Encyclopedia of short RNAs

Alessandra Breschi

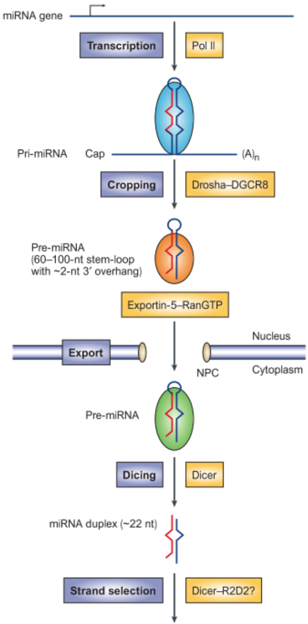
ENCODE AWG call

November 10th, 2015

What are miRNAs?

- non-coding RNAs
- small (approx 22 nt)
- Are derived from a primary transcript which carries one or more hairpin structures which are cleaved to produce the mature miRNA

Kim, 2005, Nat. Rev.



Mapping parameters for ENCODE3 CSHL short RNA-seq data

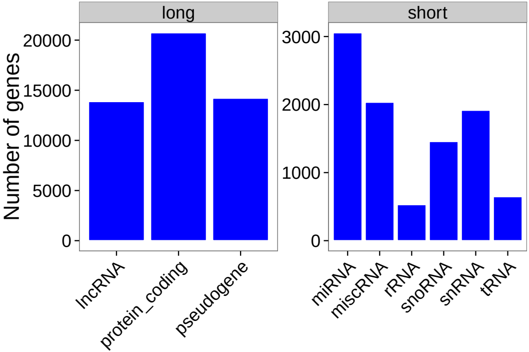
The currently submitted data are with CSHL short RNAseq pipeline. (mapping and quantification)
Comparison with a subset of UCI miRNA pipeline

Step/parameter	**CSHL**	**UCI**
Adapter trimming	STAR	cutadapt
Map to	genome+ann	genome
Number of multimaps	20	10
Number of mismatches	3	0
Multimap score distance	1	0
Seed length	16	16
Annotated SJ detection	NO	NO
Novel SJ detection	5nt	5nt

Annotation

GENCODE v19 (2013-12-05, hg19)

- 3,055 hairpins (pre-miRNAs), 20 on chrY



miRBase v19 (2012-7-23, hg19)

- 1,595 hairpins (pre-miRNAs), 2 on chrY
- 2,233 mature miRNAs
- 638 hairpins can give 2 mature miRNAs

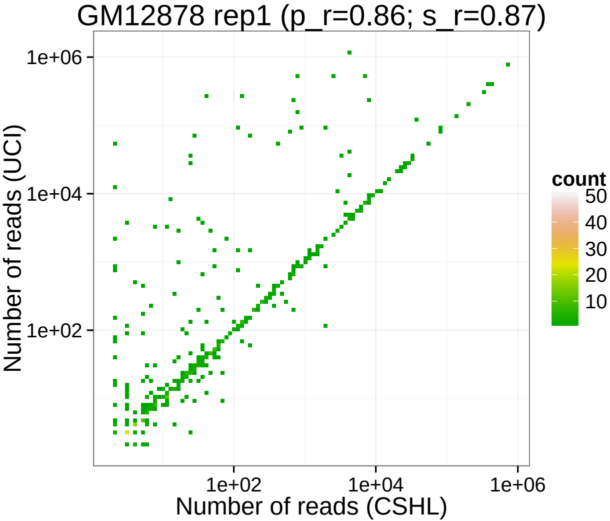
Comparison between GENCODE and miRBase:

- GENCODE does NOT annotate mature miRNAs
- 1,345 hairpins in common
- 250 miRBase only
- 1,710 GENCODE only

GENCODE annotation of miRNAs (`http://uswest.ensembl.org/info/genome/genebuild/ncrna.html`)
miRNAs are **predicted by BLASTN** of genomic sequence slices against miRBase sequences. All species are used. The BLAST hits are clustered and filtered by E value and the aligned genomic sequence is then checked for possible secondary structure using **RNAFold**. If evidence is found that the genomic sequence could form a stable hairpin structure, the locus is used to create a miRNA gene model. The resulting BLAST hit is used as supporting evidence for the miRNA gene.

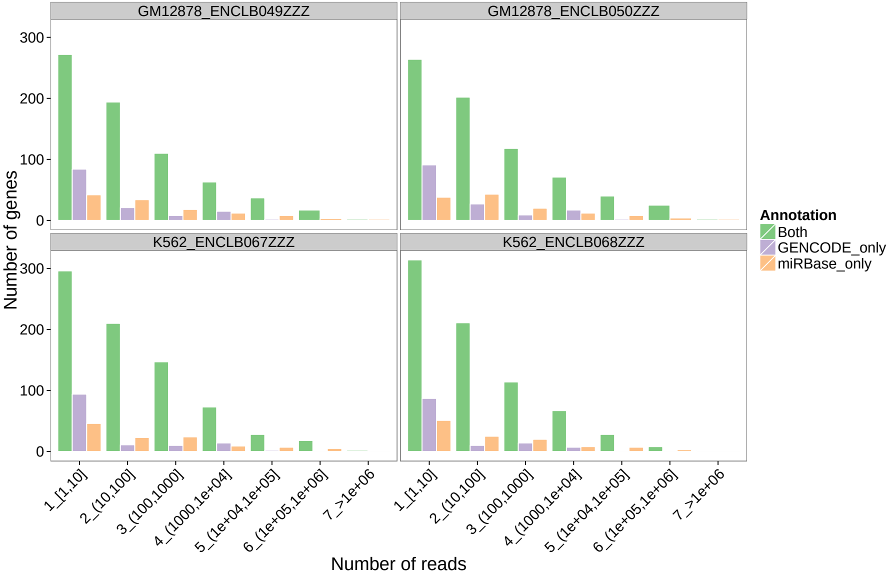
Note: The **miRNA identifier** and name are only associated to the resulting Ensembl miRNA if they are of the same species.

The two pipelines are well correlated

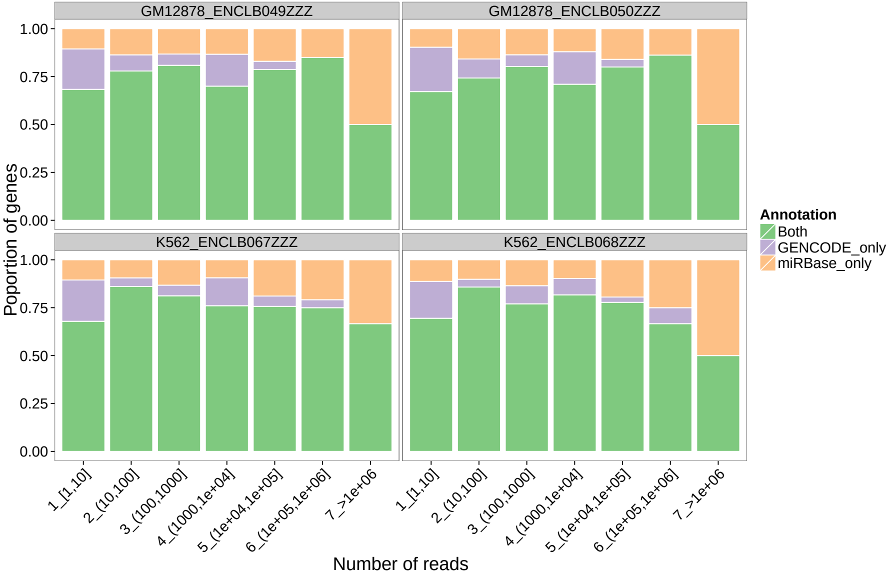


825 hairpins

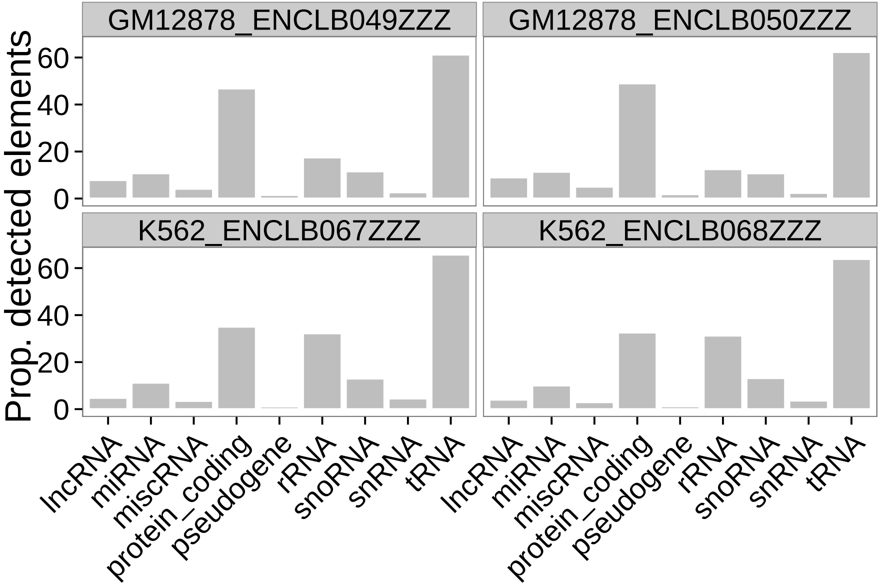
Hairpins common to both annotations are the most detected (UCI pipeline)



Hairpins common to both annotations are the most detected



Detected elements (CSHL pipeline)



Discussion points

- Pre-miRNA (hairpin) annotation: miRBase vs GENCODE
  - Only GENCODE
  - Only miRBase (remove miRNA from GENCODE and concatenate miRBase to the remaining GENCODE annotation)
  - Union of GENCODE and miRBase, ask GENCODE if this is possible, or implemented in ENCODE for each new GENCODE release
- Mature miRNAs annotation: not in GENCODE.
- Elements to report in the matrix:
  - All GENCODE elements (+ miRNAs as decided in the point above)
  - Only miRNAs matrix as separate matrix
  - Only short RNAs matrix as separate matrix
- ...

Acknowledgments

Guigó lab

- Roderic Guigó
- Sarah Djebali

CRG
Centre
for Genomic
Regulation

Gingeras lab

- Tom Gingeras
- Carrie Davis
- Alex Dobin

CSH

Mortazavi lab

- Ali Mortazavi
- Rabi Murad

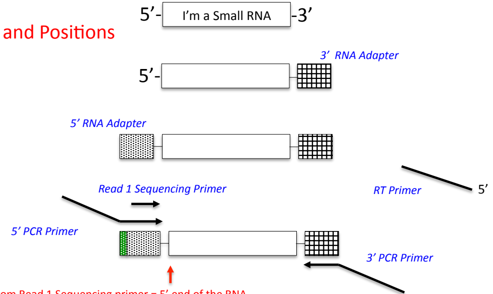
UCI

Suppplementary slides

5’ and 3’ Ligation with the Barcode in the 3’ read position

(Used in ENCODE Phase III)

Review Primers and Positions



First nucleotide read off from Read 1 Sequencing primer = 5’ end of the RNA

5’ RNA Adapter: 5’ –
3’ RNA Adapter: 5’ -
RT Primer: 5’ –
5’ PCR Primer: 5’- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC
3’ PCR Primer: 5’- CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTC
Sequencing Primer: 5’-

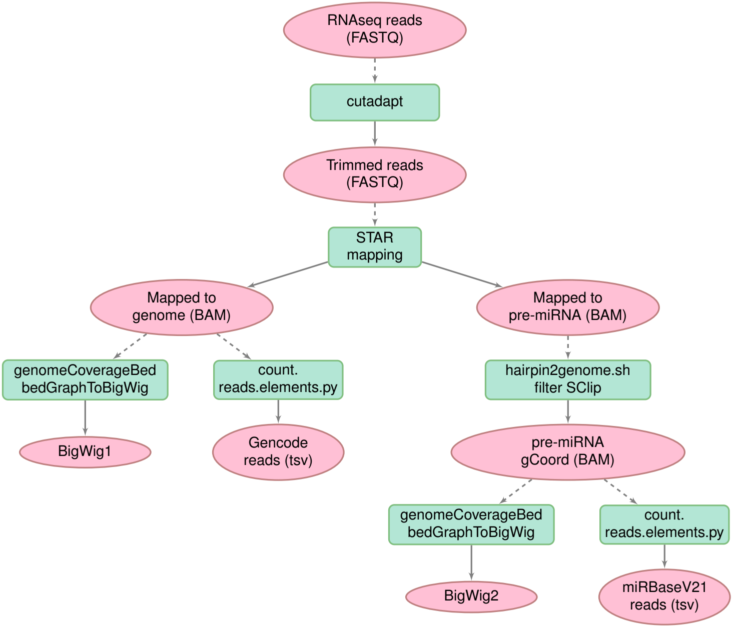
7

Carrie Davis (CSHL)

Reads statistics

labExpId	cell	totReads	adaptReads	adaptProp	finalReads	finalProp
ENCLB049ZZZ	GM12878	109,952,304	98421452	89.5	99630202	90.6
ENCLB050ZZZ	GM12878	111,410,391	107566717	96.5	91840606	82.4
ENCLB067ZZZ	K562	118,643,226	92478132	77.9	102624989	86.5
ENCLB068ZZZ	K562	96,745,250	77901801	80.5	88657954	91.6

UCI short RNA-seq pipeline for CSHL data


RNAseq reads (FASTQ)
cutadapt
Trimmed reads (FASTQ)
STAR mapping
Mapped to genome (BAM)
Mapped to pre-miRNA (BAM)
genomeCoverageBed bedGraphToBigWig
count. reads.elements.py
hairpin2genome.sh filter SClip
BigWig1
Gencode reads (tsv)
pre-miRNA gCoord (BAM)
genomeCoverageBed bedGraphToBigWig
count. reads.elements.py
BigWig2
miRBaseV21 reads (tsv)