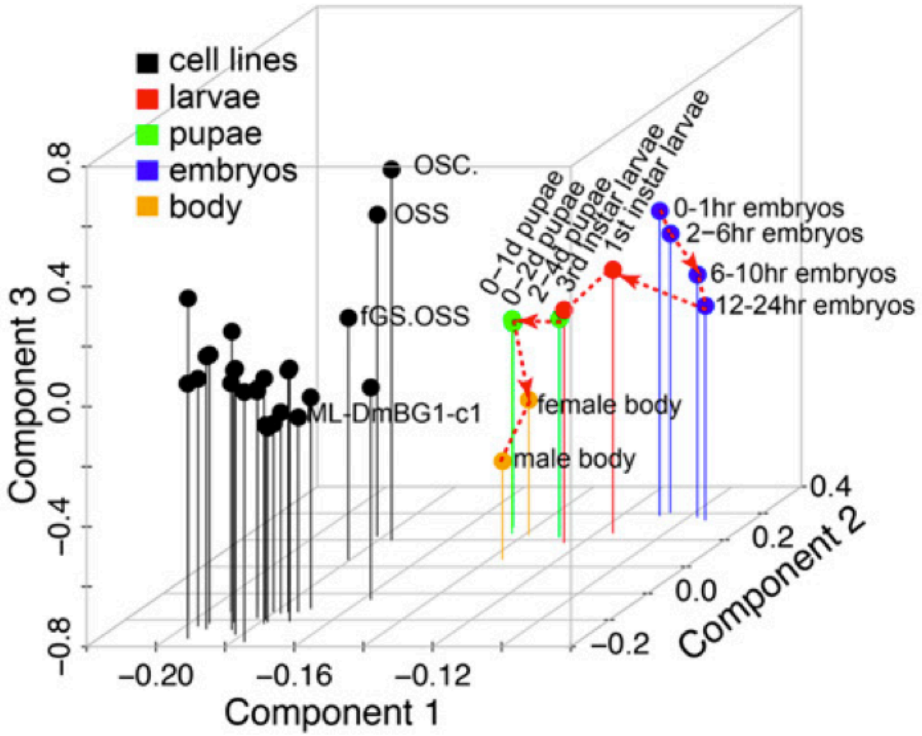
[a]Not all distinguishing miRNAs are represented in this table.

[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 than to normal tissues.

(Wen et al. Genome Research 2014)

microRNA discovery by small RNA-seq: challenges

NGS can detect hundreds of millions of small RNAs in one run

- however, many of the sequenced RNAs are degradation products from:
  - rRNAs, tRNAs, mRNAs, snRNAs, snoRNAs
  - un-annotated transcripts
- when the RNAs are mapped to the genome, they often map to millions of loci
- only a few hundreds of these loci are in fact microRNA genes
- thus, the non-trivial task of accurately classifying microRNA gene loci remains!

miRDeep: first algorithm to discover microRNAs in small RNA-seq data

- first and most widely used algorithm for microRNA discovery (>800 studies)

- probabilistic (reports probability that a given sequence is a microRNA)

- independent of:

    species

    conservation information

    genome annotation

    state of genome assembly

- incorporates our knowledge of microRNA biogenesis

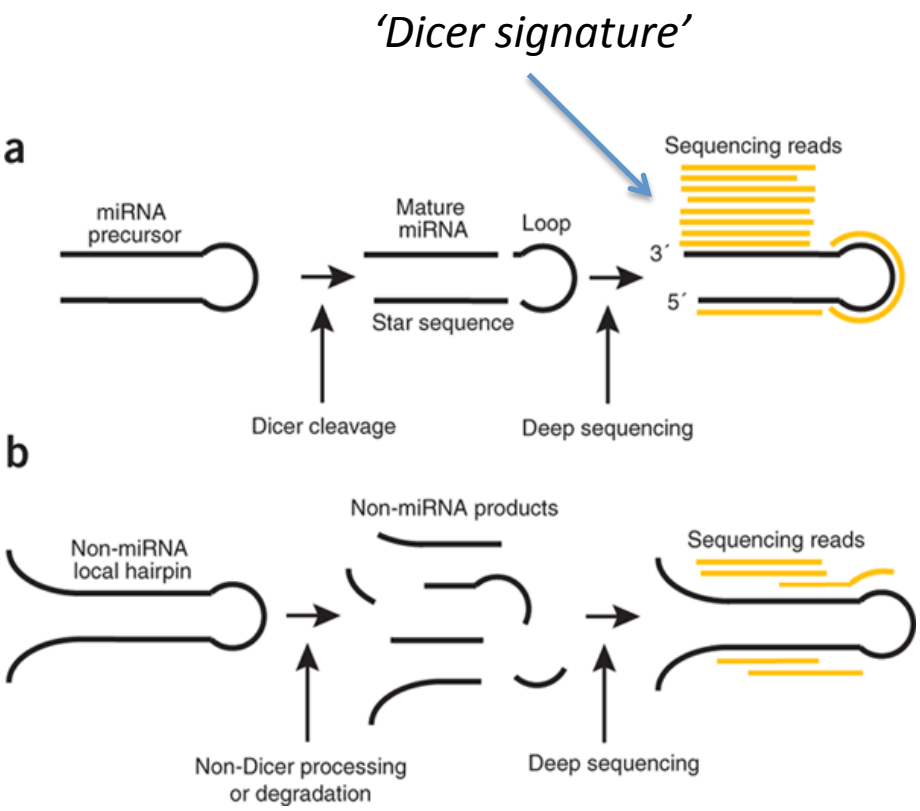
Key idea behind miRDeep(2)

Novel microRNAs are discovered in a three step process:

1: frequently sequenced RNAs are identified *('read stacks')*

2: the read stacks should overlap an RNA hairpin structure

3: the position of the stacks in hairpin should conform to Dicer processing (*'Dicer signature'*, a)



(Figure from Friedländer *et al.*, Nature Biotech. 2008)

Log-odds scoring function

$$\text{Score} = \log \frac{P(pre|data)}{P(bgr|data)} \quad (1)$$

$$P(pre|data) = P(data|pre)\ P(pre)/P(data) \quad (2)$$

$$P(data|pre) = P(sig|pre)\, P(star|pre)\, P(mfe|pre)\, P(rel|pre)\, P(con|pre) \quad (3)$$

$$P(bgr|data) = P(data|bgr)\ P(bgr)/P(data) \quad (4)$$

$$P(data|bgr) = P(sig|bgr)\, P(star|bgr)\, P(mfe|bgr)\, P(rel|bgr)\, P(con|bgr) \quad (5)$$

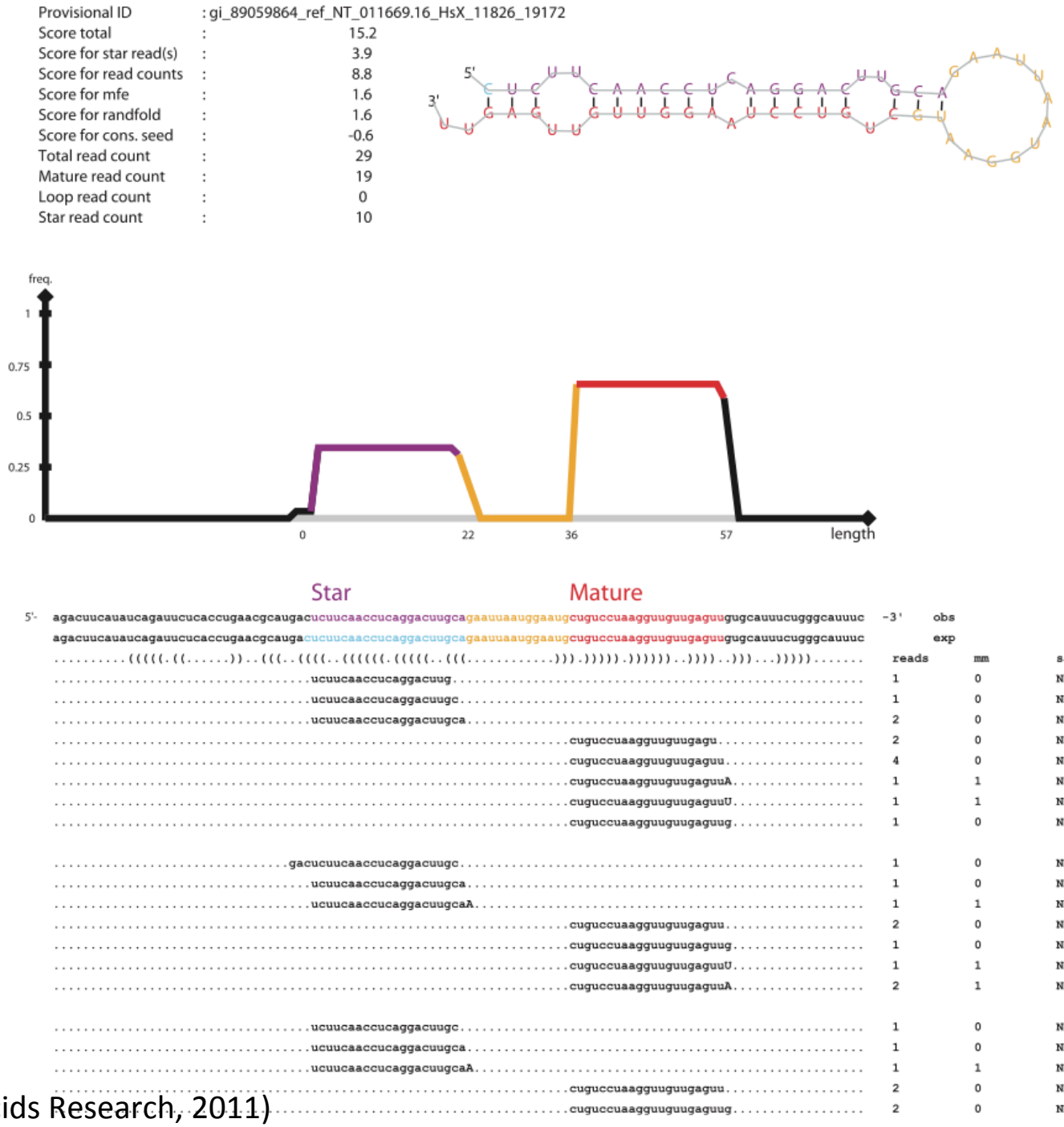
Pre: the hairpin is a genuine microRNA

Bgr: the hairpin is a (non-microRNA) background hairpin

- 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. miRCat2 which also finds other small RNAs.



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

Other strange small RNAs that show up in sequencing data

mirtrons

piRNAs

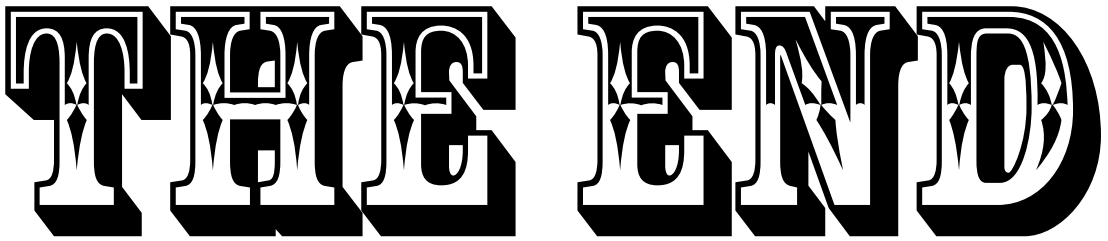
tRNA fragments

yRNAs

tasi-RNAs

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