Identification and Classification of ncRNAs using Deep Sequencing data

Sachin Pundhir

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

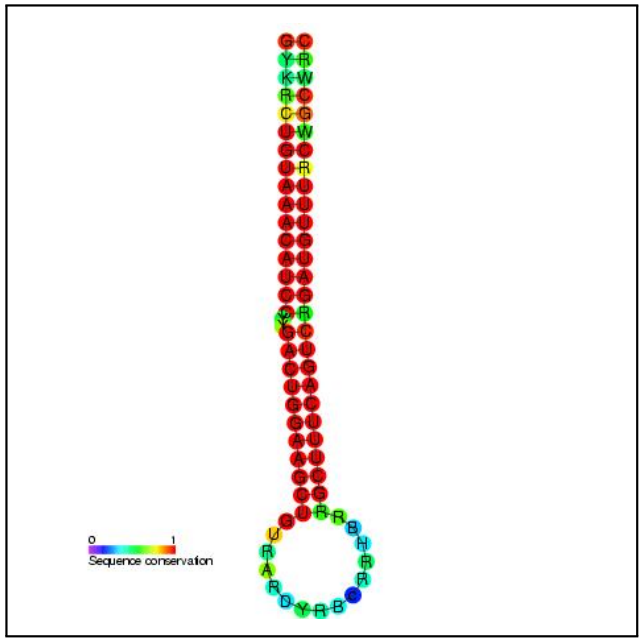
Feb 17, 2011

Non-coding RNA

Functional RNA molecules that are not translated into protein.

Notable examples are:

- tRNA: transfer RNA
- rRNA: ribosomal RNA
- piRNA: piwi-interacting RNA
- siRNA: small interfering RNA
- snRNA: small nuclear RNA
- snoRNA: small nucleolar RNA

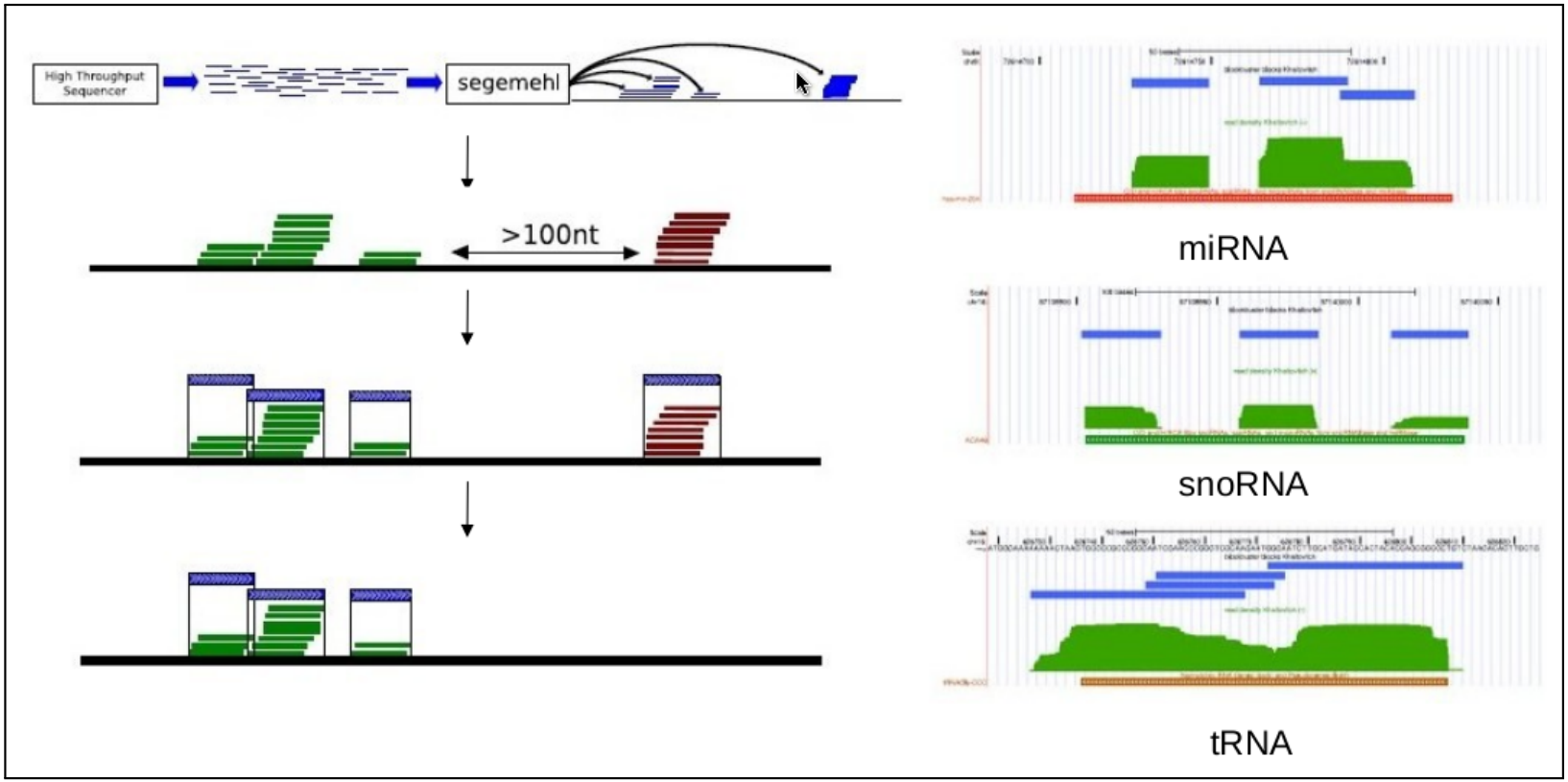


High-throughput sequencing

- Parallelize the sequencing process, producing millions of sequences at once (454, Solexa and SOLiD sequencing).
- Unlike microarray expression profiling, HTS can provide information about all RNA species present in a sample without any sequence information.
- Excellent tool for studying species where limited sequence information is available.
- Can aid in infering secondary structure and discovery of novel non-coding RNAs.

Read processing pattern

Recent studies (Jung et al. 2010, Langenberger et al. 2010) have shown a striking relationship between the mapped read patterns and the ncRNA classes.

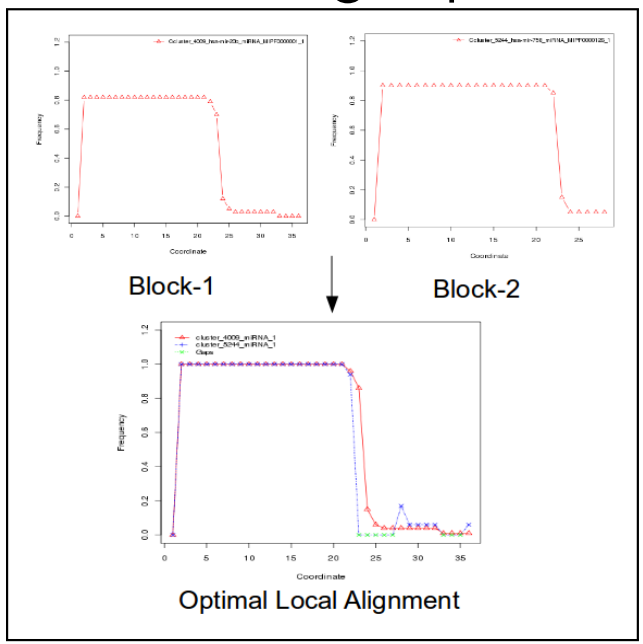


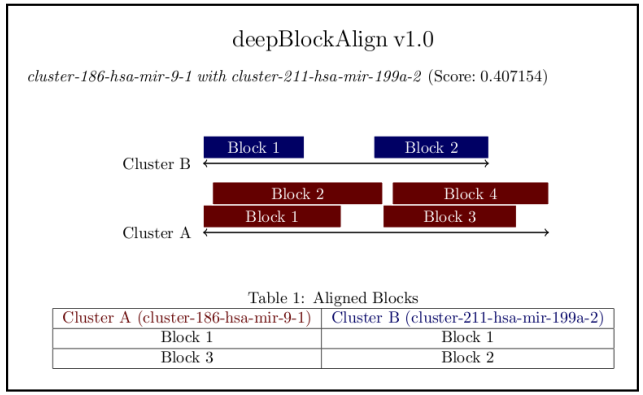
Alignment of block groups

Developed a set of programs for optimal alignment of block groups.

- blockAlign: Based on Smith-Waterman algorithm and determines optimal global alignment between all read blocks from two group of blocks.

- deepBlockAlign: Based on Sankoff algorithm and optimally align two group of blocks by taking into account
  a) block alignment scores; and
  b) distance between blocks.





Dataset

The analysis was done on two deep-sequencing datasets:

- Illumina sequencing data from Human postnatal brain in prefrontal cortex (GSE18012).

	Total	miRNA	snoRNA	tRNA	Unannotated
All	8272	318	196	298	7460
Block (1)	6655	116	54	86	6399
Block (>1)	1617	202	142	212	1061

- Illumina sequencing data from Human embryoid body cells.

	Total	miRNA	snoRNA	tRNA	Unannotated
All	2091	368	132	356	1235
Block (1)	1249	145	52	43	1009
Block (>1)	842	223	80	313	226

Block analysis

All the blocks within 8272 block groups from Human brain data were analyzed for Entropy (randomness in read processing pattern).

$$I = -\sum_{i=1}^{A} q_i \; log_2(q_i) \tag{1}$$

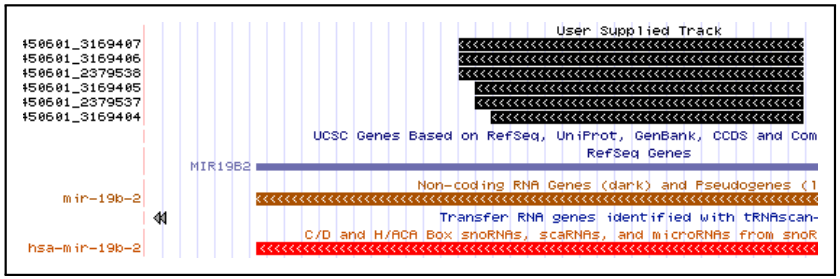
where;

$$q_i = \frac{n_i}{N} \; ; \; \sum_{i=1}^{A} n_i = N \tag{2}$$

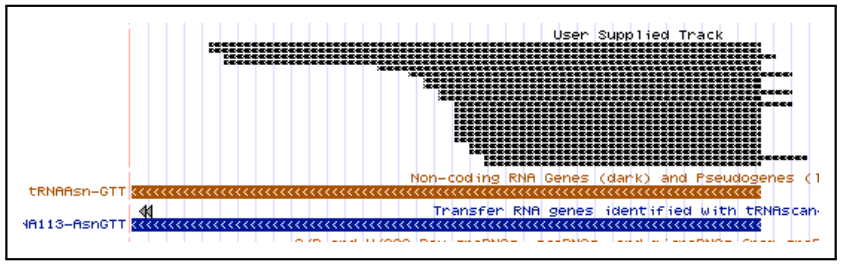
I = Entropy in read processing pattern,
n_i = Total start tags at position i,
N = Total tags in a block group,
A = Total positions in a block group having start tags.

Block analysis

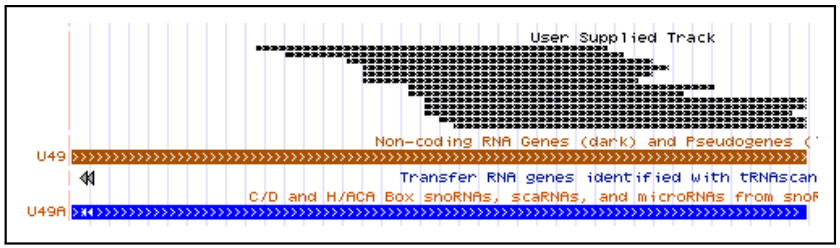
miRNA (Entropy=0)



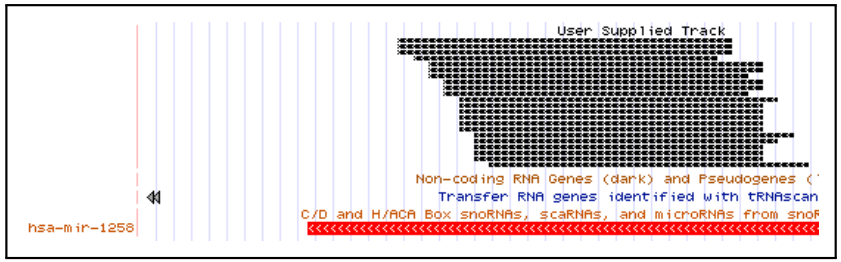
tRNA (Entropy=1.2)



snoRNA (Entropy=3)

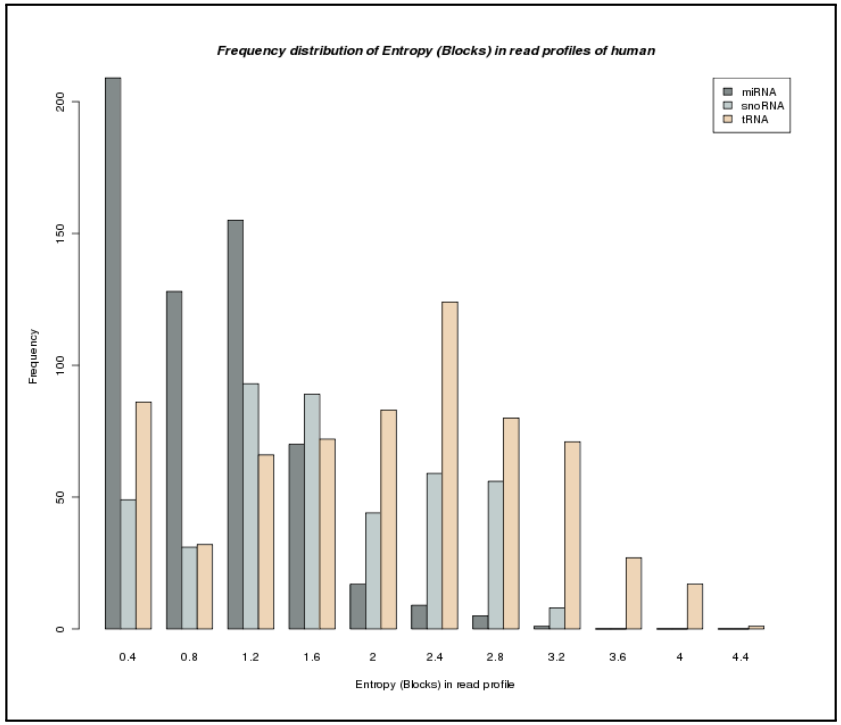


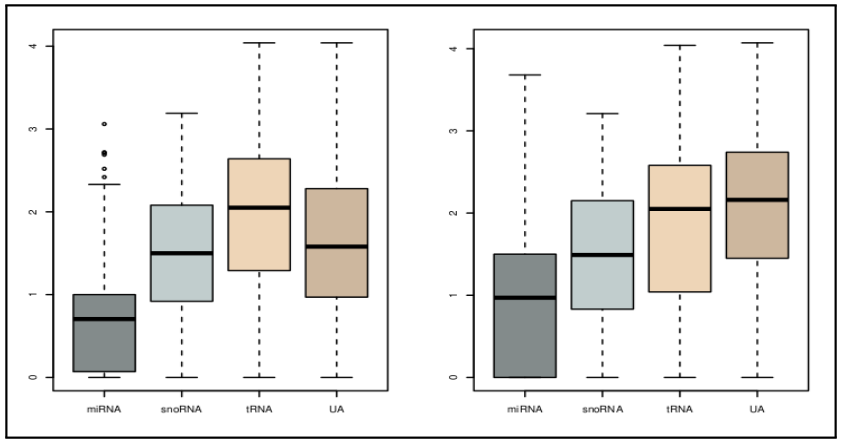
miRNA (Entropy=2.4)



Block analysis

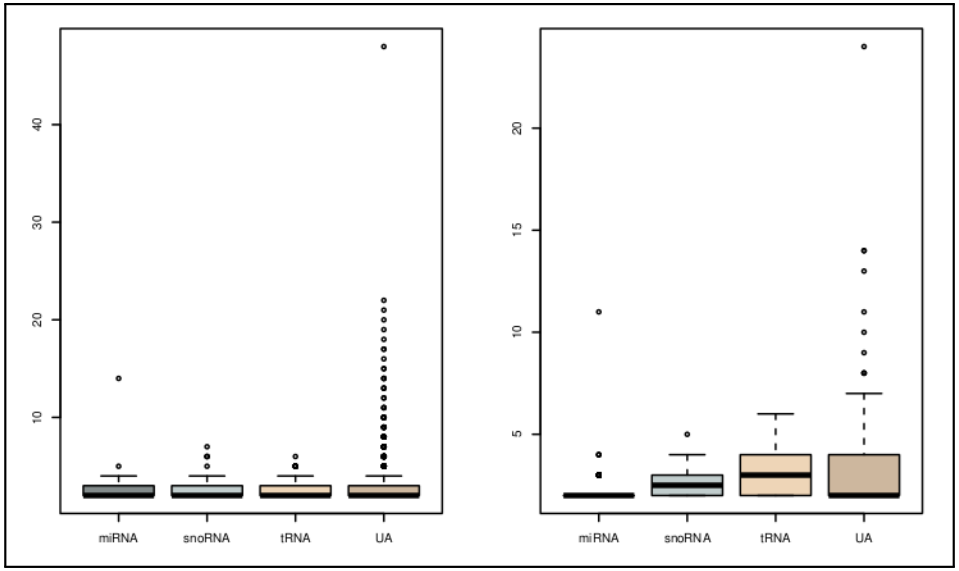
- Most of the blocks in miRNA have well defined processing pattern, Entropy ranging from 0 to 1.2.
- Read processing pattern for snoRNA and tRNA is highly variable with quite a few also having low entropy.



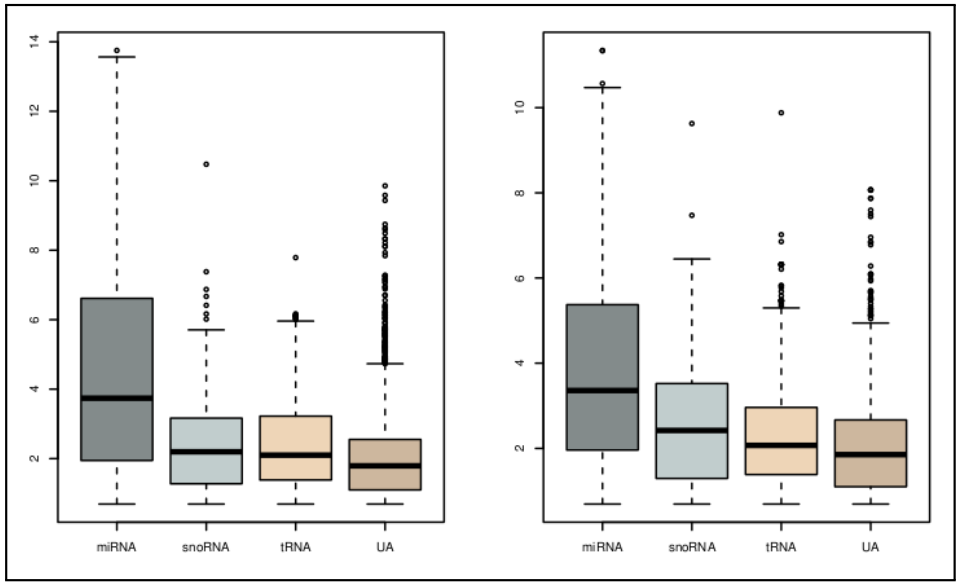


Block analysis

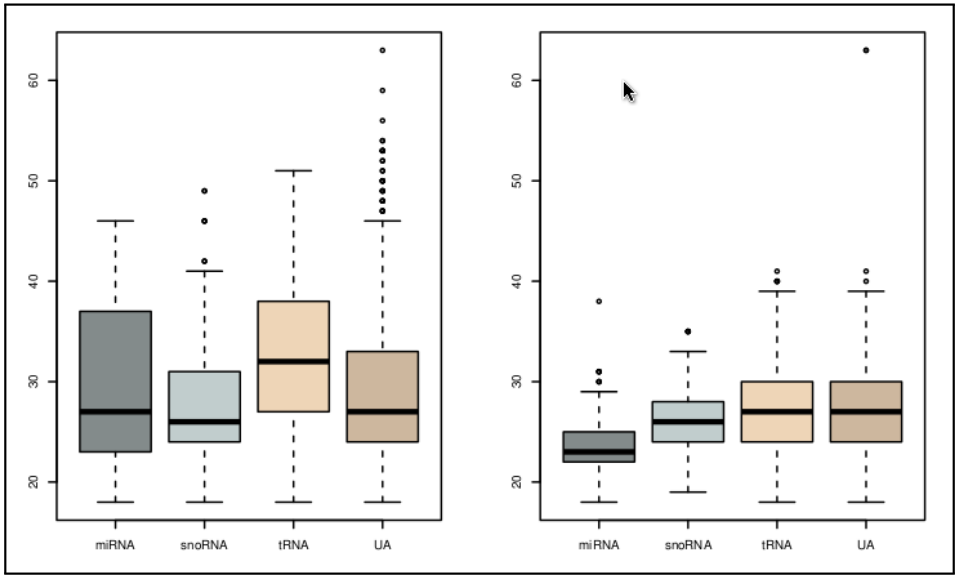
Block Count



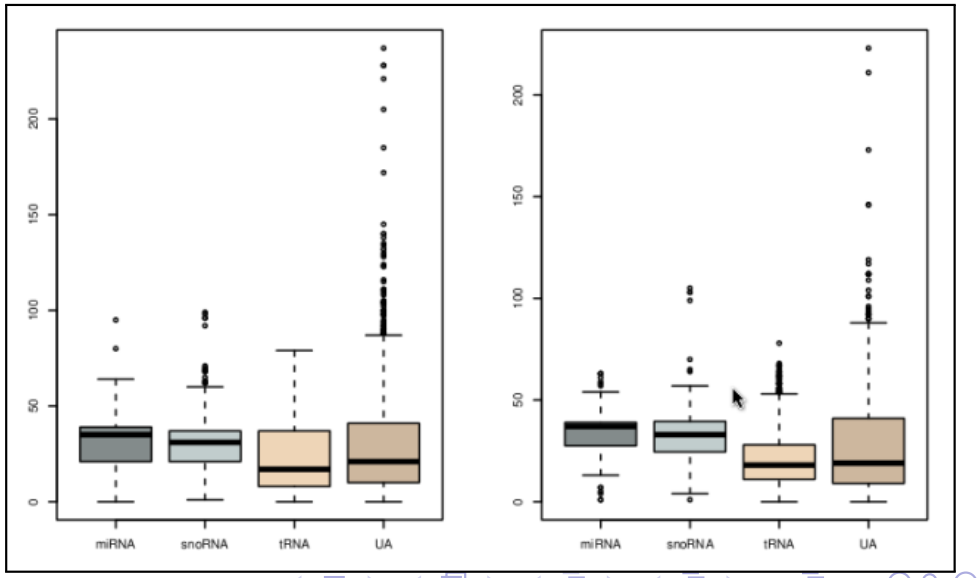
Total Reads in Block



Block Size

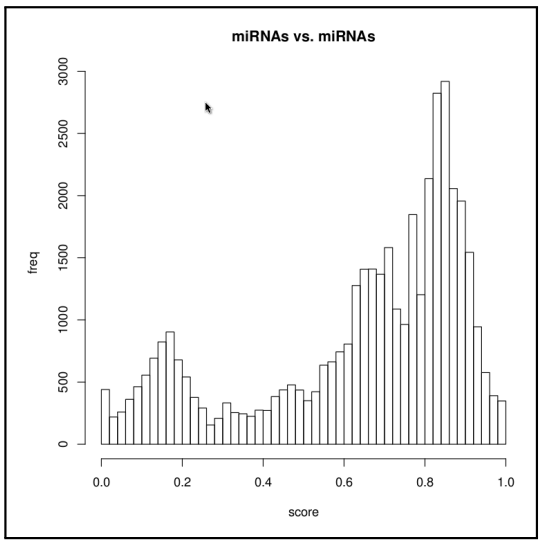


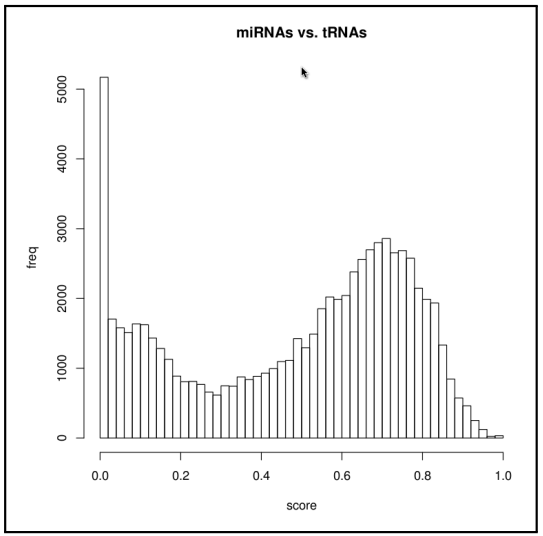
Distance between Blocks

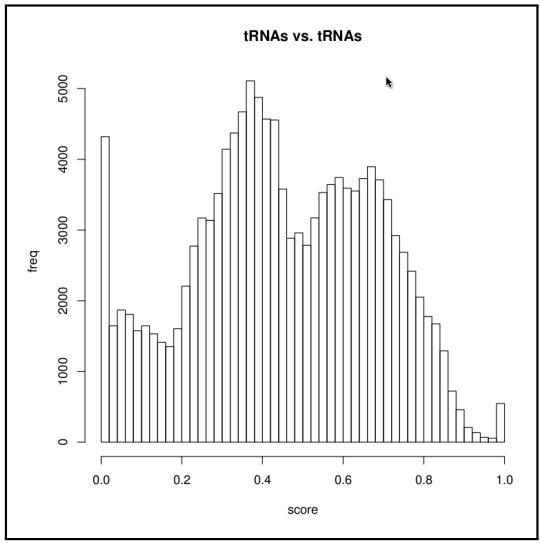


deepBlockAlign: performance measure analysis

- A distinct score distribution was not observed between miRNA Vs miRNA and miRNA Vs tRNA alignment.
- If observed, we can associate a statistical significance term like E-value/P-value to our prediction results.







deepBlockAlign: performance measure analysis

- In view of the absence of a statistical significance term like E-value/P-value, performance of deepBlockAlign was measured by using rank measure.
- All the 555 block groups having >1 block from Human brain dataset were aligned with each other using deepBlockAlign.
- Next, a block group showing best hit score with a block group from the same class was considered a 'True Positive' else 'False Negative'.

Table: deepBlockAlign results on Human brain dataset

	miRNA	snoRNA	tRNA	Total	PPV	R_K
miRNA	172	17	12	201	0.67	
snoRNA	44	77	21	142	0.71	0.6
tRNA	41	14	157	212	0.83	

deepBlockAlign: performance measure analysis

Table: deepBlockAlign results on Human embryo dataset

	miRNA	snoRNA	tRNA	Total	PPV	R_K
miRNA	208	6	8	222	0.84	
snoRNA	21	39	19	79	0.67	0.76
tRNA	19	13	281	313	0.91	

- To verify that the results for tRNA are not an artifact due to the presence of multiple copies of same tRNA in the genome, we compared all tRNA block groups from brain dataset with the embryo dataset.

	tRNA (brain)	tRNA (embryo)	Total
tRNA (brain)	157	142	212
tRNA (embryo)	149	281	313

deepBlockAlign: performance measure analysis

- Four out of six most frequent amino acid residue label in 149 and 142 True Positive tRNAs were common.

tRNA (Amino acid)	Frequency (brain)	Frequency (embryo)
Lysine	20	20
Arginine	16	12
Glycine	14	12
Leucine	12	15

cluster-4379-chr1-Lys-TTT with cluster-1138-chr1-Lys-TTT (Score: 0.739091)

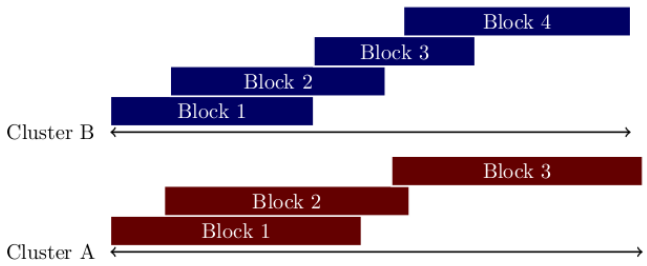


Table 1: Aligned Blocks

Cluster A (cluster-4379-chr1-Lys-TTT)	Cluster B (cluster-1138-chr1-Lys-TTT)
Block 1	Block 1
Block 2	Block 2
Block 3	Block 4

cluster-1385-chr16-Lys-CTT with cluster-3131-chr6-Gln-CTG (Score: 0.735154)

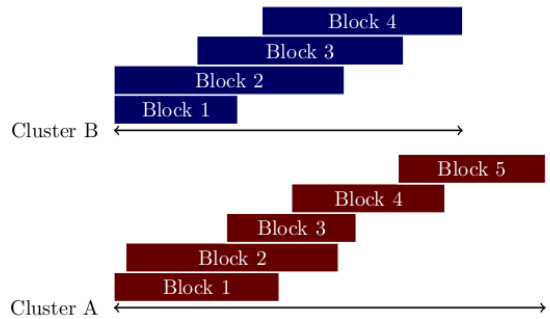
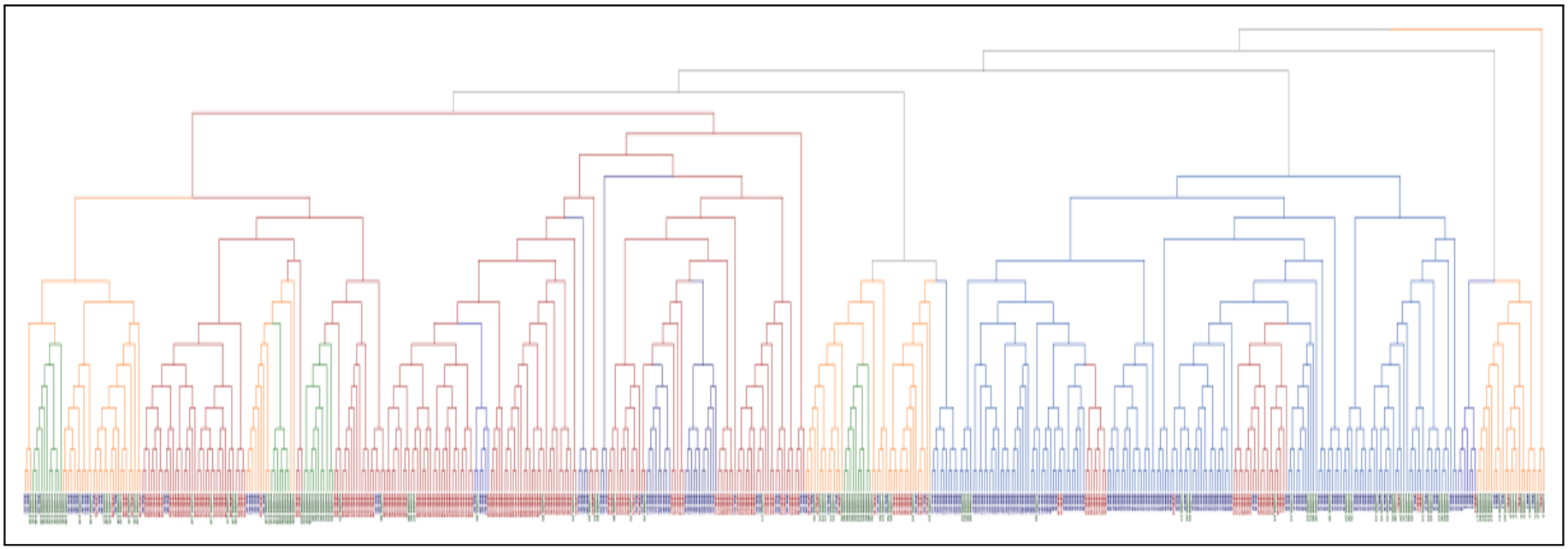


Table 1: Aligned Blocks

Cluster A (cluster-1385-chr16-Lys-CTT)	Cluster B (cluster-3131-chr6-Gln-CTG)
Block 1	Block 1
Block 2	Block 2
Block 3	Block 3
Block 4	Block 4

deepBlockAlign: performance measure analysis

- On hierarchial clustering of ncRNAs on the basis of their alignment score, we observed distinct clusters for miRNA, snoRNA and tRNA.



miRNA, snoRNA, tRNA

Conclusions

- Block parameters like entropy, count, size, distance and reads are specific for ncRNA classes (miRNA, snoRNA and tRNA) with some overlaps.
- A distinct scoring distribution between ncRNA may be possible by further optimizing the parameters.
- Performance evaluation of deepBlockAlign using rank measure analysis seems promising.
- On hierarchial clustering, distinct clusters for miRNA and snoRNA and tRNA were observed.

Conclusions

- deepBlockAlign is meant to align two block groups with similar read processing pattern irrespective of their sequence annotation, this is significant in the light of recent findings which suggest various tRNAs having miRNA like processing pattern.
- Since, deepBlockAlign does not require sequence information for predictions, it can complement well with the other ncRNA predictions tools that are based on sequence information.

Acknowledgements

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