Processing Patterns: Significance and Analysis

Sachin Pundhir

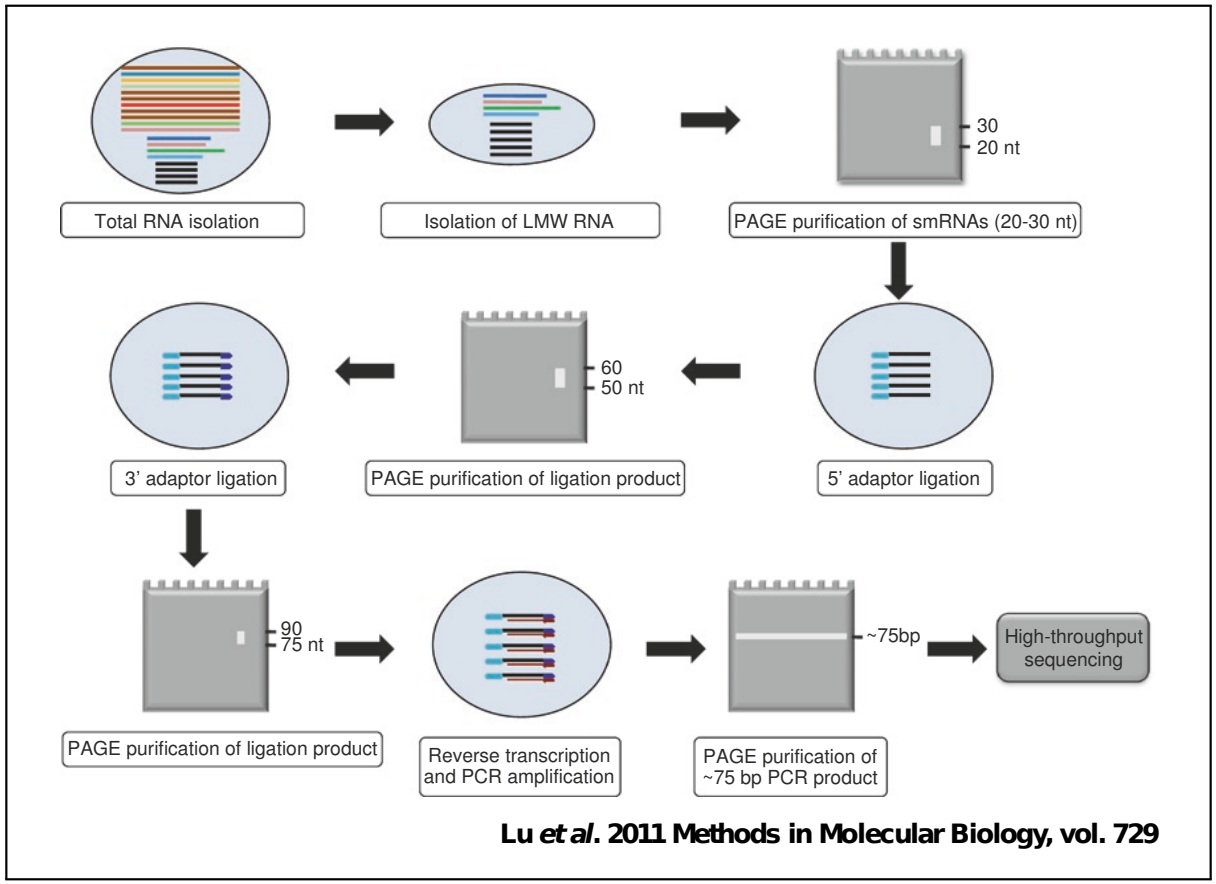
Center for Non-Coding RNA in Technology and Health,
LIFE, University of Copenhagen, Denmark

October 1, 2011

Overview

- Short-RNA sequencing
- Read Processing Pattern?
- Read Processing Pattern: significance
- Read Processing Pattern: quantification
- Read Prcosessing Pattern: alignment and comparison
- DeepBlockAlign: performance evaluation
- Results: one-sample comparison
- Results: multi-sample comparison

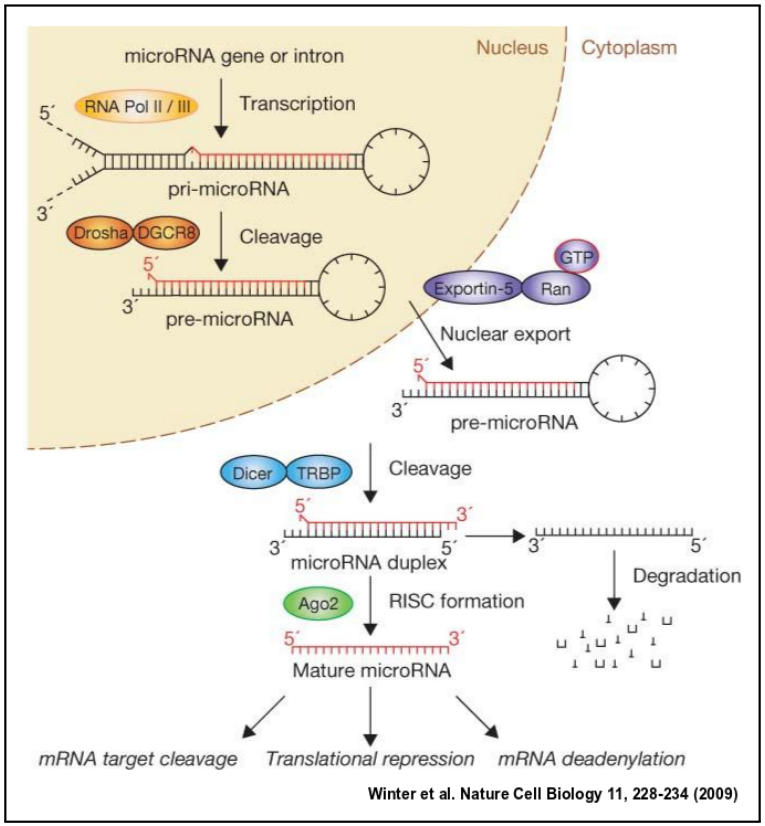
Short-RNA sequencing



- Produce millions of sequences at once eg. 454, Illumina and SOLiD sequencing.
- Variants include poly-A, total RNA, degradome-seq and others.
- Excellent tool for studying species where limited sequence information is available.

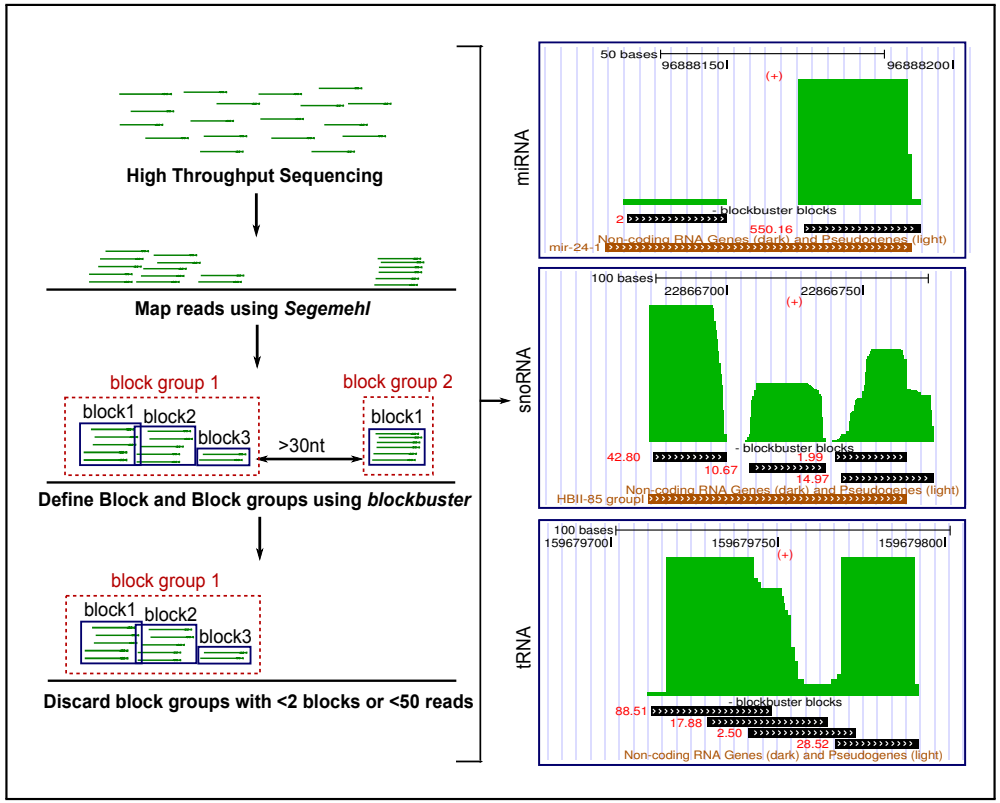
Read Processing Pattern?

- Post-transcriptional processing of transcripts generate short-RNA sequence fragments.
- When mapped to the genome, they form distict patterns termed as 'Read Processing Patterns' eg. miR-miR* pattern from miRNA.



Read Processing Pattern: significance

- Conveys information about the structure of parent transcript and modality of processing.
- Study can lead us to identify commonality and diversity in read processing machanisms.



Read Processing Pattern: quantification

Parameters to quantify Read Processing Patterns (block group)

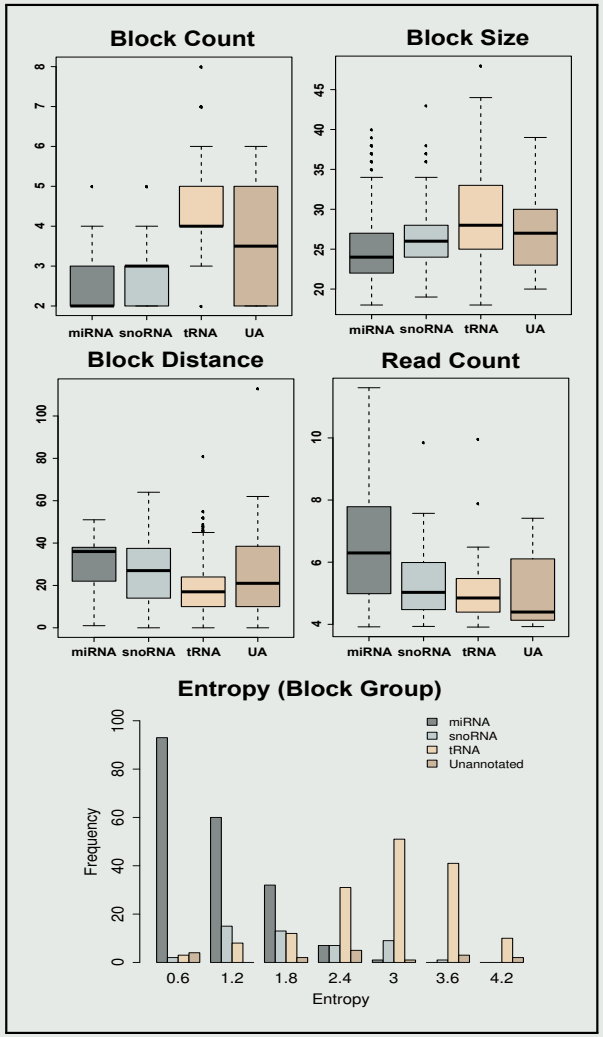
- Block count
- Block size
- Block distance
- Read count
- Entropy (I)

Entropy

$$I = -\sum_i q_i \ log_2(q_i) \quad (1)$$

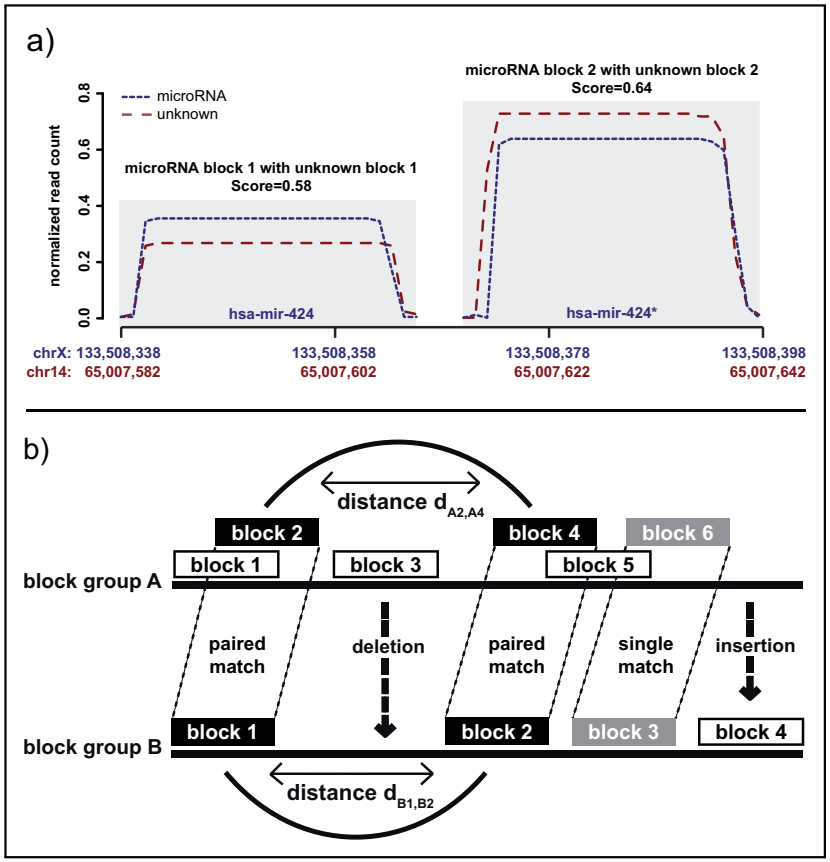
where;
- q_i is the fraction of reads in a block group that starts at position i, and;
- the sum run over all positions where read starts within a block group.

Parameter distribution



Read Processing Pattern: alignment and comparison

DeepBlockAlign: a tool to align and compare two read processing patterns. It employs a two-tired strategy.



a) Block alignment

- Each position i in block $\vec{X}$ is represented as the normalized difference between total reads and start reads, such that $x_i = (x_{1i} - x_{2i})/N_X$.
- Next, a Needleman-Wunsch like algorithm is employed to determine optimal alignment between two block profiles, $\vec{X}$ and $\vec{Y}$.
- In other words, we compute the block scores based on the comparison of *block size*, *read count* and *entropy*.

b) Block group alignment

- A variant of the Sankoff (1985) algorithm is used to compute optimal alignment between block groups.
- Here, a similarity measure is used that combines the *block score* and *block distance*.

DeepBlockAlign: performance evaluation

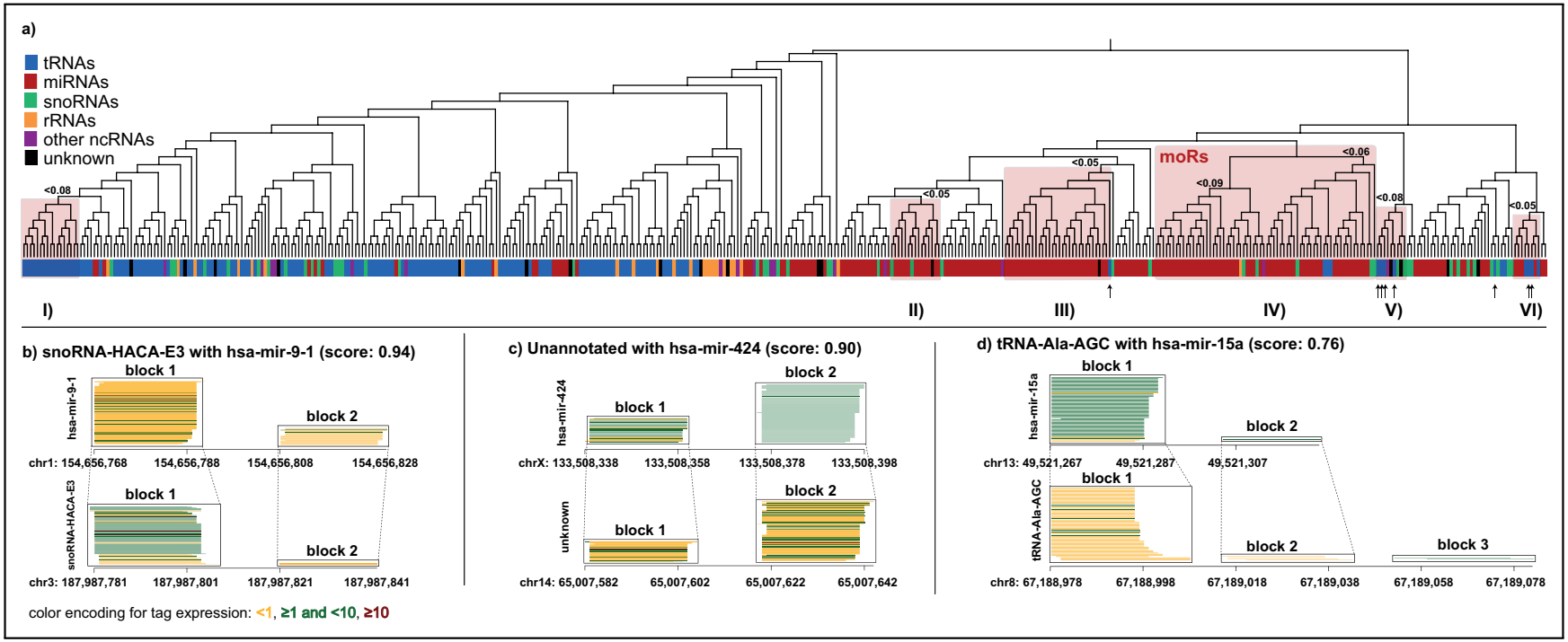
Dataset

Dataset	**#reads**	**#block groups**					
		miRNA	snoRNA	tRNA	Others	Unannotated	Total
Human_eb	7,351,304	193	47	157	40	18	455

Workflow

- All vs All alignment of 455 block groups using *DeepBlockAlign*.
- Average linkage hierarchical clustering using *pvclust*. *pvclust* computes *p*-value for each cluster using multiscale bootstrap resampling and indicates how strong the cluster is supported by the data.
- Analyze all clusters with p-value < 0.1.

Results: one-sample comparison



- Two well-seperated clusters, one with mainly miRNA and other with tRNAs.
- No well defined cluster for snoRNA.

Results: one-sample comparison

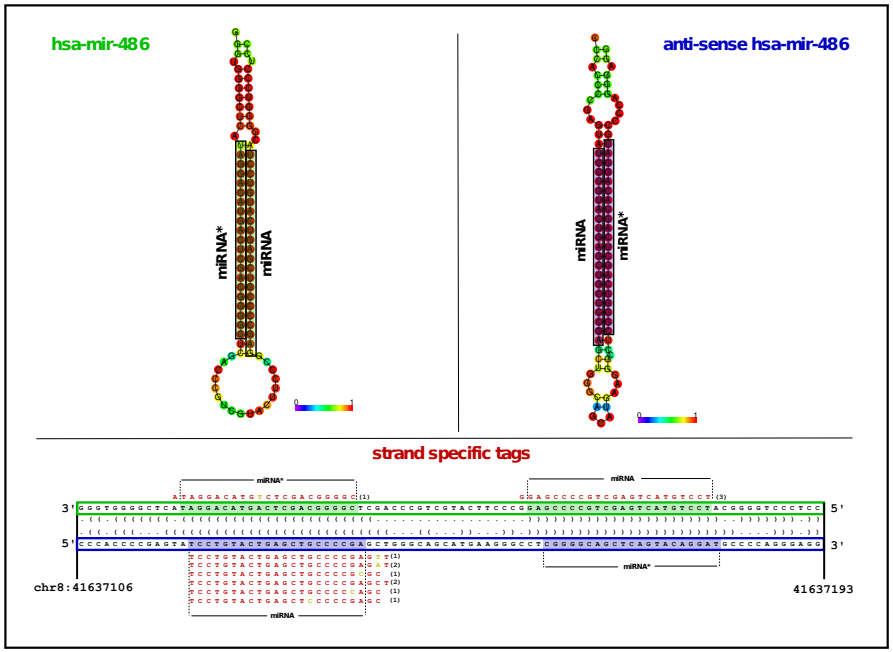
miRNA cluster

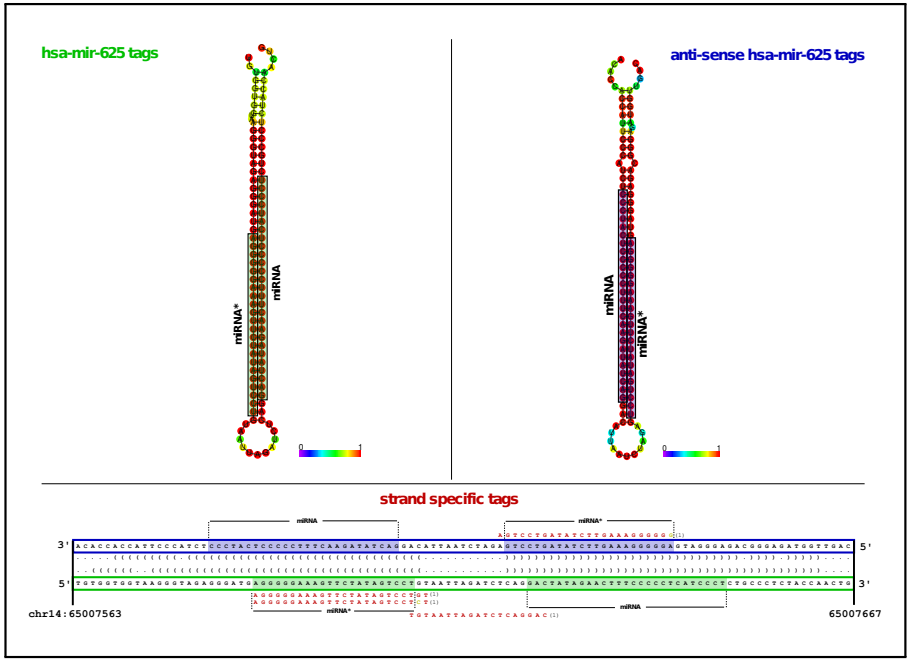
- microRNA-offset RNA (moRs) made distinct sub-cluster within miRNA cluster.
- 18 snoRNAs, 13 tRNAs and Six unannotated block groups fall within miRNA cluster.
- Five snoRNAs (ACA36b, ACA45, U27, U44 and HBI-100) have been reported in earlier studies to generate products with miRNA-like functions.
- Eight tRNAs (sharing four different anticodons) have been reported in literature.
  - tRNA Ala (AGC), tRNA Ser (AGA) - Lee et al., 2009
  - tRNA Lys (TTT), tRNA Gln (CTG) - Col et al., 2009

Results: one-sample comparison

miRNA cluster (contd..)

- Two unannotated block groups lie in an antisense direction to already annotated miRNAs (hsa-mir-486, hsa-mir-625).

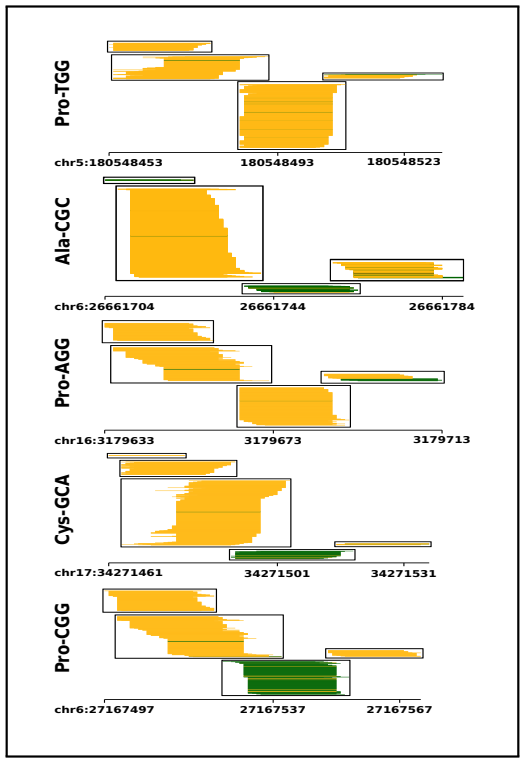




Results: one-sample comparison

tRNA cluster

- tRNAs having different anti-codons (TGG, CGC, GCA, CGG, AGG), but highly similar processing patterns.
- Six unannotated block groups were tRNA-derived pseudo genes.
- Two further loci correspond to deleted miRBase miRNAs (hsa-mir1974 and hsa-mir-1978).



Results: multi-sample comparison

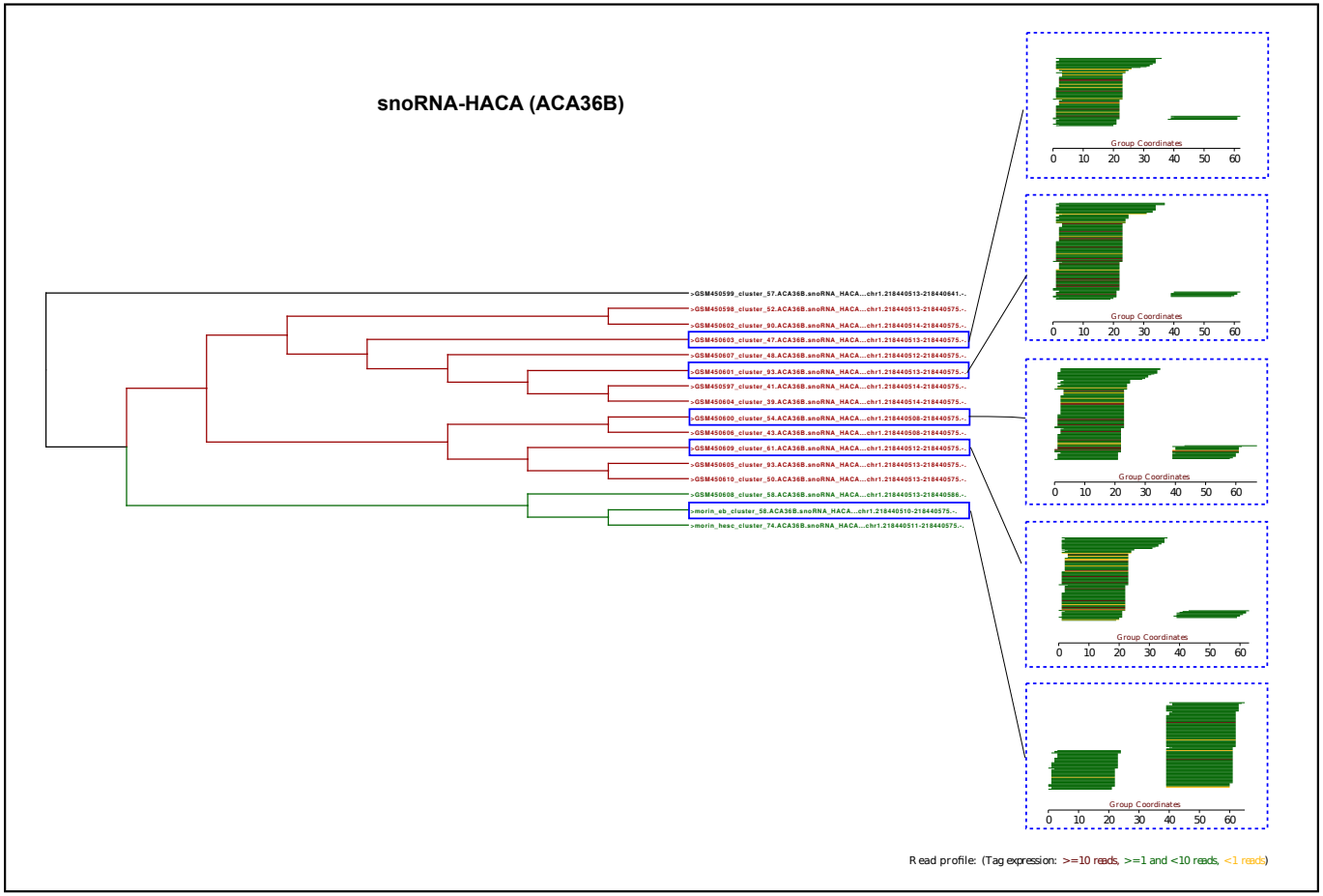
- retrieve read processing patterns conserved across multiple GEO samples.
- perform all vs all alignment followed by hierarchical clustering.
- study loci with different read processing pattern across samples.

Dataset:

Organism	**Tissue**	**#Samples**
Human	Brain	14
Human	Embryo	2

Results: multi-sample comparison

Distinct read processing pattern in samples from human brain and embryo.



Conclusions and Outlook

Conclusions

- Block entropy, count, size, distance and reads are specific for ncRNA classes (miRNA, snoRNA and tRNA).
- Comparison of read processing patterns is useful to identify
  - common processing for specific and/or across different ncRNA classes.
  - novel processing patterns.
  - difference in processing across samples.
- Since, deepBlockAlign does not require sequence information, it can complement well with the sequence-based ncRNA predictions tools.

Outlook

- Associate statistical significance (P- or E-values) with results.
- Application on larger and more diverse short-read dataset.
- A multiple-alignment version of program to align multiple block groups or read processing patterns.

Acknowledgements

- David Langenberger, University of Leipzig
- Steve Hoffmann, University of Leipzig
- Claus Ekstrøm, University of Copenhagen
- Peter Stadler, University of Leipzig
- Jan Gorodkin, University of Copenhagen
- All colleagues, University of Copenhagen
- LIFE, University of Copenhagen

– We have two Ph.D. positions at RTH –