UNIVERSITÄT
LEIPZIG

Hierarchical classification of ncRNAs in small RNA seq data based on block patterns

35th Winterseminar Bled

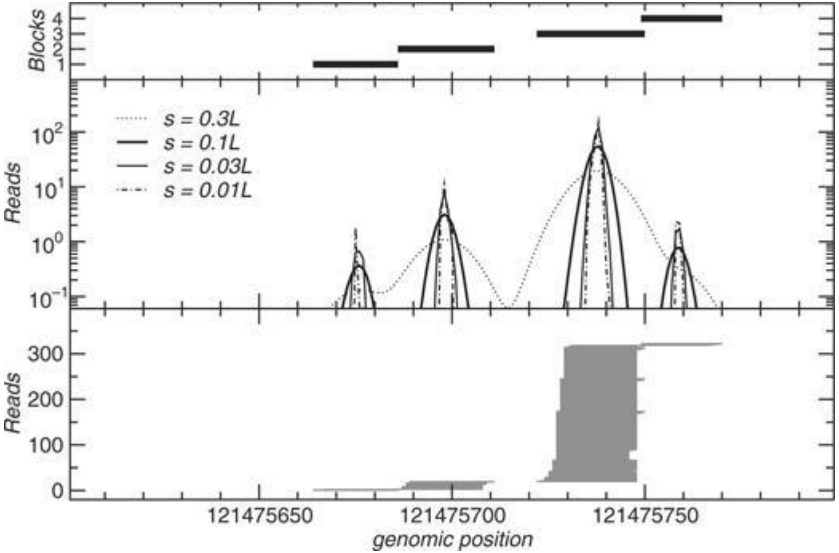
Tobias Hagemann

Bioinf Leipzig

12.02.2020

Basics

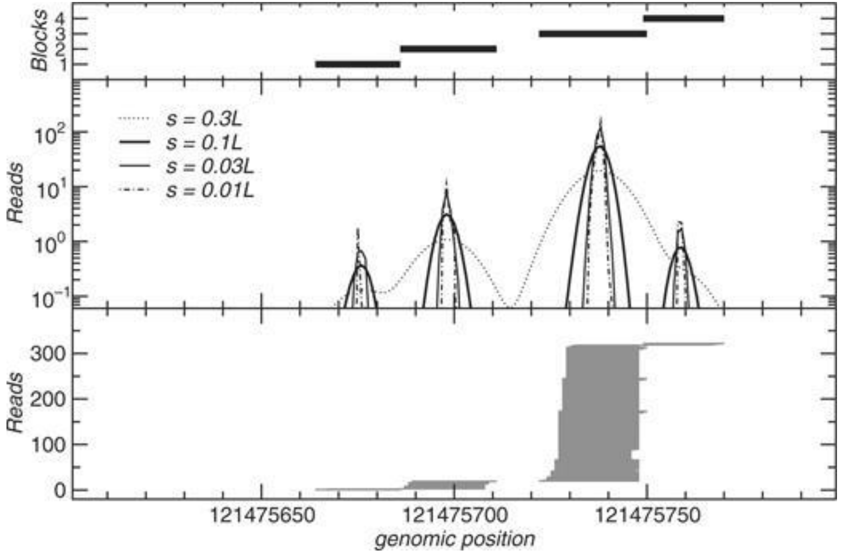
BLOCKBUSTER [1]



[1] Langenberger et al. (2009)

Basics

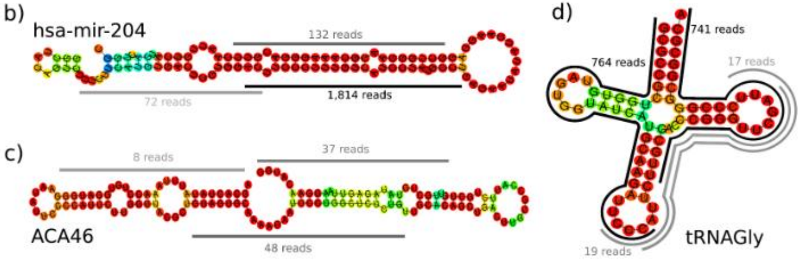
BLOCKBUSTER [1]



1) Langenberger et al. (2009)

DARIO [2] [3]

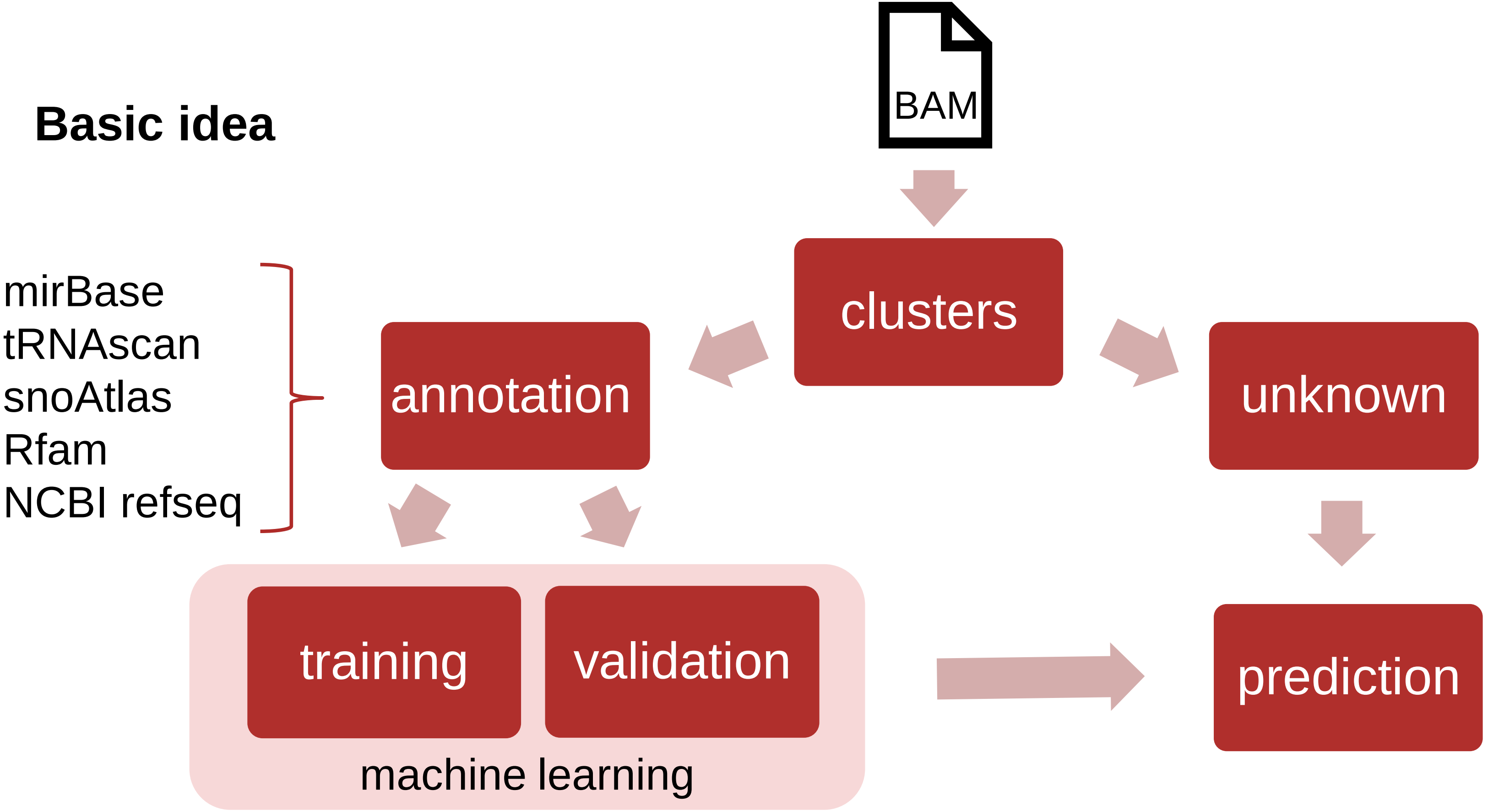
- Train and classify transcribed regions
- Use block patterns
- For tRNA, miRNA, snoRNA



2)

2) Langenberger et al. (2010)
3) Fasold et al. (2011)

Basic idea


BAM
clusters
annotation
unknown
mirBase
tRNAscan
snoAtlas
Rfam
NCBI refseq
training
validation
machine learning
prediction

Learning on bam files

Account for differences in

- Mapping
- Sequencing
- Species
- Tissue
- …

Learning on bam files

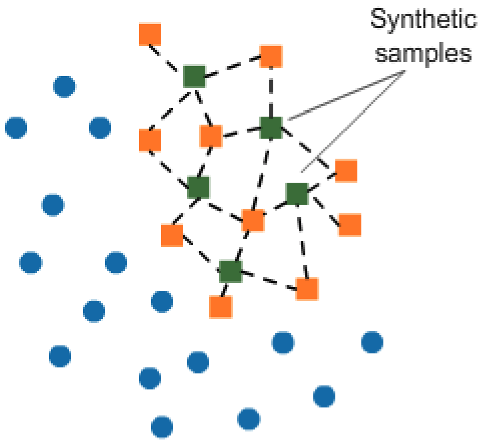
Account for differences in

- Mapping
- Sequencing
- Species
- Tissue
- …

Small sample size &
class imbalance

Synthetic minority oversampling technique (SMOTE)

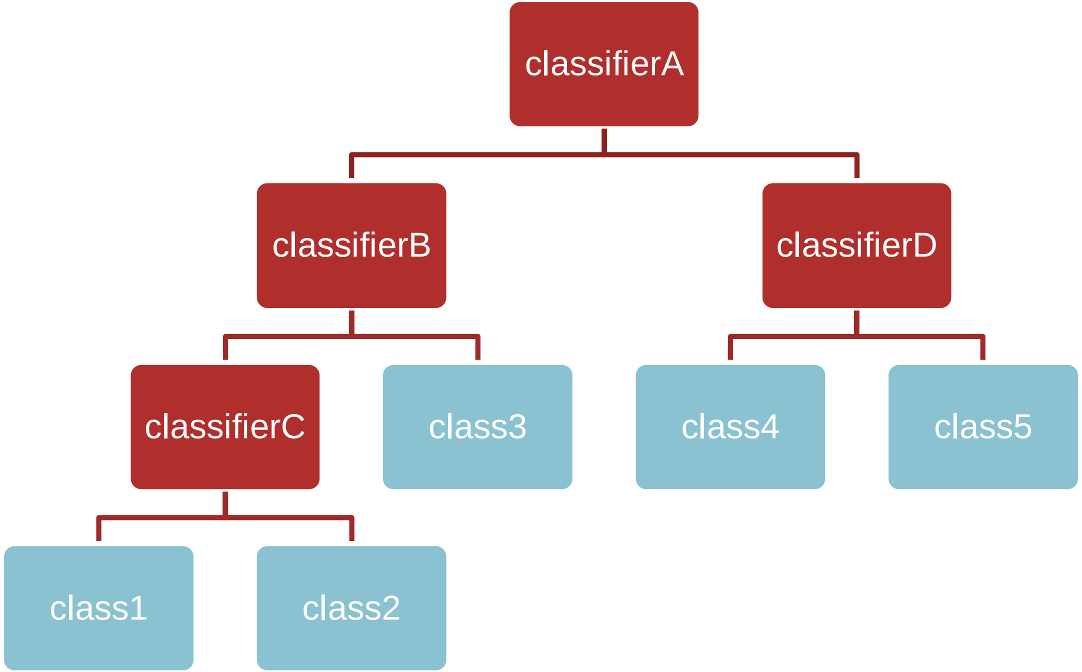
Class	sample size before SMOTE	Sample size after SMOTE
tRNA	326 (36,2%)	326
snoRNA	218 (24,2%)	326
mRNA	210 (23,3%)	326
miRNA	68 (7,5%)	326
rRNA	57 (6,3%)	326
∑	879 (97,5%)	1630



https://towardsdatascience.com/super-bowl-prediction-model-99048f366fed

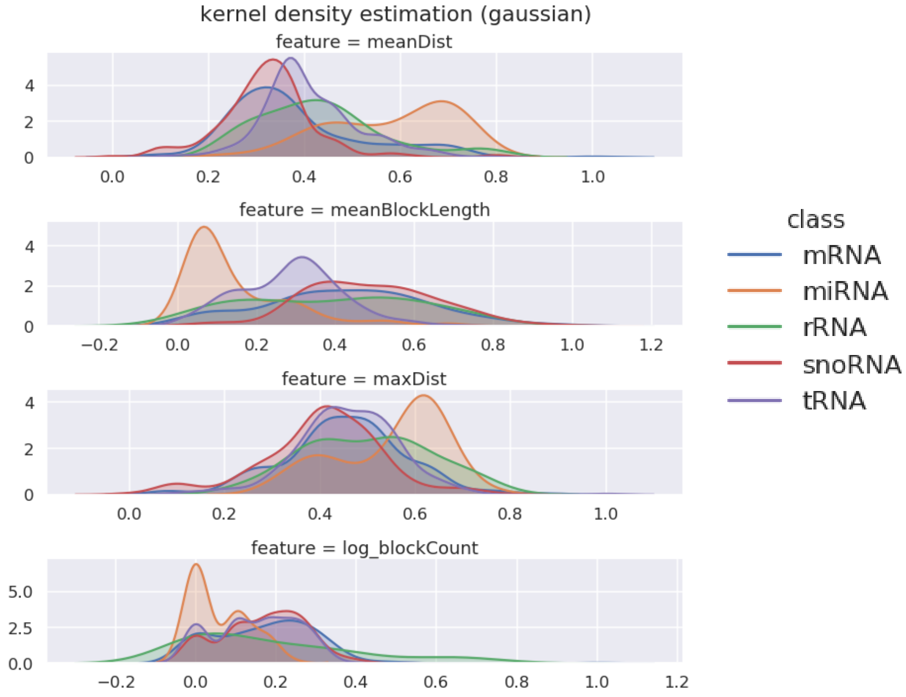
Hierarchical classification

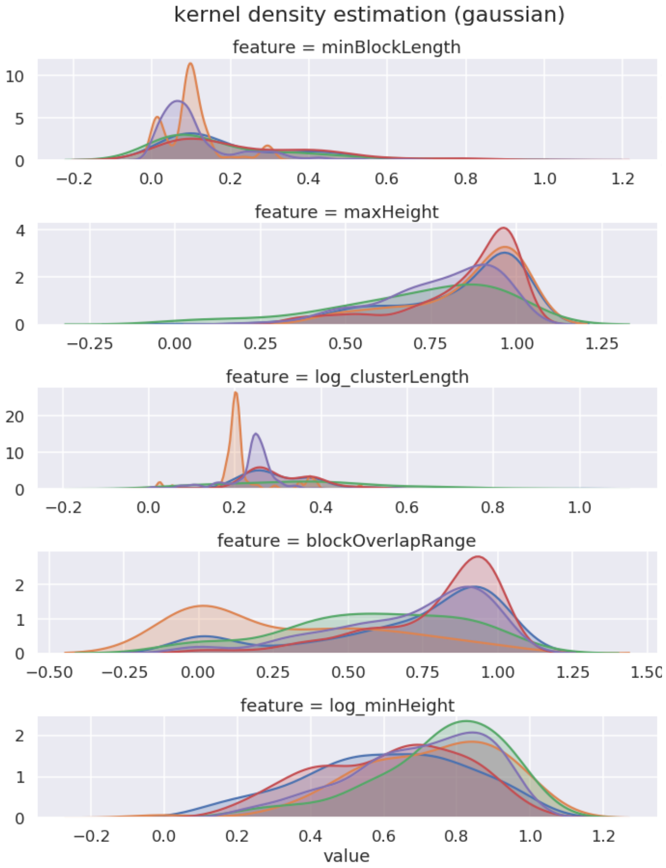
- “Local classifier per parent node“ approach [4)]
- Feature selection and clustering at every parent node



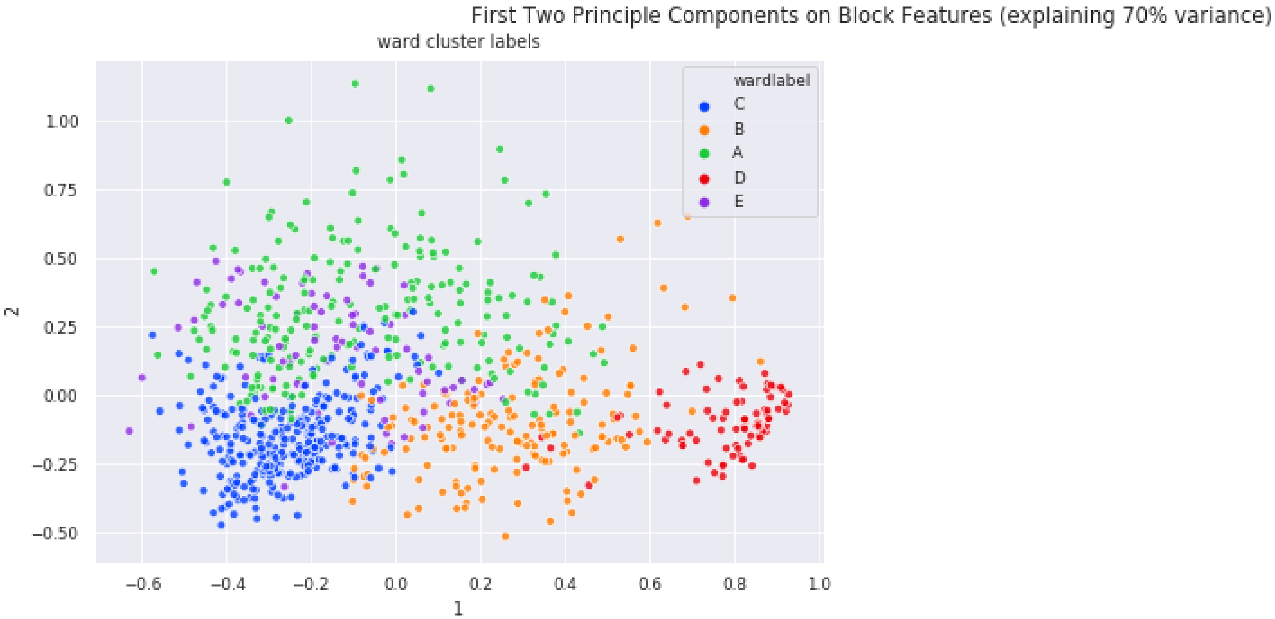
4) Silla & Freitas (2001)

Block feature analysis

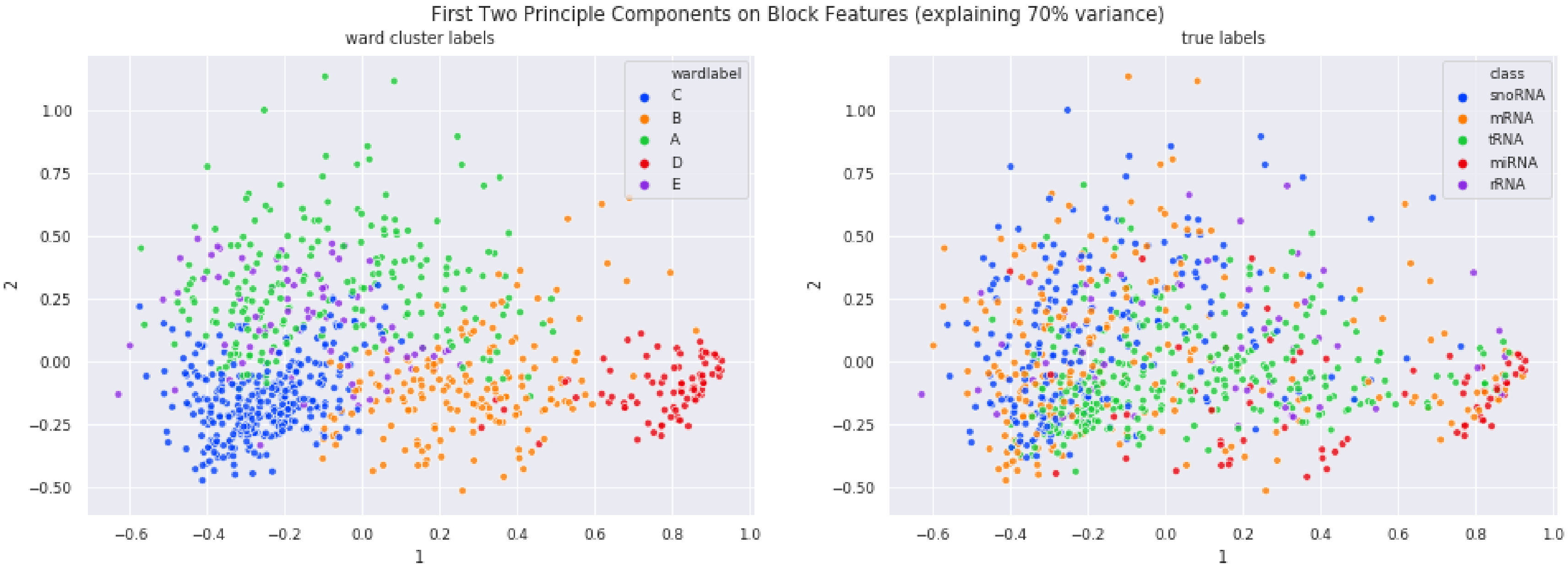




Block feature analysis

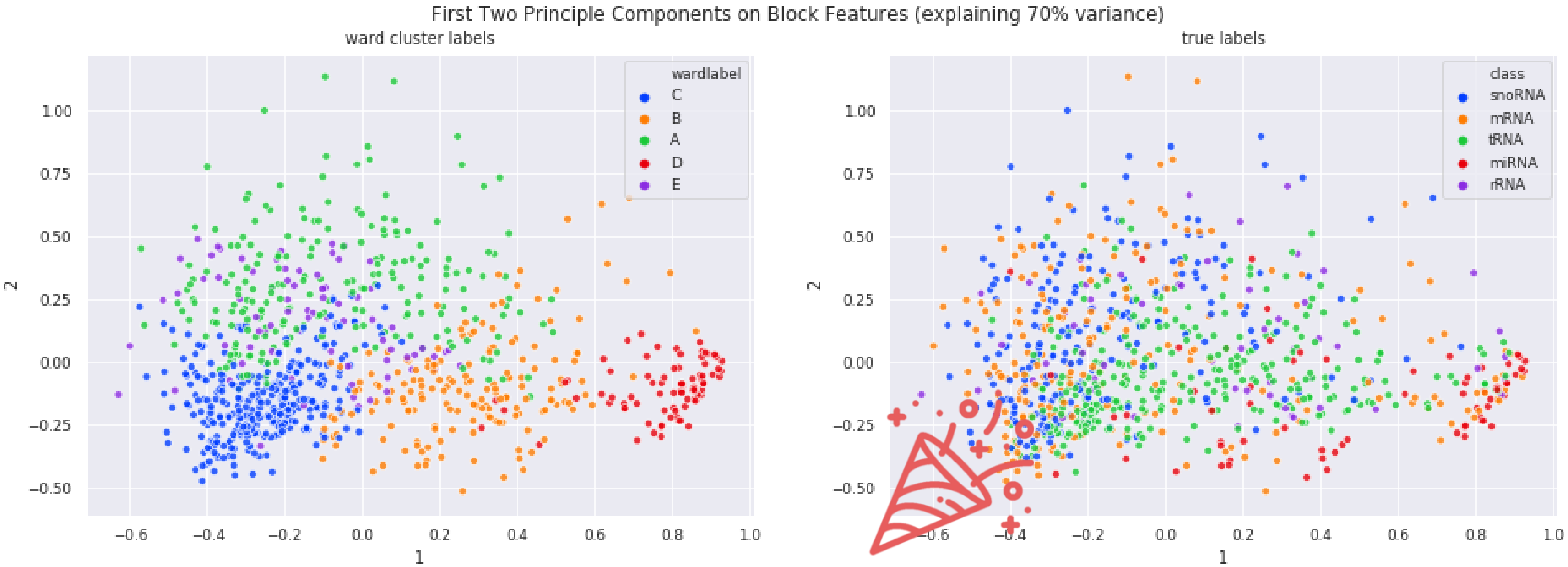


Block feature analysis



UNIVERSITÄT
LEIPZIG

Block feature analysis



Block feature engineering on cluster level

Feature type based on …	Possible Features
Sequence	• GC content/block • K-tuple composition/block (Shannon entropy)
Mapping	• Mismatches/block • Average quality/block
Structure	• Mfe • Self-containment • Base pair entropy • Accessibility (at or in between blocks) • Hairpins/block

Novel structures
but
Less conservation

Better performance
but
overfitting

Sources

1) Langenberger, David, et al. "Evidence for human microRNA-offset RNAs in small RNA sequencing data." *Bioinformatics* 25.18 (2009): 2298-2301
2) Langenberger, David, et al. "Identification and classification of small RNAs in transcriptome sequence data." *Biocomputing 2010*. 2010. 80-87.
3) Fasold, Mario, et al. "DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments." *Nucleic acids research* 39.suppl_2 (2011): W112-W117.
4) Silla, Carlos N., and Alex A. Freitas. "A survey of hierarchical classification across different application domains." *Data Mining and Knowledge Discovery* 22.1-2 (2011): 31-72.

UNIVERSITÄT
LEIPZIG

THANKS TO ...

Stephanie Kehr

Peter Stadler

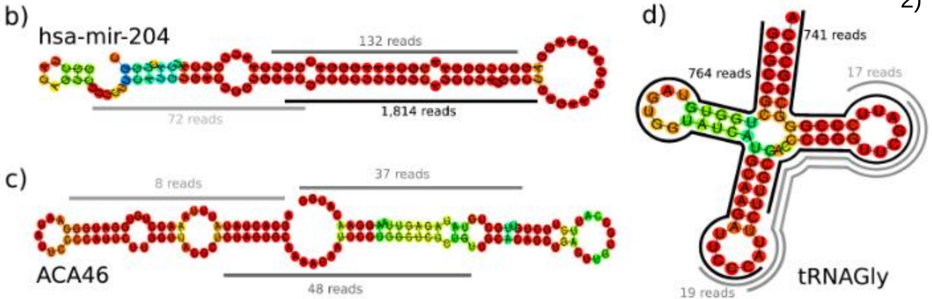
Viewpoint Hunters

You!

DARIO performance

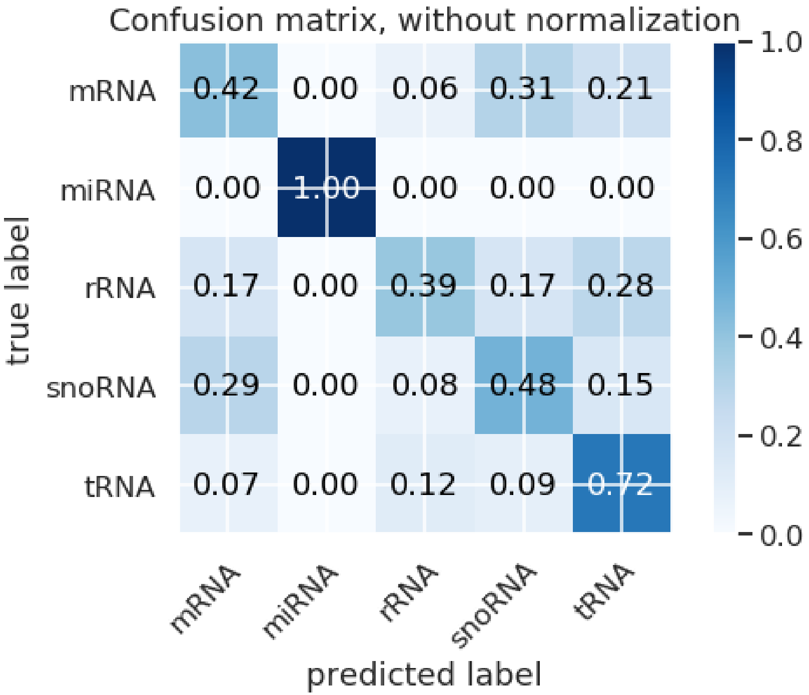
class	PPV	Recall
tRNA	0,932	0,918
miRNA	0,860	0,633
snoRNA	0,819	0,694

2)

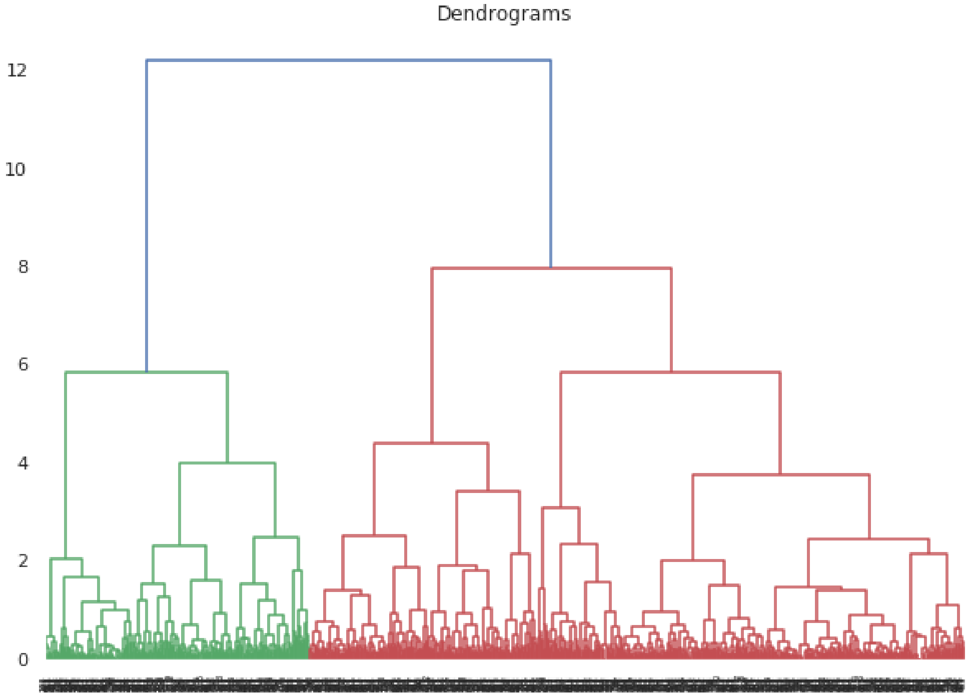


2) Langenberger et al. (2010)

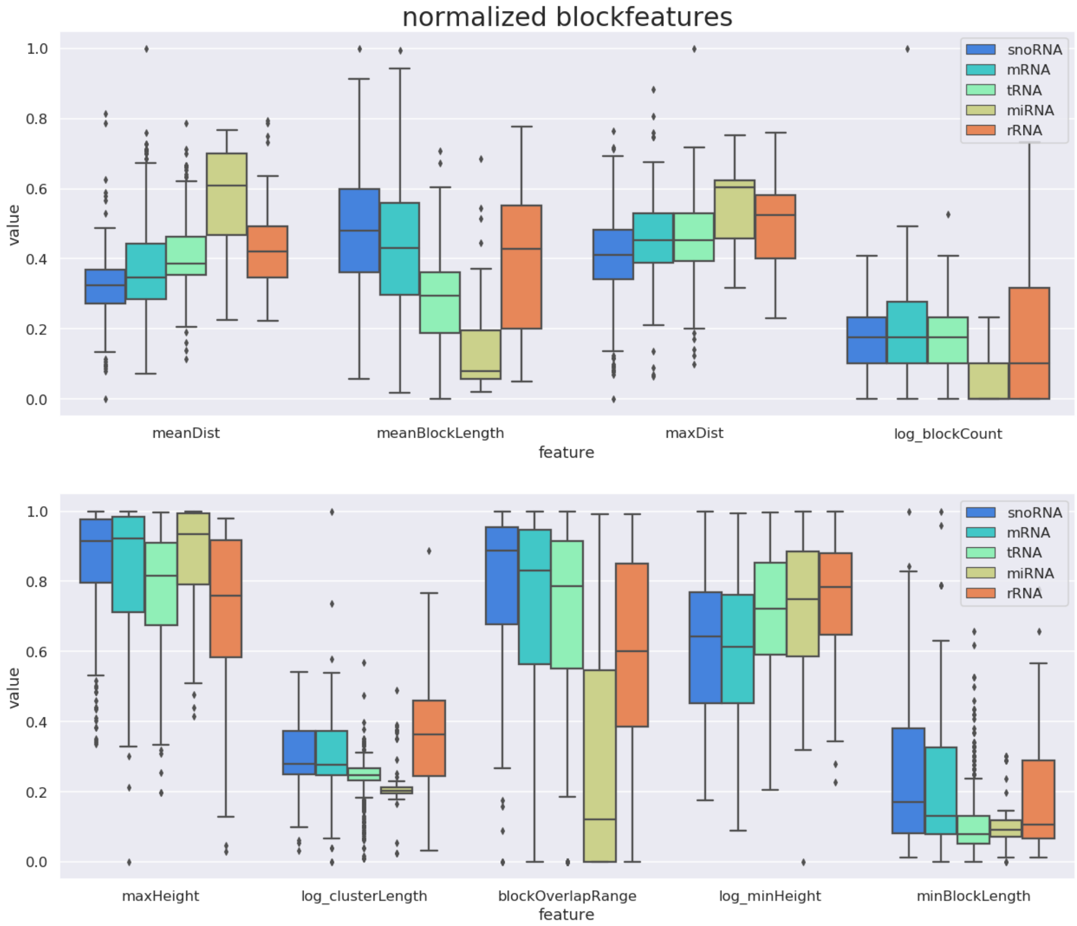
Random forest prediction on 5 classes



Ward clustering



Block features (DARIO)


normalized blockfeatures
value
meanDist
meanBlockLength
maxDist
log_blockCount
feature
snoRNA
mRNA
tRNA
miRNA
rRNA
maxHeight
log_clusterLength
blockOverlapRange
log_minHeight
minBlockLength