BlockClust: efficient clustering and classification of non-coding RNAs from short read RNA-Seq profiles

Pavankumar Videm[1], Dominic Rose[1,2], Fabrizio Costa[1], Rolf Backofen[1,3-5]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany.
[2]Munich Leukemia Laboratory (MLL), Munich, Germany.
[3]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Germany.
[4]Centre for Non-coding RNA in Technology and Health, Bagsvaerd, Denmark.
[5]Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Germany.

Introduction

Sequence and secondary structure analysis can be used to assign putative functions to non-coding RNAs (ncRNAs).
However sequence information is changed by post-transcriptional modifications [1] and secondary structure is only a proxy for the true 3D conformation of the RNA polymer.
Instead we can use the pattern of processing that can be observed through the traces left in small RNA-seq reads data.

We propose to encode expression profiles in discrete structures, which can be processed using fast graph-kernel techniques.
We developed BlockClust [2] which allows both clustering and classification of small ncRNA transcripts with similar processing patterns.

Methods

Given the mapped reads we use the blockbuster tool [3] to identify consecutive reads called blocks and adjacent blocks called blockgroups. Each blockgroup is then encoded as a discrete graph. We compute several attributes for each block, between two consecutive blocks and globally over the whole blockgroup (see Figure 1). The attributes are then discretized and used as vertex labels in a graph representation. The resulting graphs are finally processed using the fast Neighborhood Subgraph Pairwise Distance Kernel (NSPDK) [4].

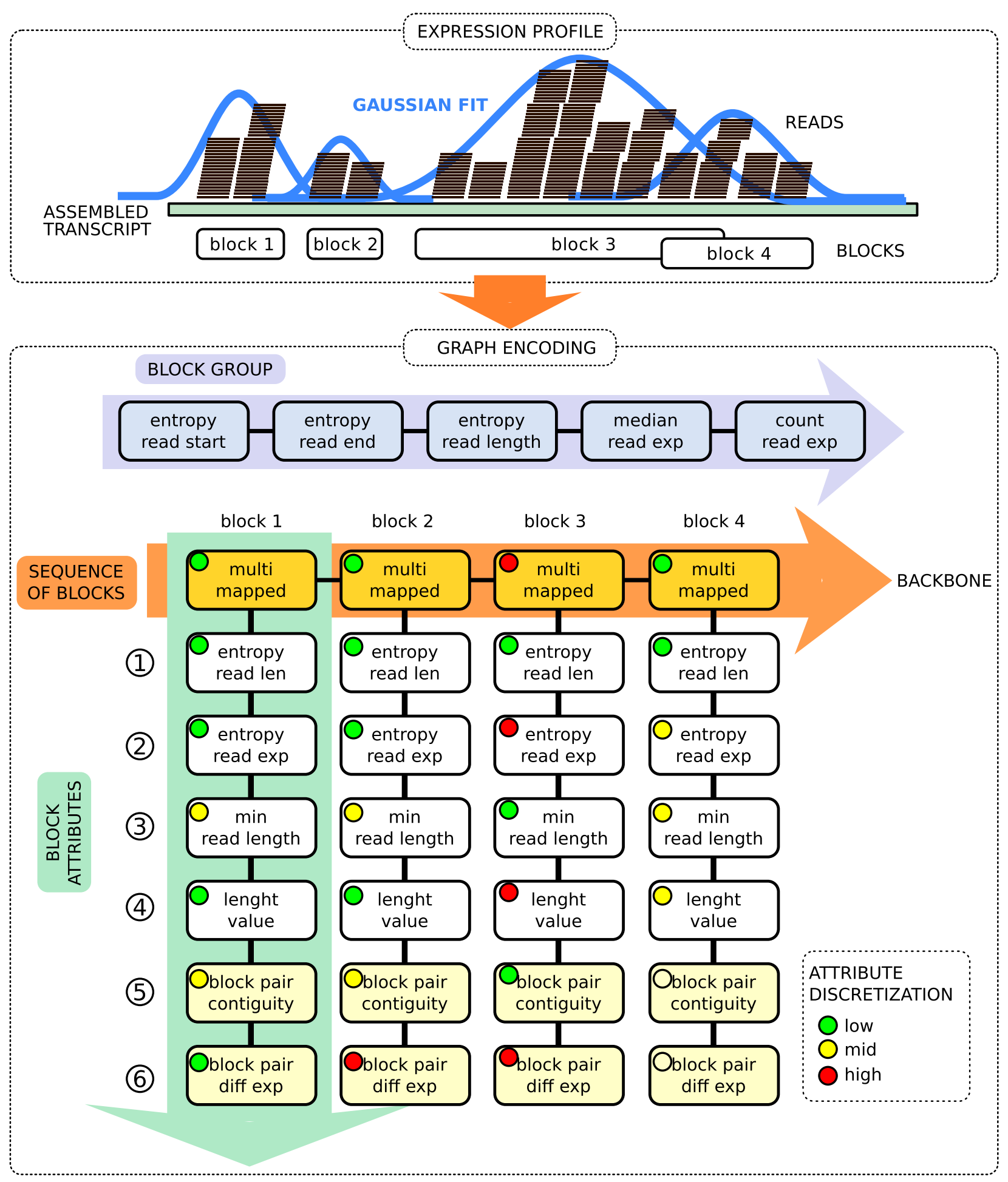


Figure 1: Read profile encoding.

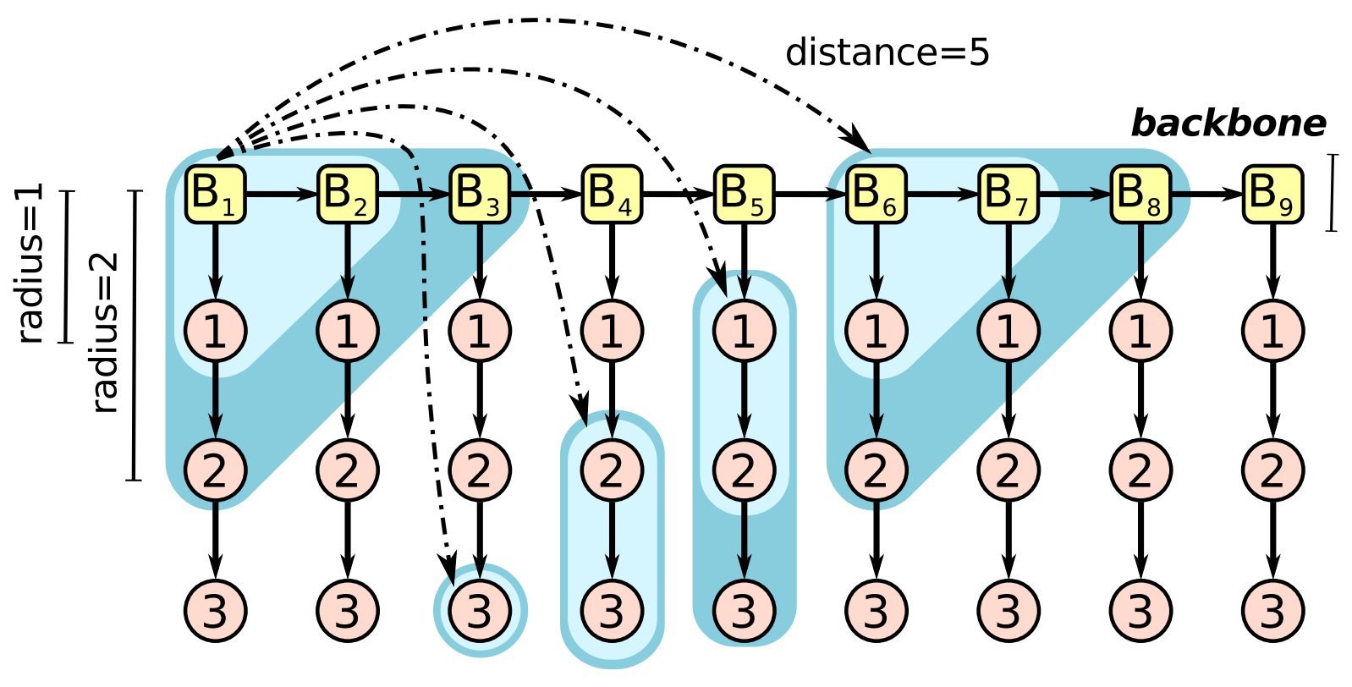


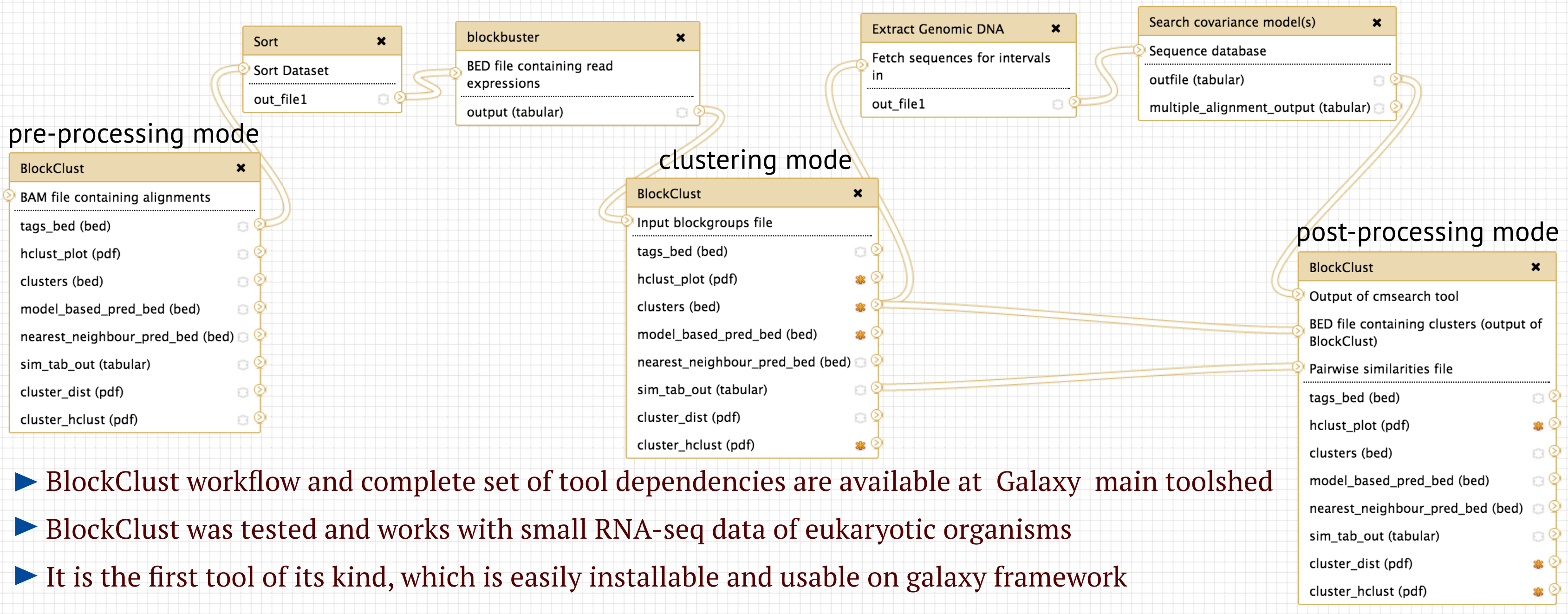
Figure 2: Combinatorial features.

The kernel evaluates the similarity between two graphs as the fraction of neighborhood subgraph pairs they have in common. The similarity notion parametrized by the maximal size of the neighborhood subgraphs and by the maximal distance allowed between the subgraphs in each pair.

Since neighborhood subgraphs can be efficiently enumerated in near linear time, the resulting approach has in practice linear complexity and can be used in large scale settings.

DevelopmentData: for training models; human embryoid body, embryonic stem cells, H1 and IMR90 cell lines.

BenchmarkData: to evaluate robustness; a comprehensive collection of 32 samples from human, mouse, fly, chimp, worm and plant in a variety of tissues and cell lines.



- ▶ BlockClust workflow and complete set of tool dependencies are available at Galaxy main toolshed
- ▶ BlockClust was tested and works with small RNA-seq data of eukaryotic organisms
- ▶ It is the first tool of its kind, which is easily installable and usable on galaxy framework

http://toolshed.g2.bx.psu.edu/view/rnateam/blockclust_workflow

Results

Performance:

▶ We measured the tendency for transcripts of functionally identical RNAs to be neighbors.
▶ We computed the AUC ROC (see Table 1) using the distance as a predictor function to evaluate the quality of the induced metric.
▶ We computed the purity of the partition generated by the Markov Cluster Process [5] to evaluate the clustering quality.

▶ Binary classification models were built for miRNA, tRNA and C/D-box snoRNA classes.

Mode →		Clustering			Classification	
ncRNA class	#transcripts	AUC	#clusters	Purity	PPV	Recall
miRNA	168	0.896	10	0.855	0.901	0.886
tRNA	173	0.741	17	0.837	0.899	0.796
C/D-box snoRNA	78	0.731	7	0.683	0.870	0.474
H/ACA-box snoRNA	4	0.838	0	0	-NA-	-NA-
rRNA	20	0.872	2	0.956	-NA-	-NA-
snRNA	7	0.637	0	0	-NA-	-NA-
Y RNA	8	0.685	0	0	-NA-	-NA-
Weighted average	458	0.805	36	0.813	-NA-	-NA-

Table 1: Performace of BlockClust averaged over 10 random test splits of DevelopmentData

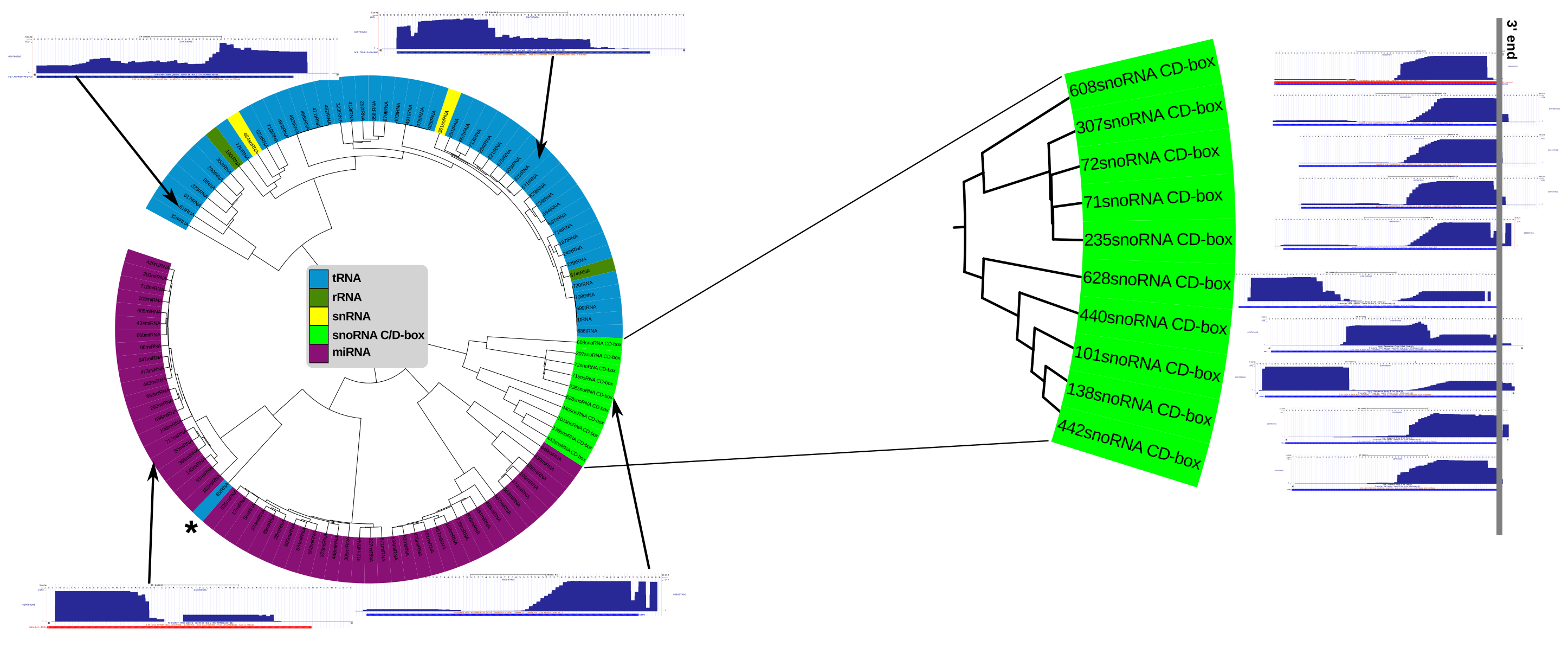
Comparision with other tools:

We compared BlockClust on the BenchmarkData to existing tools that can process read profiles of small ncRNAs from RNA-seq data: deepBlockAlign [6] for clustering and DARIO [7] for classification.
BlockClust achieved a 60-fold speedup (50 seconds vs. 58 minutes on a dataset of ≈600 profiles) w.r.t. deepBlockAlign.

Mode →	Clustering		Classification			
Tool name →	deepBlockAlign	BlockClust	DARIO		BlockClust	
ncRNA class	AUC	AUC	PPV	Recall	PPV	Recall
miRNA	0.714	**0.925**	0.85	0.81	**0.88**	**0.89**
tRNA	0.701	**0.795**	0.92	**0.88**	**0.95**	0.80
C/D-box snoRNA	0.615	**0.762**	0.46	**0.52**	**0.74**	0.39
H/ACA-box snoRNA	0.720	**0.859**	-NA-	-NA-	-NA-	-NA-
rRNA	0.759	**0.873**	-NA-	-NA-	-NA-	-NA-
snRNA	0.610	**0.698**	-NA-	-NA-	-NA-	-NA-
Y RNA	0.656	**0.694**	-NA-	-NA-	-NA-	-NA-
Weighted average	0.700	**0.839**	-NA-	-NA-	-NA-	-NA-

Table 2: Performace comparision of BlockClust vs. deepBlockAlign and DARIO

Clustering example:



- ▶ Hierarchical clustering plot on one of the BenchmarkData samples.
- ▶ **tRNA:** mixture of tRNA halves, 5 - or 3 -derived fragments.
- ▶ **miRNA:** typical miRNA and miRNA* blocks.
- ▶ **C/D-box snoRNA:** step-wise extension for towards 3'-end.

References

[1] Sven Findeiss, David Langenberger, Peter F. Stadler, and Steve Hoffmann. Traces of post-transcriptional RNA modifications in deep sequencing data. Biol Chem, 392(4):305–13, 2011.
[2] Pavankumar Videm, Dominic Rose, Fabrizio Costa and Rolf Backofen. BlockClust: efficient clustering and classification of non-coding RNAs from short read RNA-seq profiles. Bioinformatics, 30(12), i274-i282, 2014.
[3] David Langenberger, Clara Bermudez-Santana, Jana Hertel, Steve Hoffmann, Philipp Khaitovich, and Peter F. Stadler. Evidence for human microRNA-offset RNAs in small RNA sequencing data. Bioinformatics, 25(18):2298–301, 2009.
[4] Fabrizio Costa and Kurt De Grave. Fast Neighborhood Subgraph Pairwise Distance Kernel. In Proceedings of the 26 th International Conference on Machine Learning, pages 255–262. Omnipress, 2010.
[5] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, 30 (7):1575–84, 2002.
[6] David Langenberger, Sachin Pundhir, Claus T. Ekstrom, Peter F. Stadler, Steve Hoffmann, and Jan Gorodkin. deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns. Bioinformatics, 28(1):17–24, 2012.
[7] Mario Fasold, David Langenberger, Hans Binder, Peter F. Stadler, and Steve Hoffmann. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res, 39(Web Server issue):W112–7, 2011