Classification and Identification of Non-coding RNAs using High Throughput Sequencing Data

David Langenberger[1], Steve Hoffmann[1], Clara Bermudez-Santana[1,2], and Peter F. Stadler[1,3−6]

[1]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University Leipzig, Germany; [2]Department of Biology, Universidad National de Colombia, Bogota, Colombia; [3]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany; [4]Fraunhofer Institut für Zell Therapie und Immunologie (IZI), Leipzig, Germany; [5]Department of Theoretical Chemistry, University of Vienna, Austria; [6]Santa Fe Institute, Santa Fe, New Mexico, USA

E-mail: {david, steve, clara, stadler}@bioinf.uni-leipzig.de
WorldWideWeb: http://www.bioinf.uni-leipzig.de/

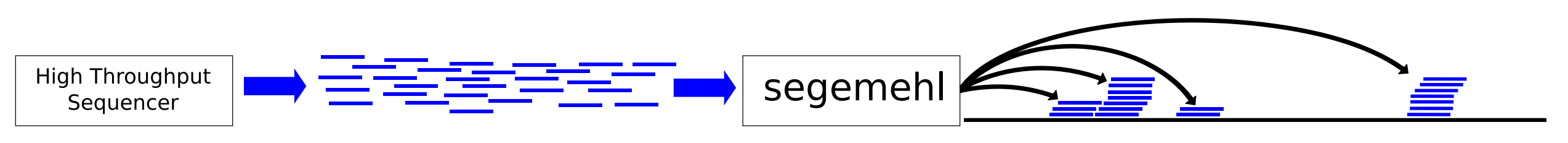
Introduction

High Throughput Sequencing (HTS) has the ability to rapidly sequence millions of individual small RNA molecules. This, for the first time, offers the opportunity to identify known and classify new small RNAs based on unbiased data. Recently, several strategies have been devised for that purpose [1,2].
Small RNAs are a diverse collection of molecules with several important biological functions. Thus, exploring small RNA biology or characterizing differential expression profiles by sequencing offers an exciting possibility to get more information about non-coding RNAs (ncRNAs).
Some of these ncRNA classes, in particular microRNAs and snoRNAs, undergo maturation processes that lead to the production of shorter RNAs. After mapping the sequences to the reference genome specific patterns of short reads can be observed. These read patterns seem to reflect the processing and thus are specific for the RNA transcripts of which they are derived from (Fig. 1).

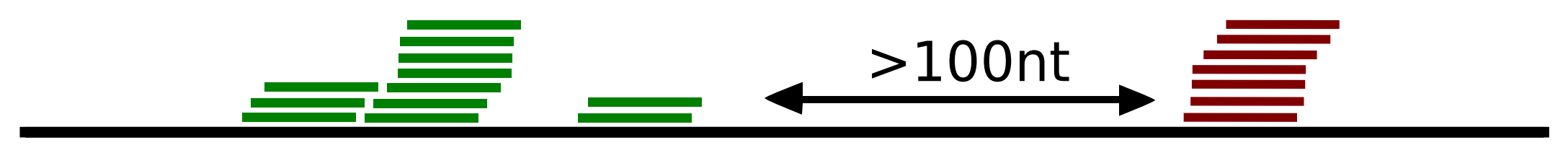
We explore here the potential of short read sequence data in the classification and identification of ncRNAs.

Data Preparation

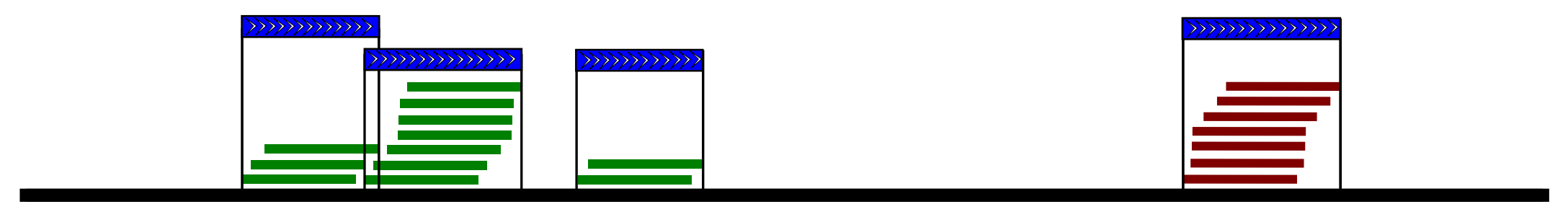
1. Map small RNAs to the human genome, using `segemehl` [3]



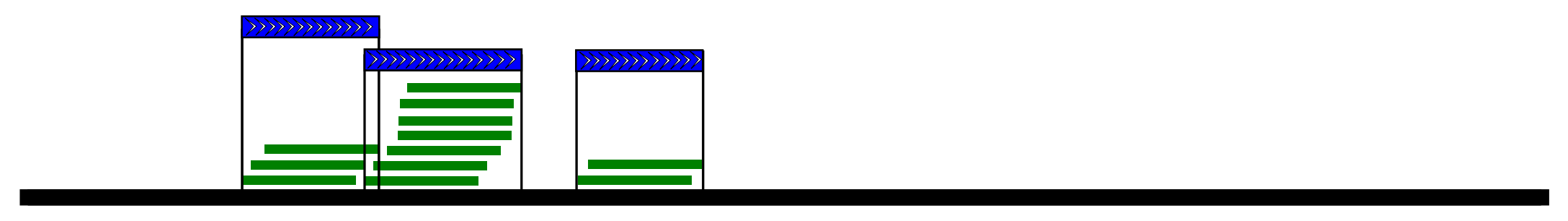
2. Cluster hits based on their genomic location (distance $<$100nt)



3. Divide consecutive reads into blocks, using `blockbuster` [4]



4. Discard clusters with $<$2 blocks and/or $<$10 reads (small information content)



We started with **355,453 unique reads** that were mapped to **2,191,220 positions** of the human genome (NCBI36). After all clustering and filtering steps we identified **852 clusters**. This set comprises **2,538 individual blocks** and **85,459 unique reads**.

Machine Learning Approach

Classifier

We implemented a highly accurate machine learning approach based on the random forest method [5] for the classification of three types of ncRNAs: microRNAs, snoRNAs, and tRNAs.

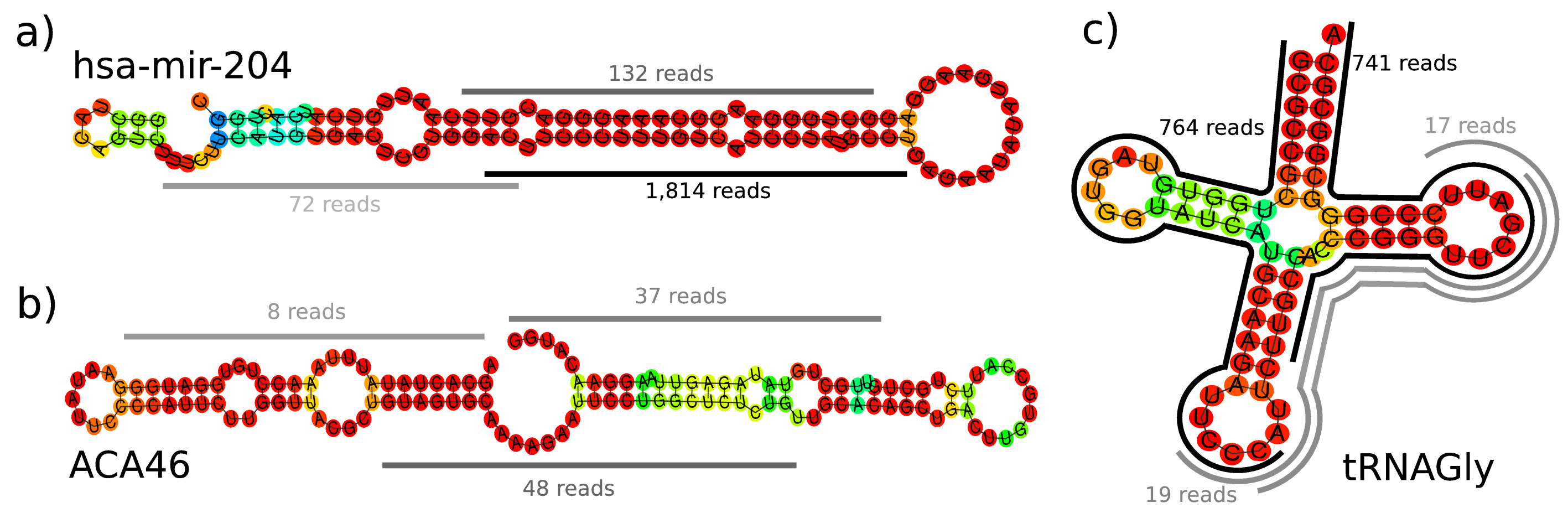


Fig. 1. Non-coding RNAs exhibit specific block patterns and secondary structures. (a) The class of miRNAs often shows a block pattern of two or three separated blocks and the sequence folds into the typical miRNA hairpin structure. (b) snoRNAs tend to have miRNA-like mature and star blocks at their 5' and 3' hairpins with minor overlaps, while (c) a series of overlapping blocks and the cloverleaf structure is striking for the tRNA class.

Training Set

The classifier was trained with all clusters found within annotated ncRNAs loci [`miRBase v.12` (727 entries), `tRNAscan-SE` (588 entries) and `snoRNAbase v.3` (451 entries)], see Table 1.

Table 1. In total 434 of 852 clusters were found within regions of annotated miRNA, tRNA and snoRNA loci.

RNA class	source	loci found	blocks/cluster (mean)	reads/cluster (median)
microRNAs	miRBase v.12	218	2.42 ± 1.04	4535.33
tRNAs	tRNAscan SE	87	3.22 ± 1.92	183.95
snoRNAs	snoRNAbase v.3	129	2.60 ± 1.66	127.5

Features

Based on visual inspection of the mapped reads (see Fig. 1), twelve features were selected to train the random forest model (see Fig. 2).

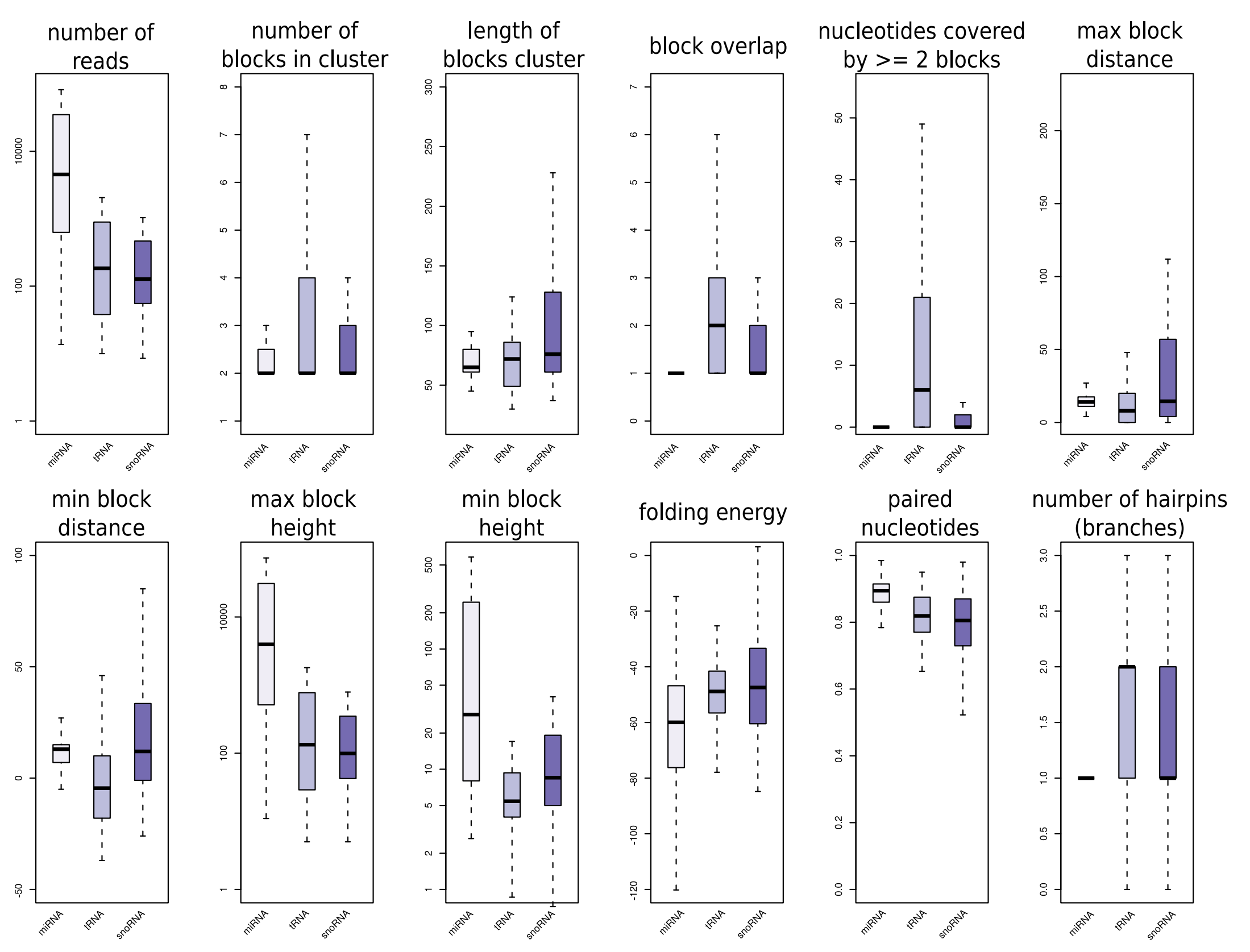


Fig. 2. Box plots for 12 different features selected to train the random forest classifier.

Performance

The random forest model was repeatedly trained with randomly chosen annotated loci (50% of each class as training set and the remaining 50% as test set) in order to determine positive predictive values (PPV) and recall rates (see Table 2).

Table 2. Positive predictive values (PPV) and recall rates. For each set size means, medians, and standard deviations are calculated from 20 randomly sampled training sets.

	PPV		recall	
	mean	sdev	mean	sdev
all	0.911	0.021	0.784	0.022
miRNA	0.953	0.021	0.892	0.020
tRNA	0.864	0.063	0.673	0.053
snoRNA	0.859	0.063	0.681	0.077

First Results and Outlook

Using our machine learning approach, we were able to predict 29 new miRNAs of which 3 (hsa-mir-1978, hsa-mir-2110, hsa-mir-1974) have already been annotated in the most recent `miRBase release (v.14)`, as well as a novel member of the mir-548 family and another locus is the human ortholog of the bovine mir-2355. Two clusters lie in antisense direction to annotated miRNA loci (hsa-mir-219-2 and hsa-mir-625). 22 clusters classified as miRNA remain to be confirmed. Furthermore, we created UCSC-tracks in order to visualize the results (see Fig. 3a) and compare our data to other HTS datasets (see Fig. 3b).

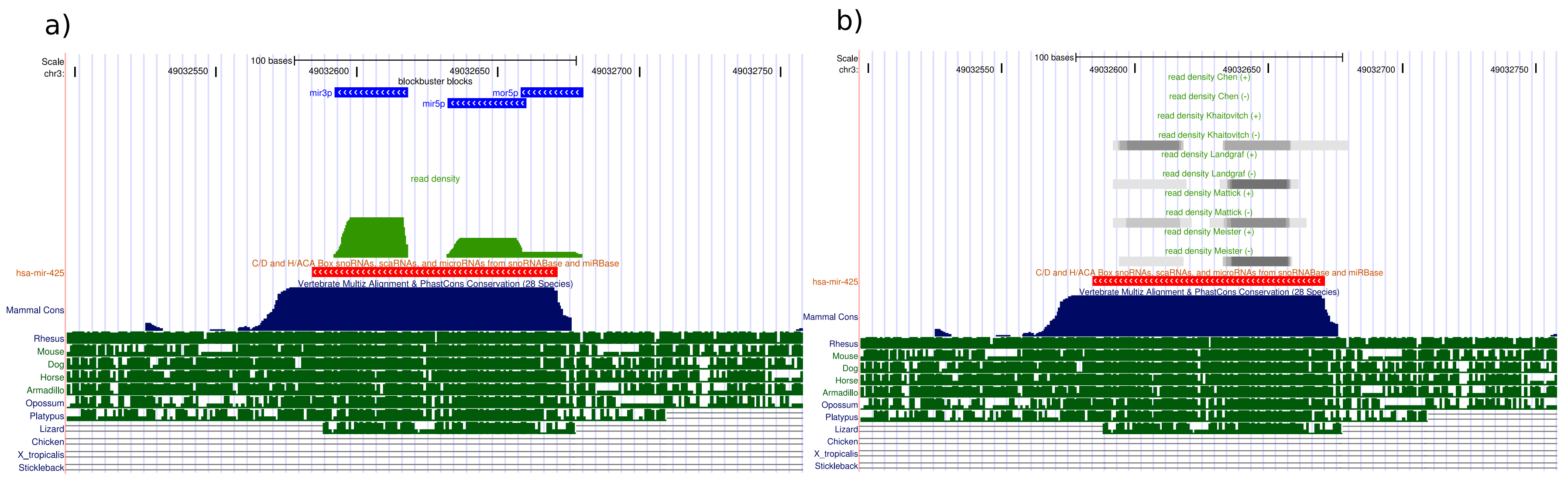

a)
b)
hsa-mir-425
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback

Fig. 3. Distribution of short reads at the hsa-mir-425 locus. (a) There are three clearly distinct blocks of reads: they correspond to moR (5'-end), miR* (center) and miR (3'-end) transcripts. The conservation pattern is shown below. (b) Comparision of our hsa-mir-425 locus with other HTS datasets.

For the future we are planning to integrate more published HTS datasets in order to improve and validate our machine learning approach. We will also use HTS data from other species for further comparative analyses of our results.

References

[1] Michael Hackenberg, Martin Sturm, David Langenberger, Juan Manuel Falcon-Perez, Ana M. Aransay. **miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments**. Nucleic Acid Research, 2009.

[2] Marc R. Friedländer, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, Nikolaus Rajewsky. **Discovering microRNAs from deep sequencing data using miRDeep**. Nature Biotech, 26(4):407-415, 2008.

[3] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia Sharma, Philipp Khaitovitch, Jörg Vogel, Peter F. Stadler,and Joerg Hackermüller. **Fast mapping of short sequences with mismatches, insertions and deletions using index structures**. PloS Comp. Biol., 2009

[4] David Langenberger, Clara Bermudez-Santana, Jana Hertel, Steve Hoffmann, Philipp Khaitovitch, Peter F. Stadler. **Evidence for Human microRNA-Offset RNAs in Small RNA Sequencing Data**. Bioinformatics, 2009.

[5] L. Breiman. **Random forests**. Machine learning, 45(1):5-32, 2001.

Printed at the Universitätsrechenzentrum Leipzig.