In-silico identification and classification of non-coding RNA using high-throughput sequences.

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

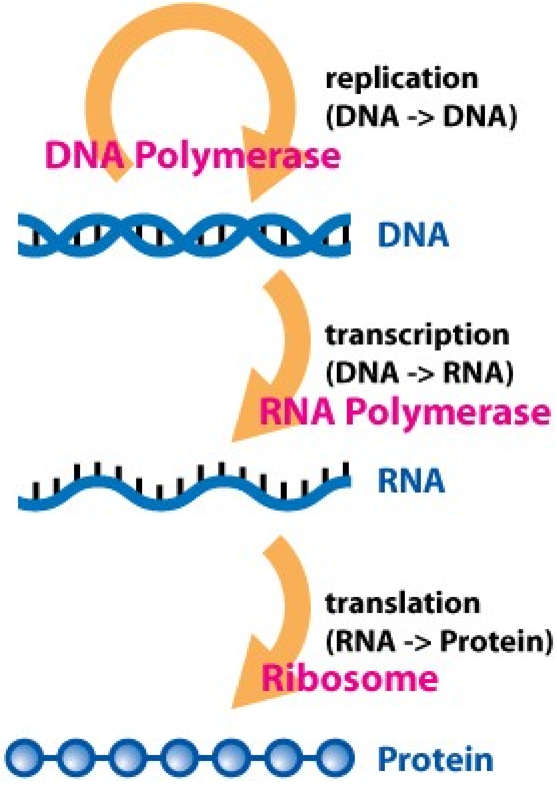
Central Dogma

The Central Dogma of Molecular Biology deals with the detailed residue-by-residue transfer of sequential information.

This information transfer occurs between three major biopolymers named DNA, RNA and Protein.

In Humans, 90% of the human genome is transcribed into RNA but only 1.4% of the genome encode for proteins.

Rest of the RNA is termed as non-coding RNA (ncRNA).

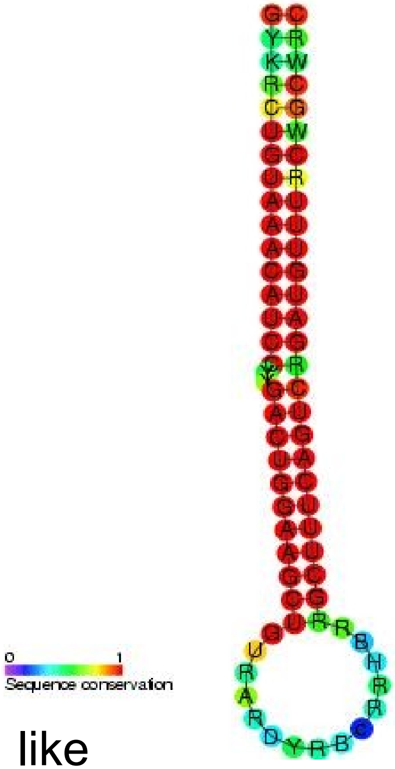


Non-coding RNAs

Functional RNA molecule that is not translated into protein.

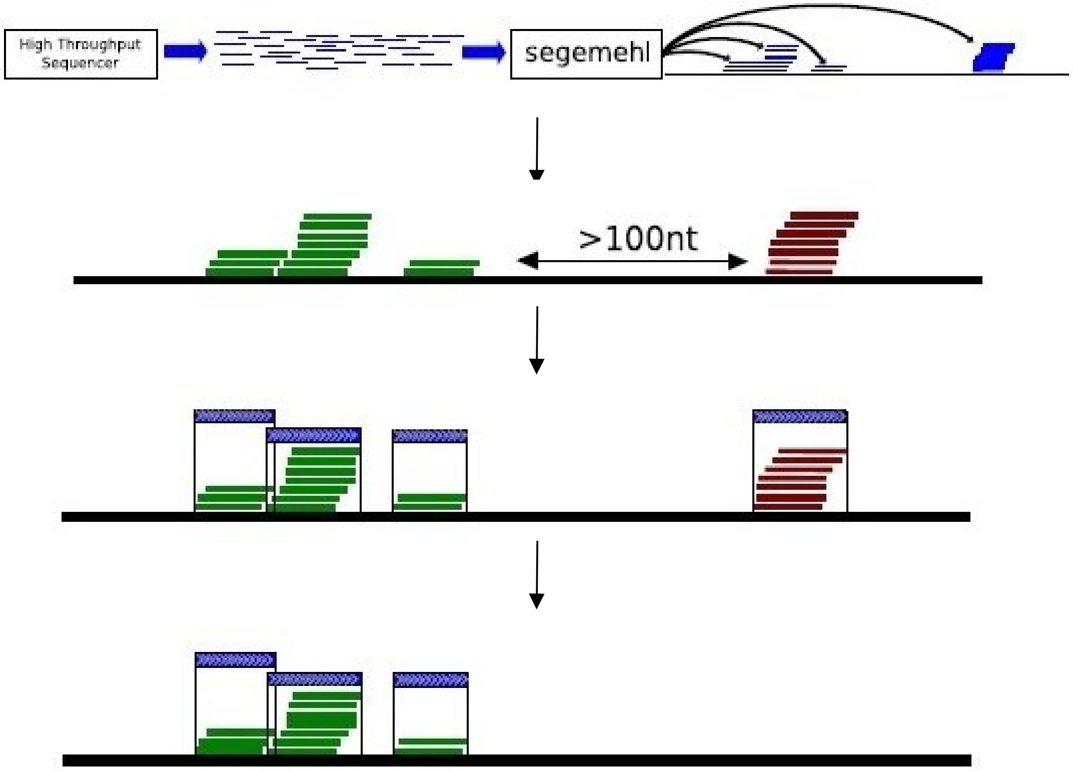
Notable examples of ncRNAs are:

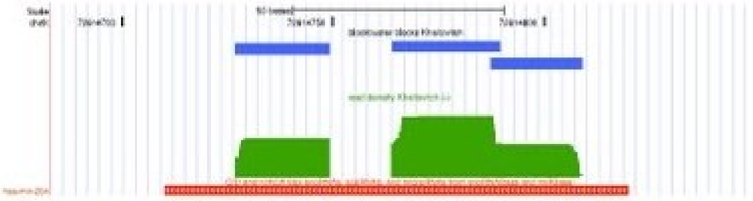
- tRNA: transfer RNA
- rRNA: ribosomal RNA
- piRNA: Piwi-interacting RNA
- siRNA: small interfering RNA
- snRNA: small nuclear RNA
- snoRNA: small nucleolar RNA
  - stRNA: small temporal RNA



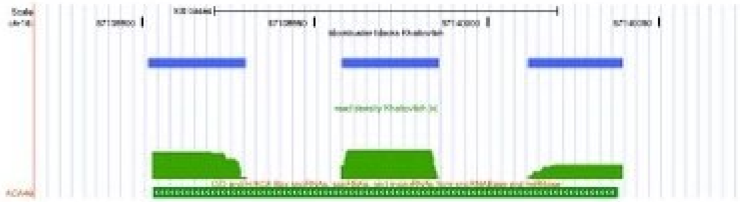
Recent studuies have shown that majority of ncRNA like miRNAs are actively involved in regulation.

Background

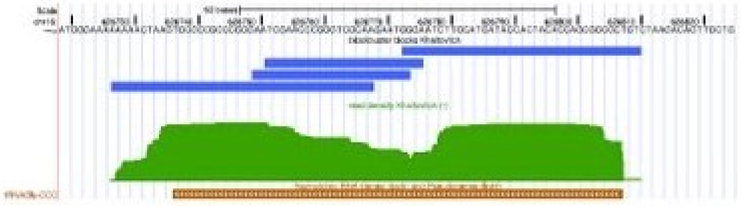




miRNA



snoRNA



tRNA

Langenberger et. al, 2010 . Identification and Classification of small RNAs in Transcriptome Sequece Data

deepAlign

A C Program to optimally align two read processing patterns without taking the expression level explicitly into account.

Use simple dynamic programming algorithm, in Needleman-Wünsch and Smith-Waterman style, by simply processing the read count at each position.

More specifically, each position in the profile has two parameters

- Start read count
- Total read count

Minimizing these parameters between two read profiles yield an alignment.

Algorithm:

Firstly, for each read cluster, total number of start reads and total reads are normalized to one, seperately.

Secondly, for two profile *X* and *Y*, each position is represented by x_i and y_j, repectively.

Where,

x_i = difference of normalized total reads and start reads at position *i* in profile *X*.
y_j = difference of normalized total reads and start reads at position *j* in profile *Y*.

Next, maximization score for comparing *X* and *Y* is calculated as:

$$s_{ij} = \max \begin{cases} 0 \\ s_{i(j-1)} + c \\ s_{(i-1)j} + c \\ s_{(i-1)(j-1)} + \Psi_\delta(x_i, y_j) \end{cases} ,$$

Where c is the gap penalty and

$$\Psi_\delta(x_i, y_j) = \begin{cases} 1 - \frac{|x_i - y_j|}{\max\{x_i, y_j\}} & \text{if } |x_i - y_j| < \delta \text{ and } n_i m_j > \theta \\ \frac{|x_i - y_j|}{\max\{x_i, y_j\}} & \text{Otherwise} \end{cases} ,$$

Where;

δ is the allowed threshold for the difference between x_i and $y_{j.}$ **set to 0.1**

n_i *and* m_j are total number of reads in profile X and Y at positions *i* and *j*.

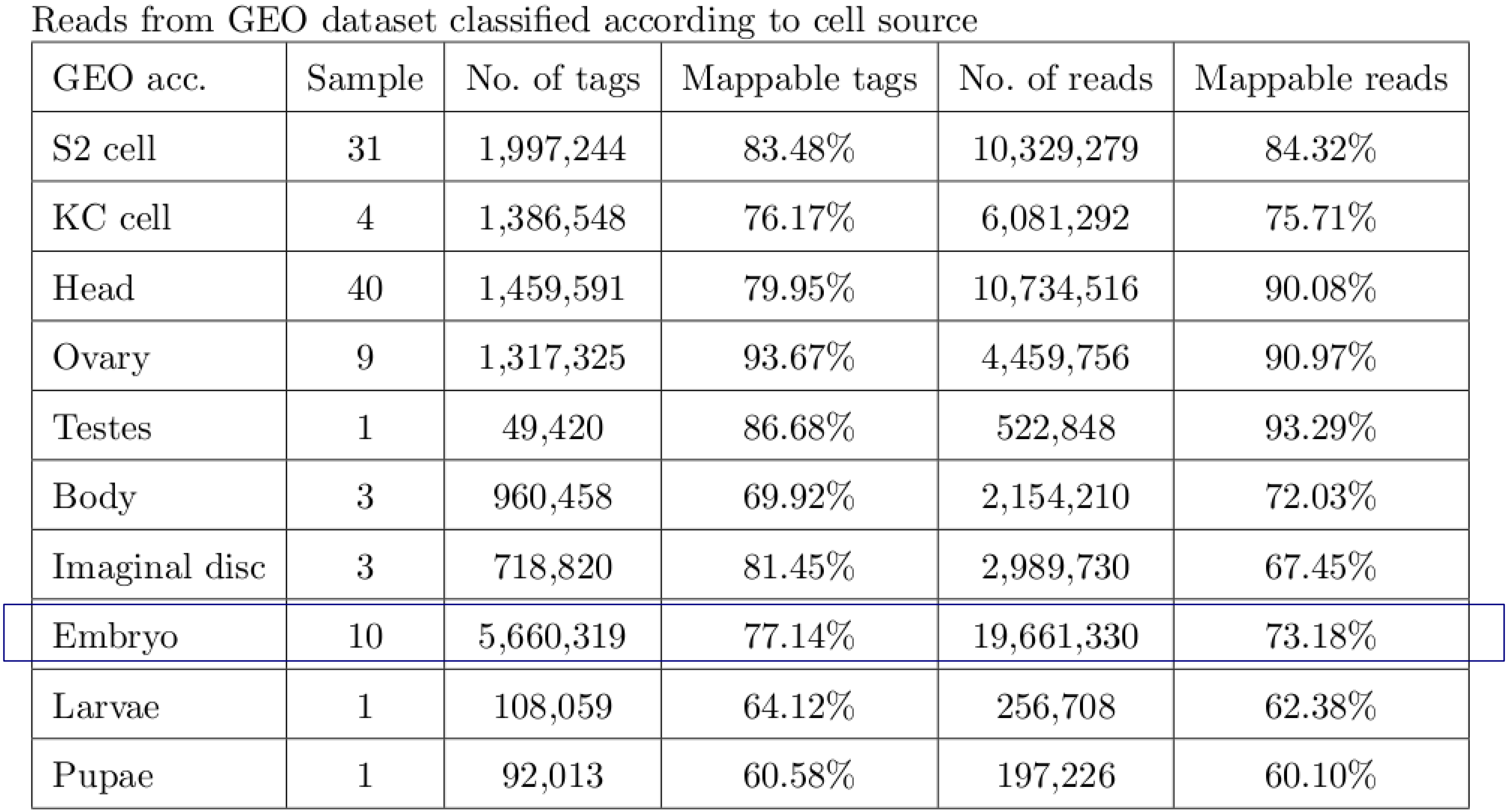
Match score for aligning x_i and y_j is the fraction similarity between the read profile at position *i* and *j.*

Likewise, mismatch score is the fraction dissimilarity between read profile at position *i* and *j*.

Mapping of read data

Downloaded *D. Melanogaster* genome from FlyBase (Apr. 2006).

Retrieved 13,335,481 sequence tags comprising of 55,894,809 reads from 12 Gene Expression Omnibus (GEO) datasets derived from 90 experiments performed on drosophila cell lines and tissues.



Reads from GEO dataset classified according to cell source

GEO acc.	Sample	No. of tags	Mappable tags	No. of reads	Mappable reads
S2 cell	31	1,997,244	83.48%	10,329,279	84.32%
KC cell	4	1,386,548	76.17%	6,081,292	75.71%
Head	40	1,459,591	79.95%	10,734,516	90.08%
Ovary	9	1,317,325	93.67%	4,459,756	90.97%
Testes	1	49,420	86.68%	522,848	93.29%
Body	3	960,458	69.92%	2,154,210	72.03%
Imaginal disc	3	718,820	81.45%	2,989,730	67.45%
Embryo	10	5,660,319	77.14%	19,661,330	73.18%
Larvae	1	108,059	64.12%	256,708	62.38%
Pupae	1	92,013	60.58%	197,226	60.10%

Next, tags mapped to fly genome using *segemehl* (Hoffman et al., 2009).

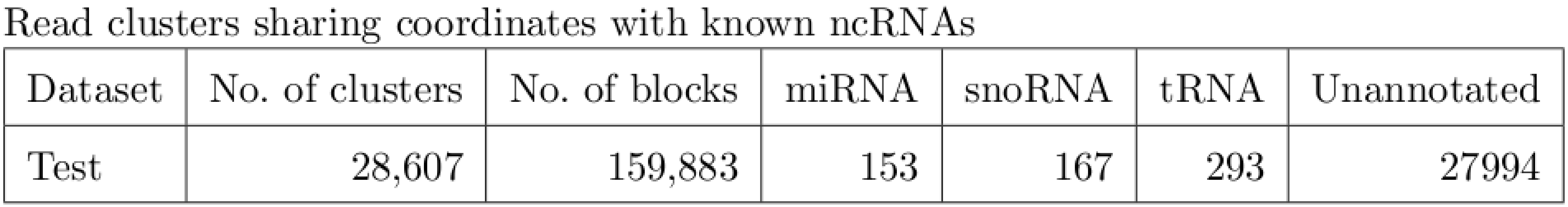
Conditions during mapping:

- Mapping accuracy >80%
- Read count for tags mapping at multiple loci were equally weighted across all their loci.
- Tags mapping at number of loci greater than their number of reads were discarded.

Clustering using *blockbuster* (Langenberger et. al. 2009):

- Defines clusters (seperated by <30nt) and blocks.
- Clusters with <10 reads and blocks with <2 reads were discarded.
- Total, 28607 read clusters and 159,883 blocks were identified.

Next, the start and end coordinates of known miRNA (176 loci), snoRNA (249 loci) and tRNA (295 loci) retrieved from miRBase, FlyBase and gtRNAdb, respectively.



Read clusters sharing coordinates with known ncRNAs

Dataset	No. of clusters	No. of blocks	miRNA	snoRNA	tRNA	Unannotated
Test	28,607	159,883	153	167	293	27994

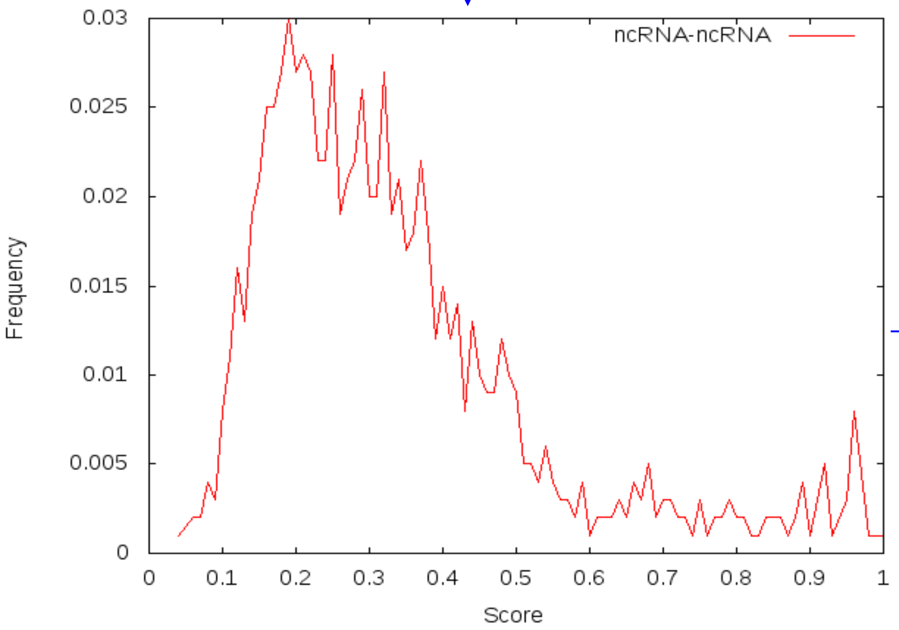
Performance evaluation

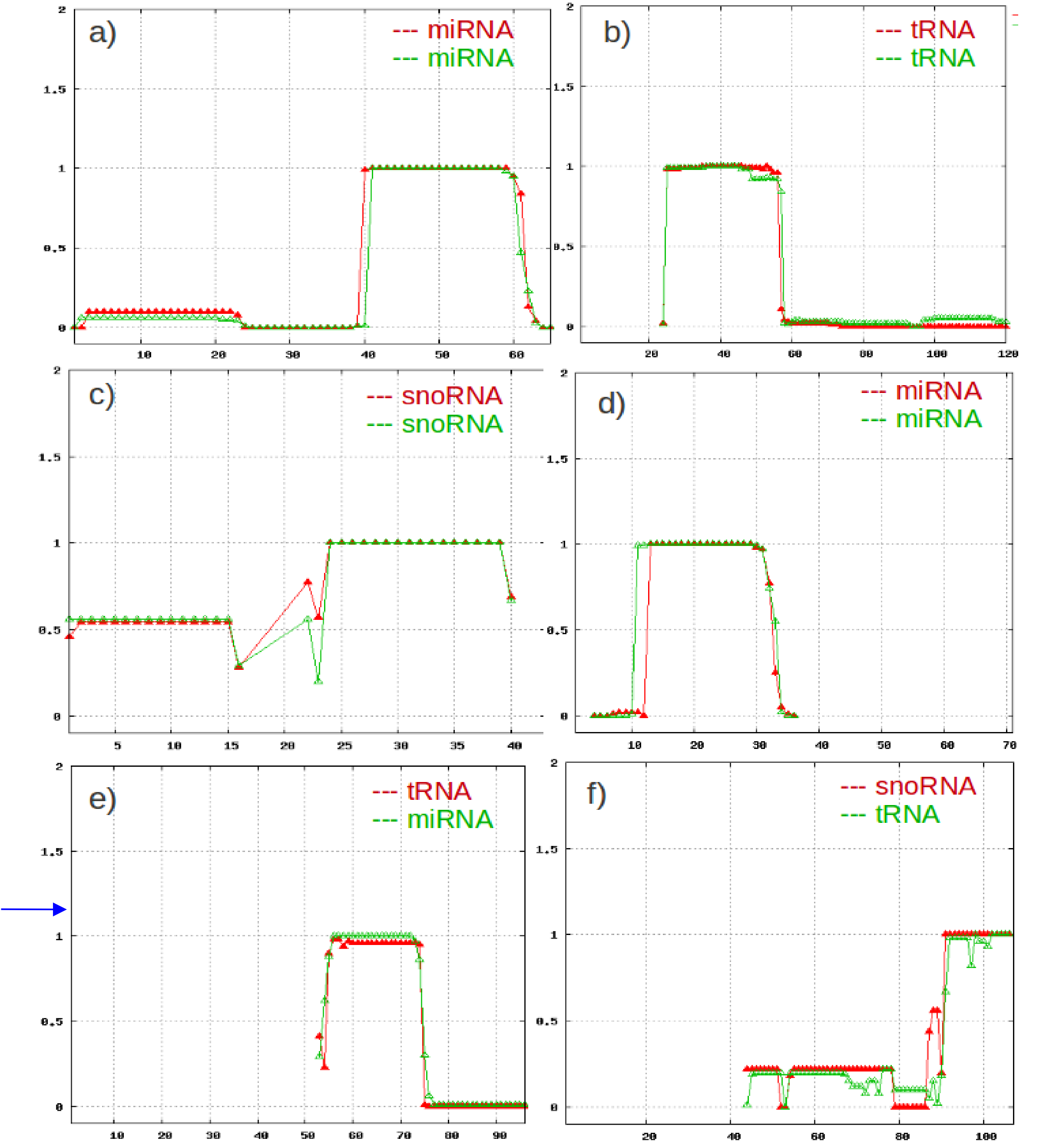
A) Evaluation of scoring distribution

For all ncRNA clusters, an All Vs All alignment was carried out using deepAlign.

The yielded scoring distribution was used to classify alignments as

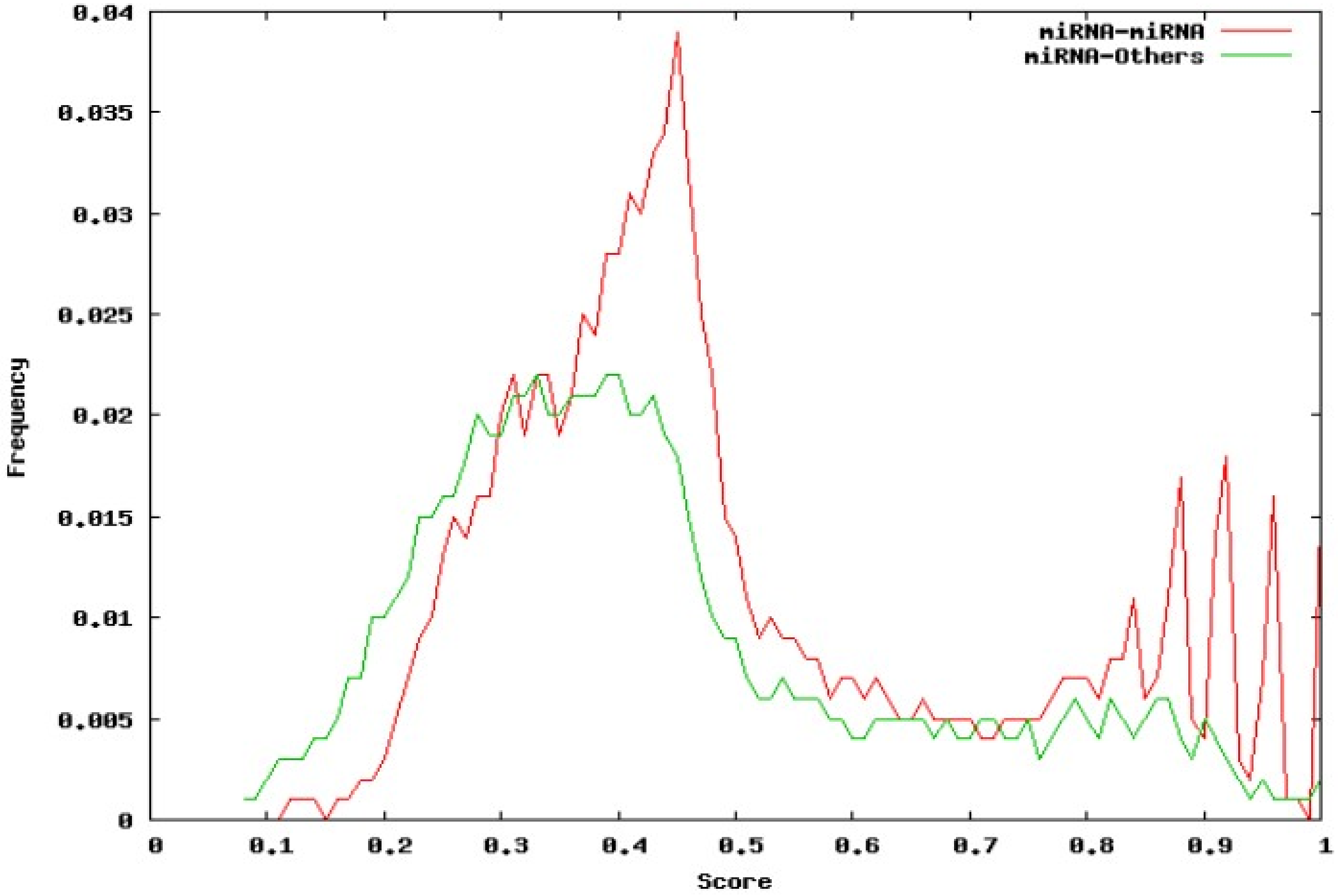
- Good (Score, >0.65 to 1)
- Average (Score, >0.35 to 0.65)
- Poor (Score, < 0.35)





B) Evaluate comparative scoring distribution between miRNA-miRNA and miRNA-others

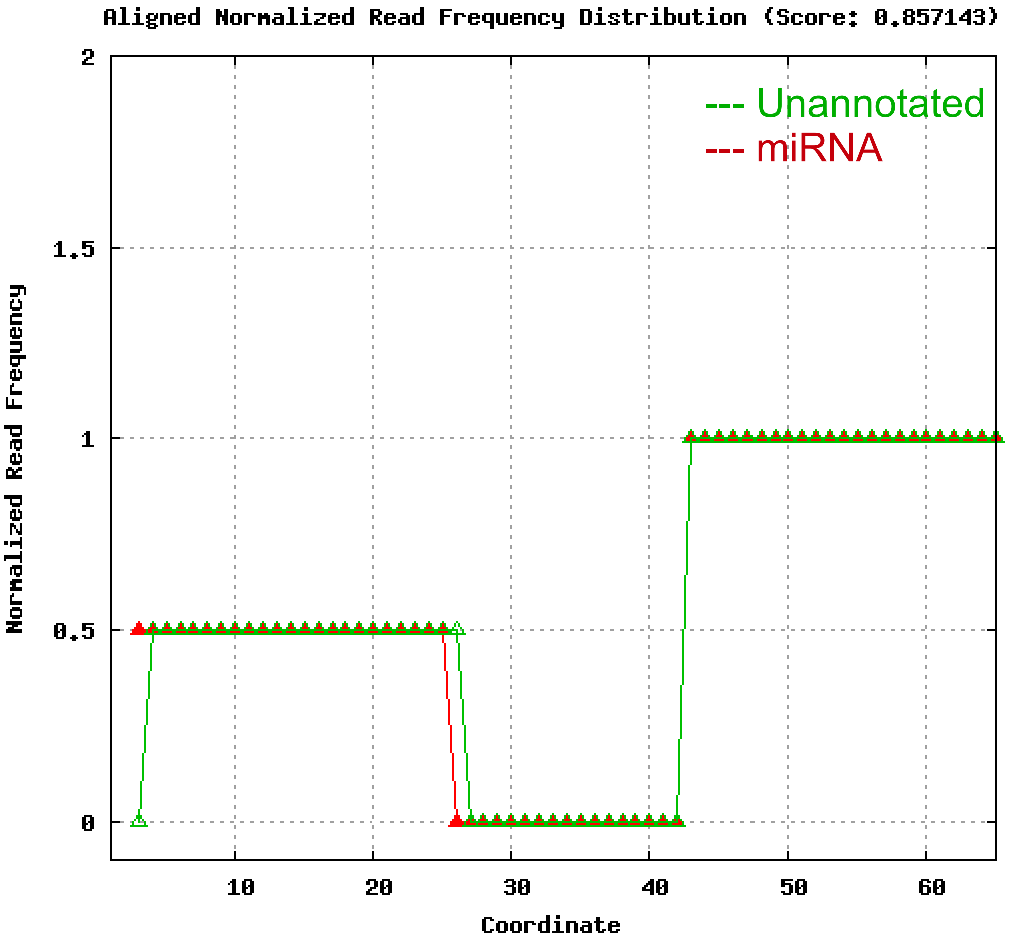
miRNA profile clusters were aligned with themselves and scoring distribution thus obtained was compared with the alignment scoring distribution of miRNA with other ncRNAs

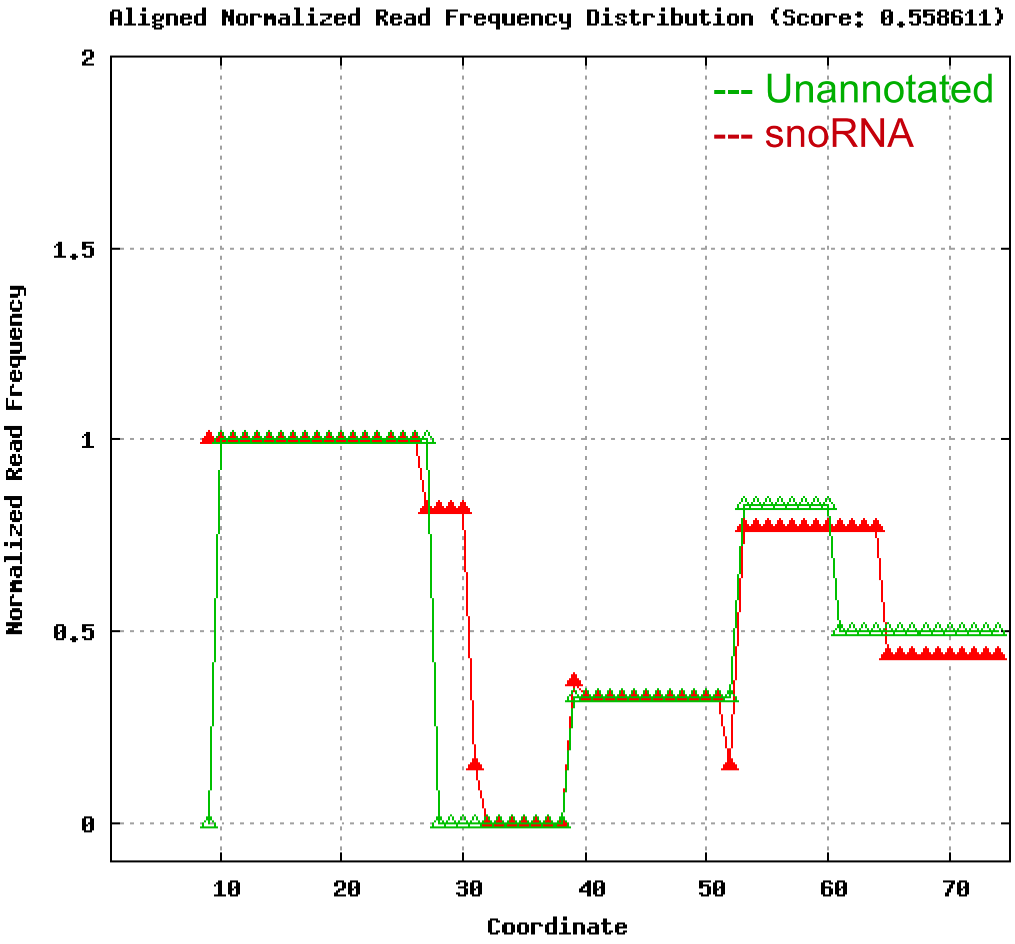


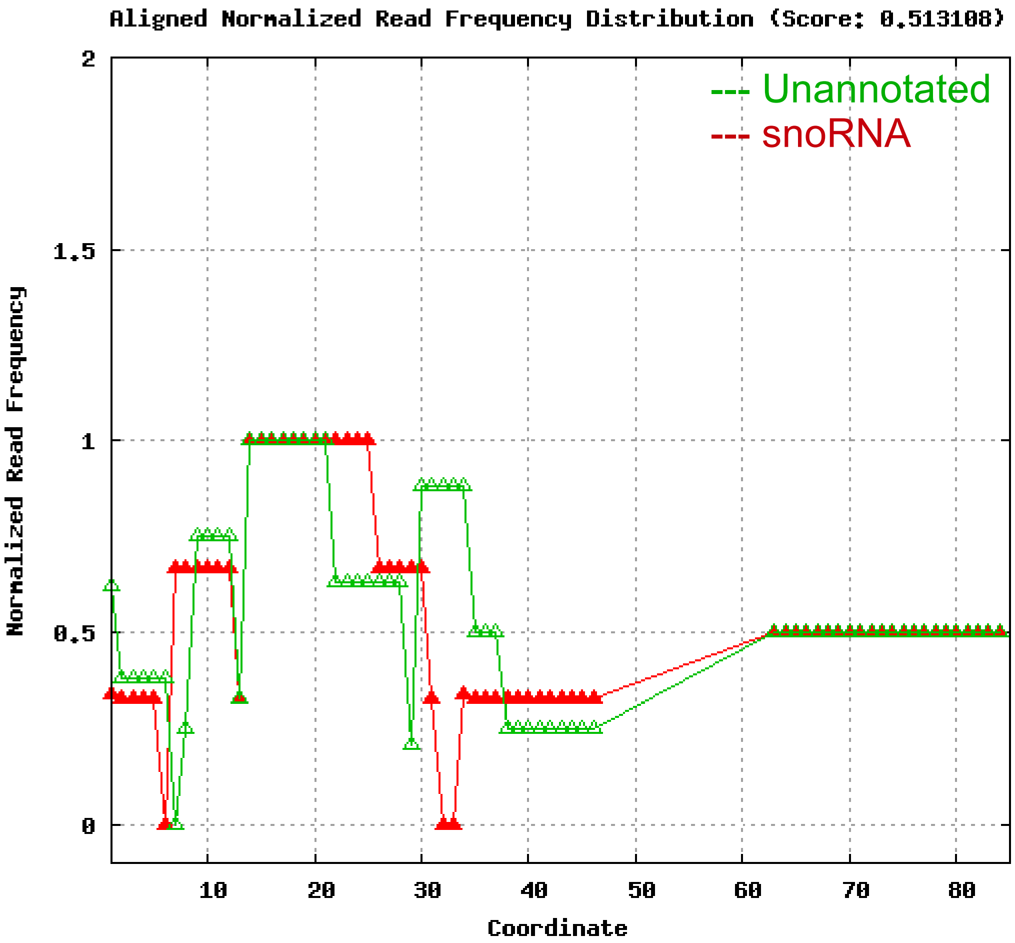
C) Evaluate performance across whole fly genome

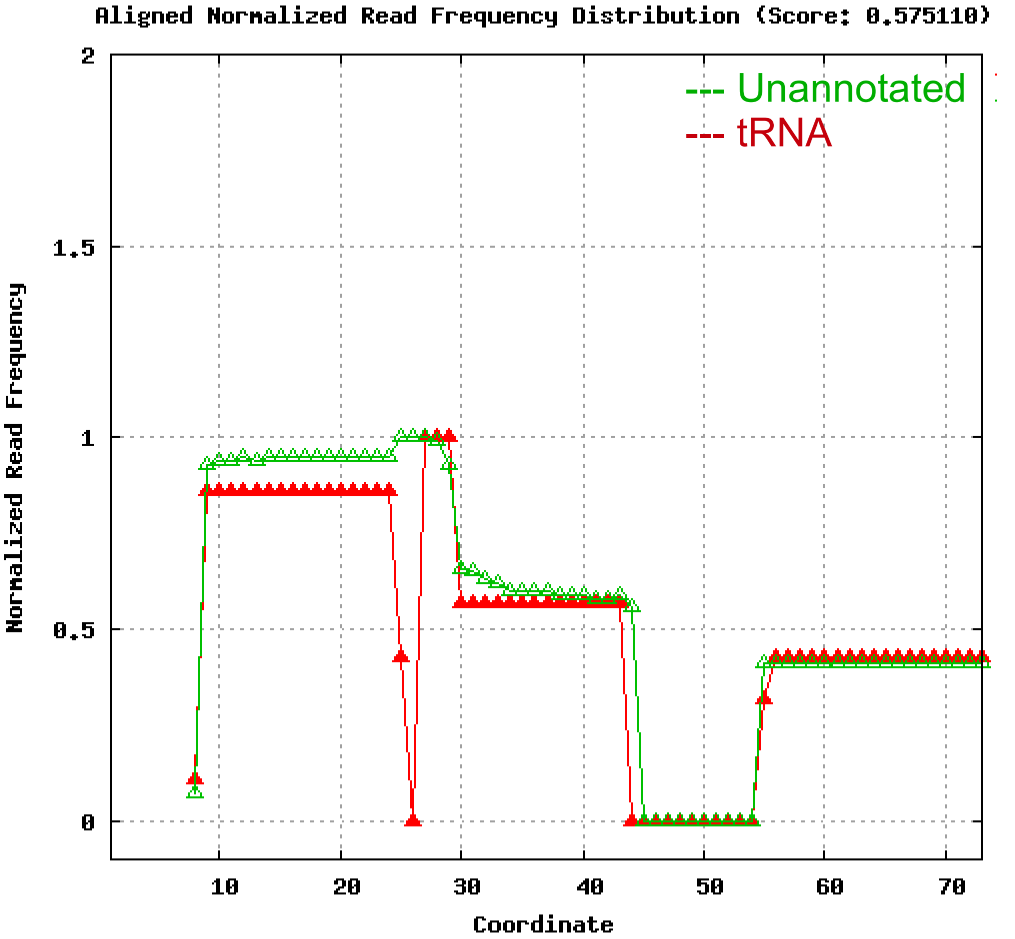
All the 28,607 read clusters were aligned with themselves and significanlty scoring alignments were studied for processing patterns.

Many of the unannotated read clusters aligned with annotated clusters with seemingly similar processing patterns.






Aligned Normalized Read Frequency Distribution (Score: 0.513108)
--- Unannotated
--- snoRNA
Normalized Read Frequency
0
0.5
1
1.5
2
10
20
30
40
50
60
70
80
Coordinate


Aligned Normalized Read Frequency Distribution (Score: 0.575110)
--- Unannotated
--- tRNA
Normalized Read Frequency
0
0.5
1
1.5
2
10
20
30
40
50
60
70
Coordinate

deepBlockAlign

The algorithm for *deepAlign* is ideal for identifying similar read processing patterns across a genome irrespective of the ncRNA annotation.

Since, recent studies have shown that ncRNA like tRNA or snoRNA may have read processing patterns like miRNA, *deepAlign* algorithm can not be applied for ncRNA classification with high accuracy.

deepBlockAlign is based on a novel algorithm that may be used in identification and classification of ncRNAs.

Follows a two step procedure:

- First, an alignment score is computed for all the read blocks in a cluster by comparing them with each other (*deepAlign*).

- Secondly, the block scores are employed to derive an optimal alignment between the two read clusters using *Sankoff* algorithm.

Acknowledgements

Jan Gorodkin, University of Copenhagen
Laurence Croft, University of Copenhagen
Anne Wenzel, University of Copenhagen
Claus Ekstrøm, University of Copenhagen
David Langenberger, University of Leipzig
Steve Hoffmann, University of Leipzig
Peter Stadler, University of Leipzig

Two Ph.D. Positions open, to be filled ASAP (www.rth.dk).