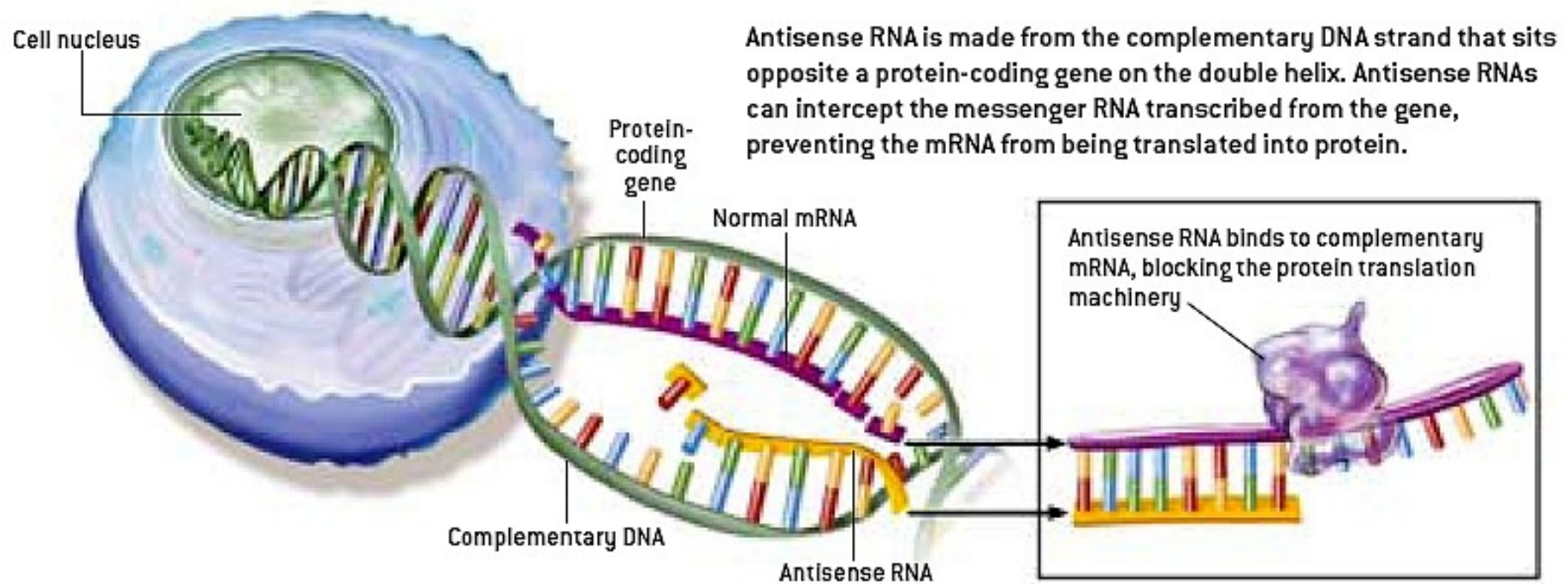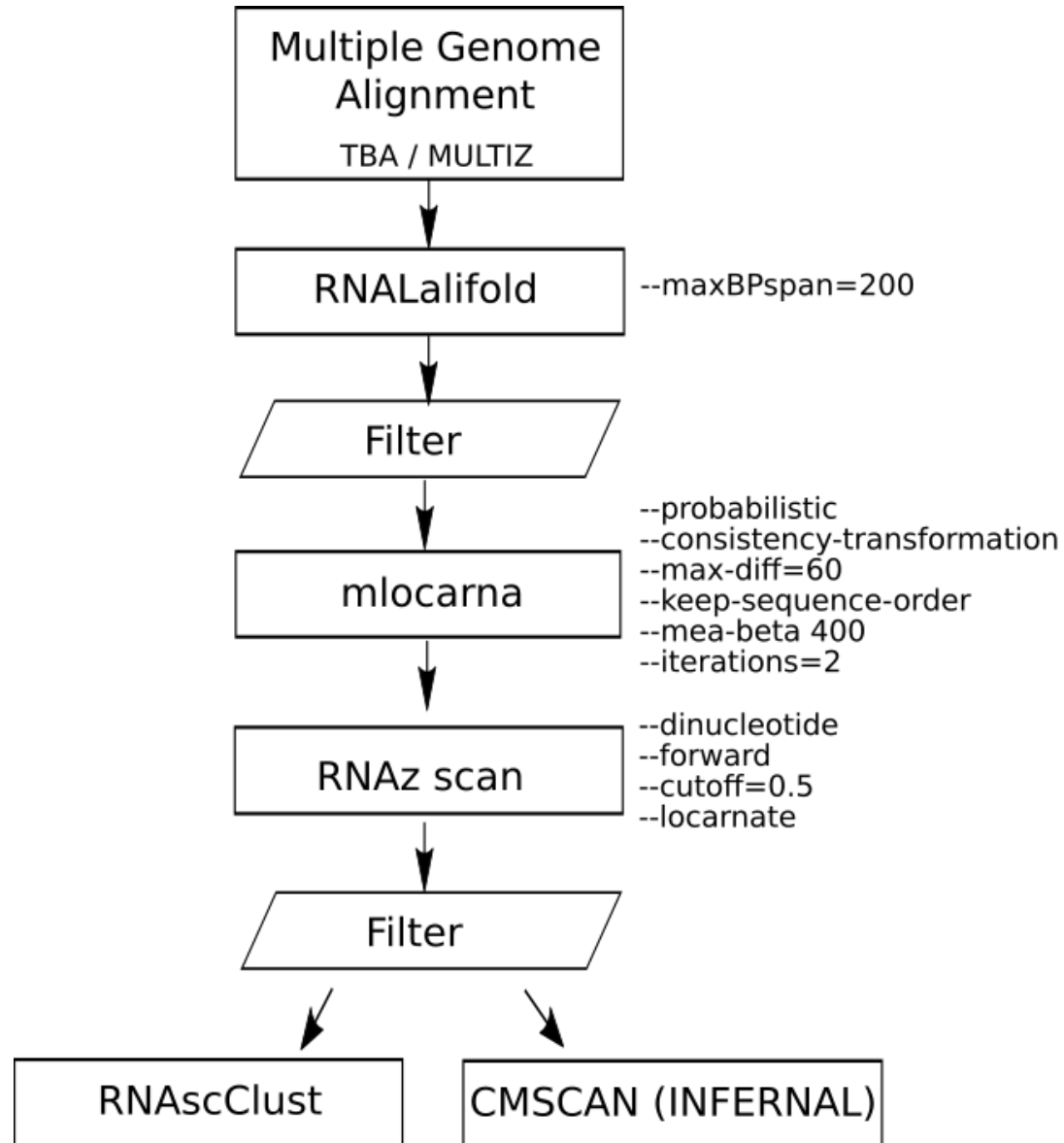# Clustering RNA secondary structures in Arabidopsis thaliana

Veerendra Gadekar
Department of Theoretical Chemistry
University of Vienna
Bled, Feb 2018

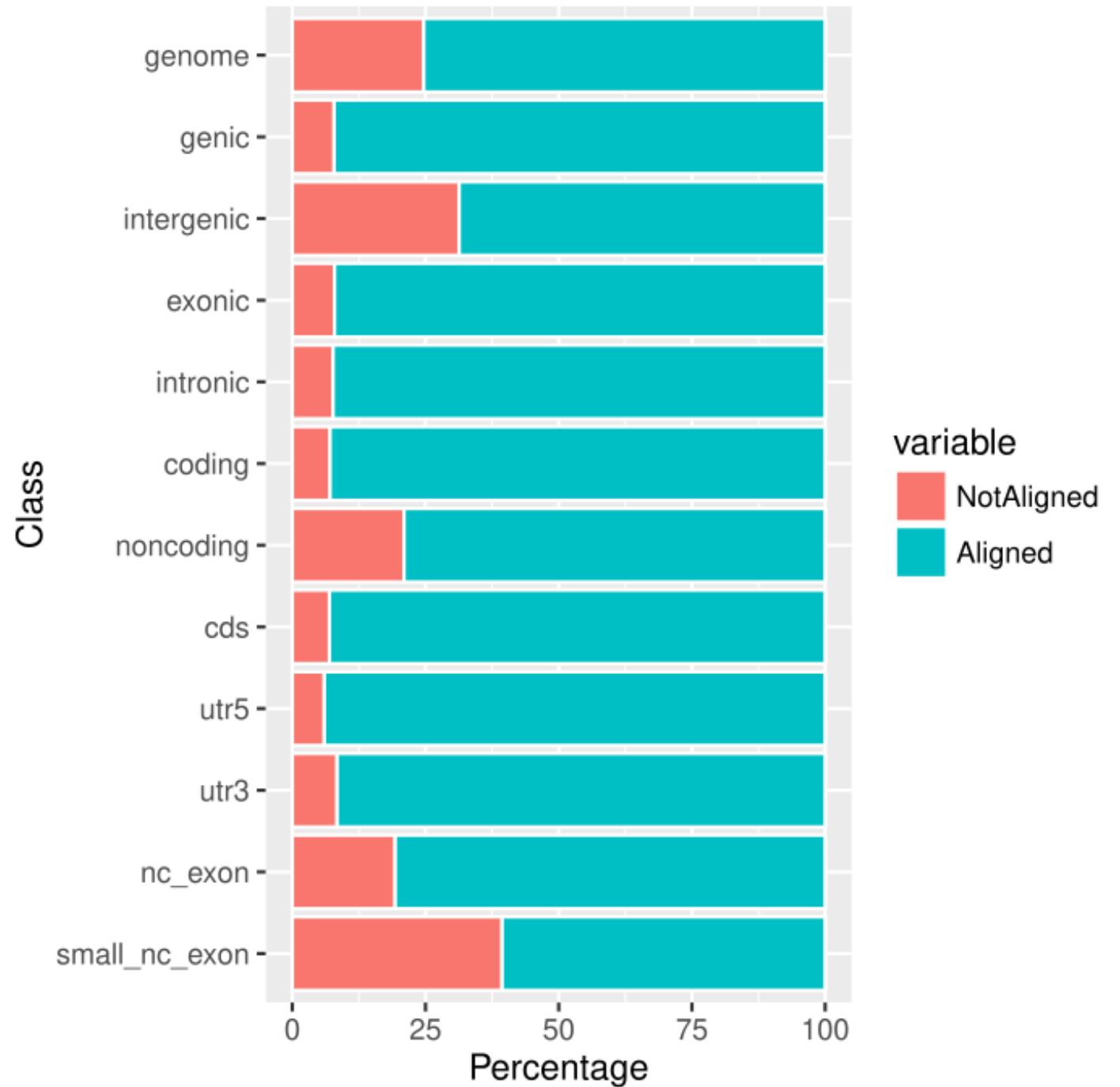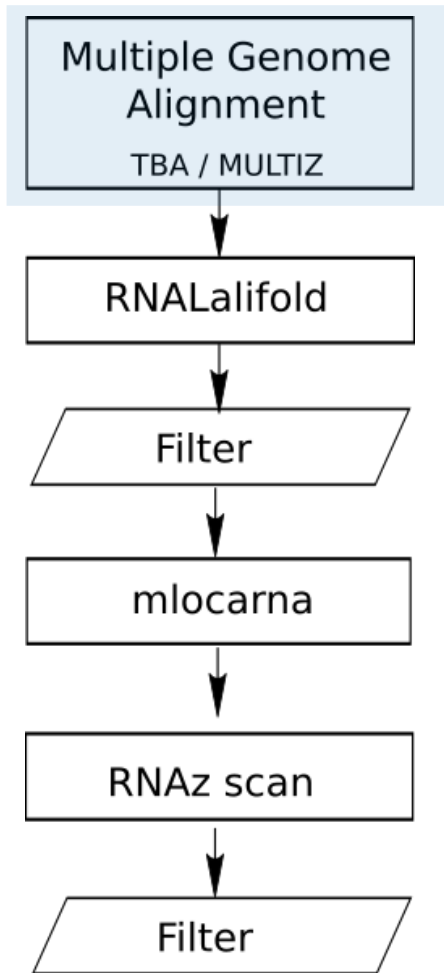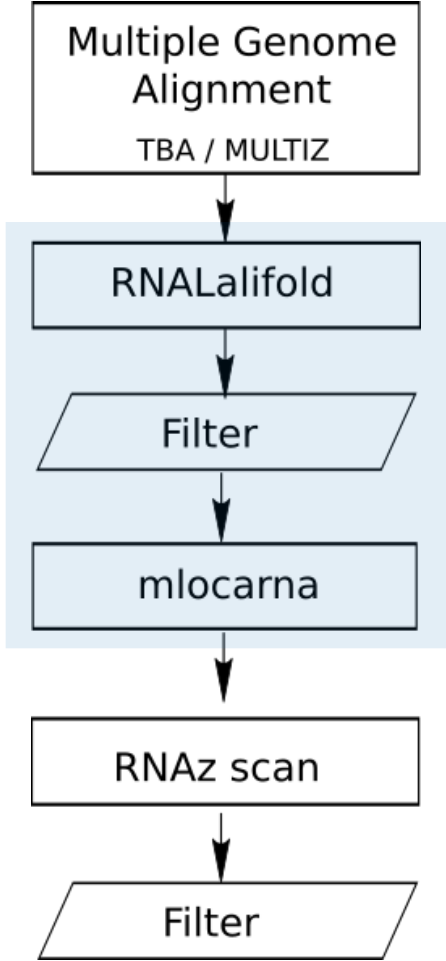# Cis-natural antisense transcripts



Cell nucleus

Protein-coding gene

Normal mRNA

Complementary DNA

Antisense RNA

Antisense RNA is made from the complementary DNA strand that sits opposite a protein-coding gene on the double helix. Antisense RNAs can intercept the messenger RNA transcribed from the gene, preventing the mRNA from being translated into protein.

Antisense RNA binds to complementary mRNA, blocking the protein translation machinery

# Pipeline

Multiple Genome Alignment
TBA / MULTIZ

RNALalifold

Filter

mlocarna

RNAz scan

Filter

A.thaliana 37

Class (top to bottom): genome, genic, intergenic, exonic, intronic, coding, noncoding, cds, utr5, utr3, nc_exon, small_nc_exon

variable
NotAligned
Aligned

Percentage
0    25    50    75    100

Multiple Genome
Alignment

TBA / MULTIZ

RNALalifold

Filter

mlocarna

RNAz scan

Filter
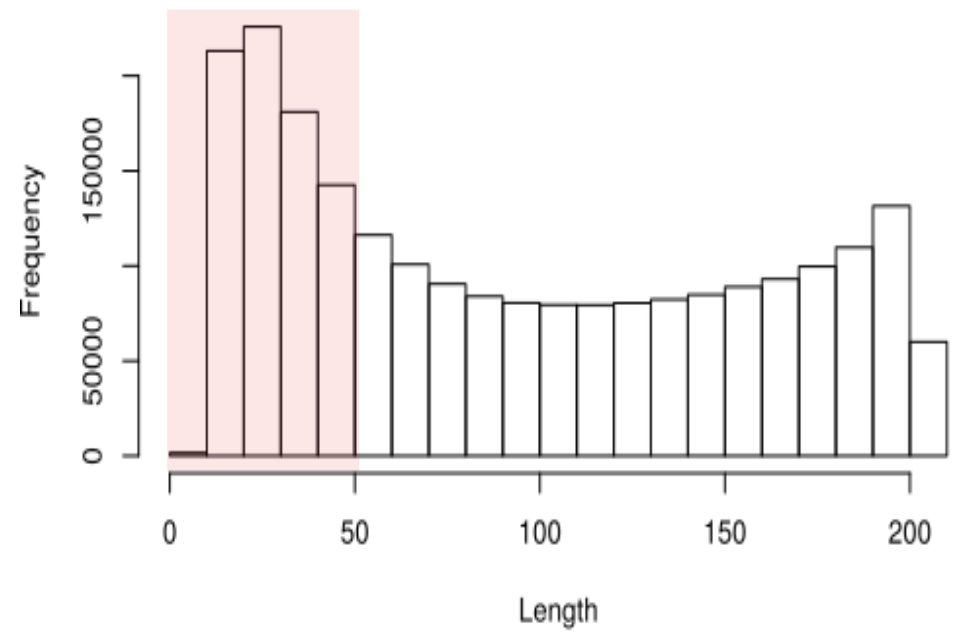
**Forward strand**
**RNALalifold Hits: 2225930**

Frequency

Length

**RNALalifold Hits Used: 590319 (2%)**

Frequency
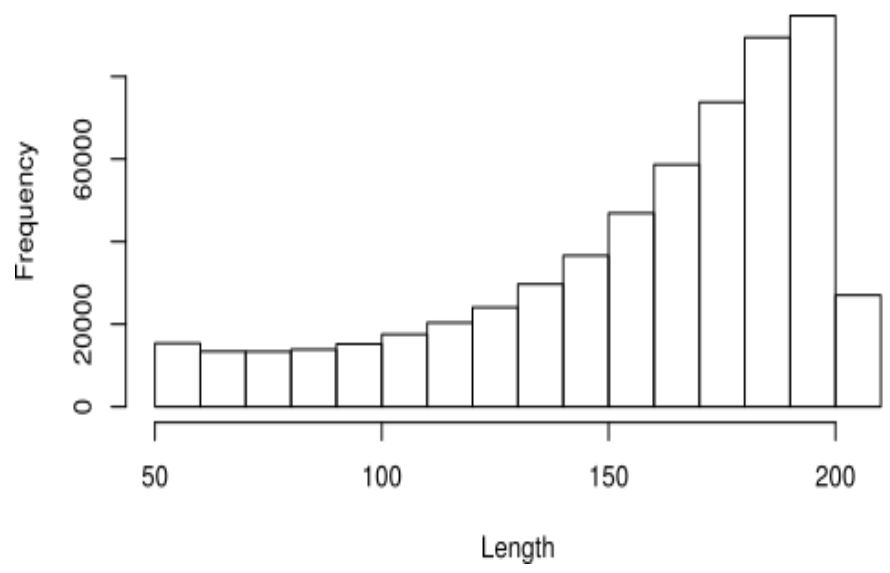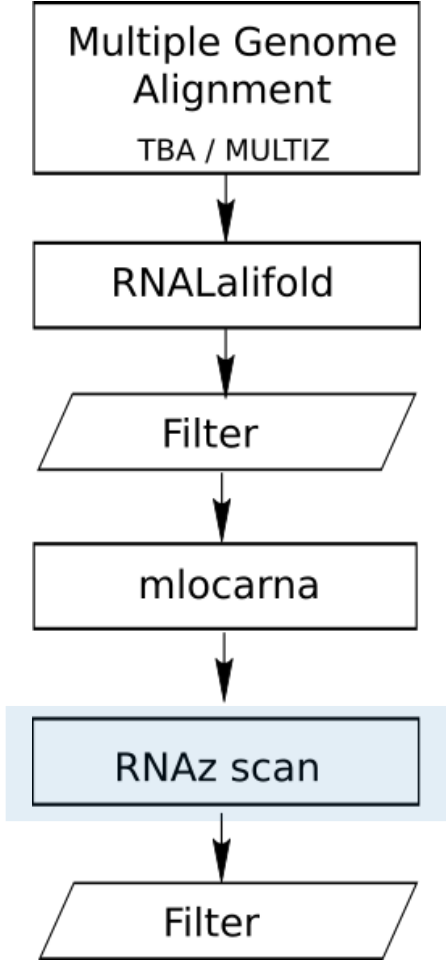
Length

**Forward strand**
**RNAz Locus: 46962** (8%)
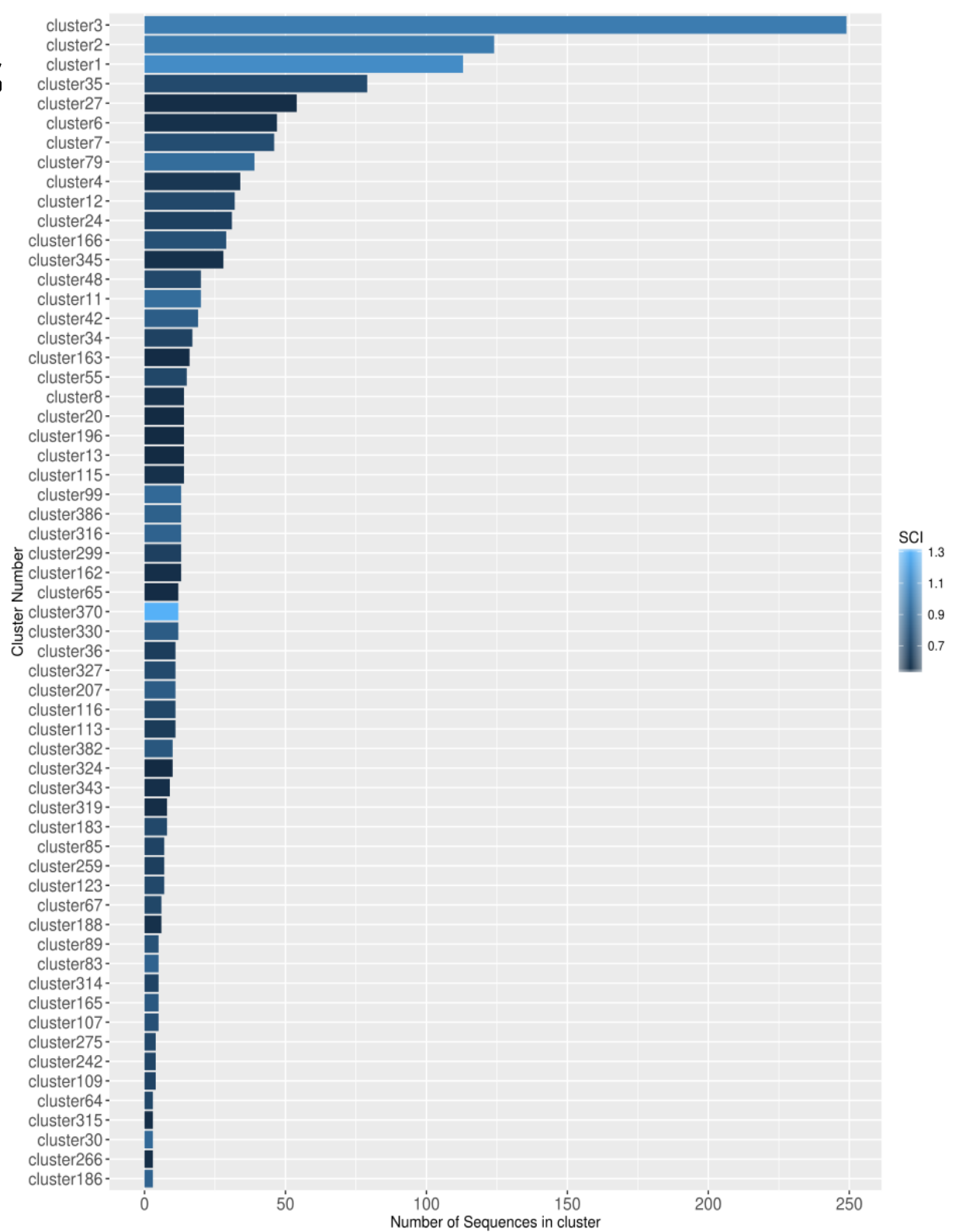>500: 206;  <=500: 46756

TAIR.10−33
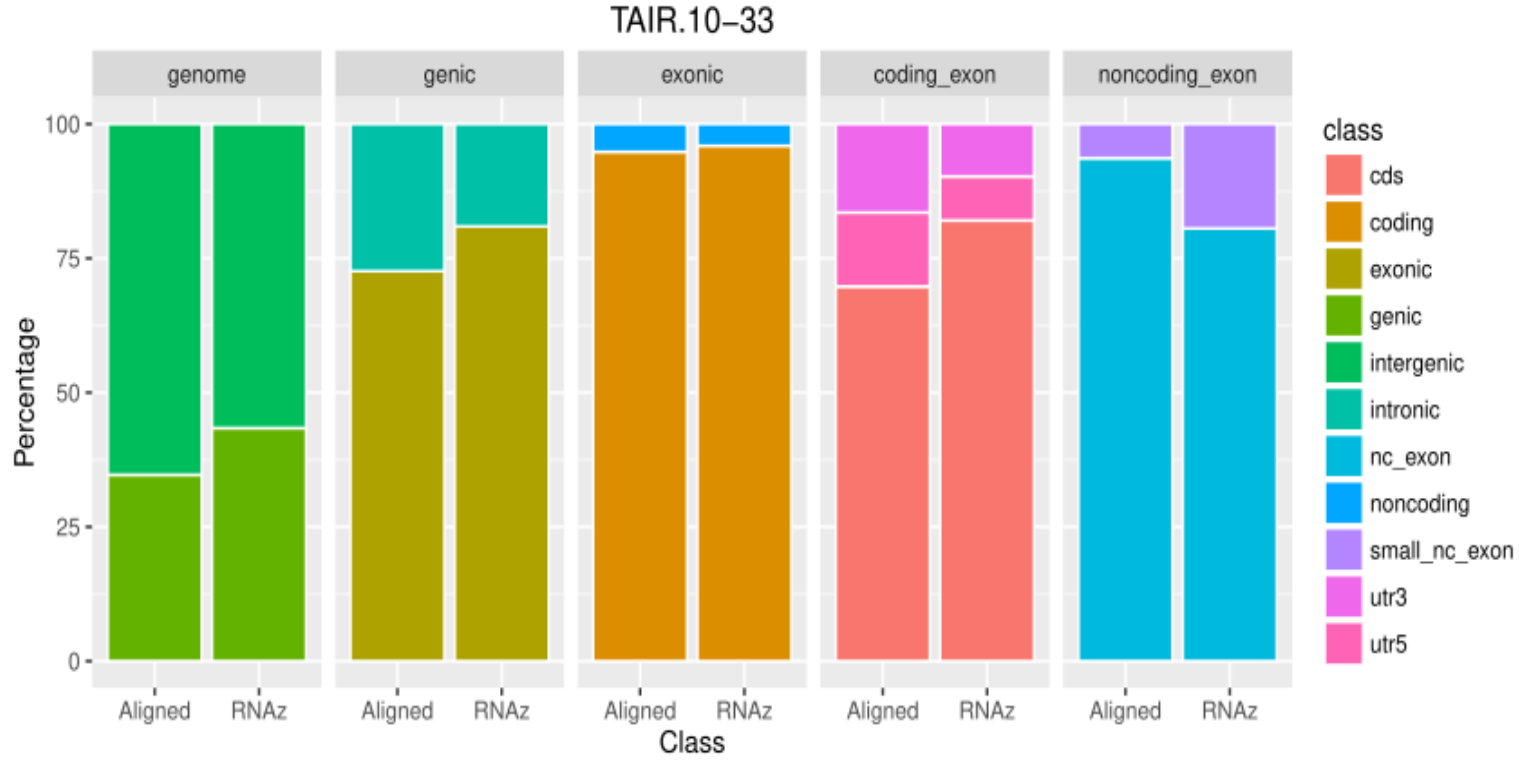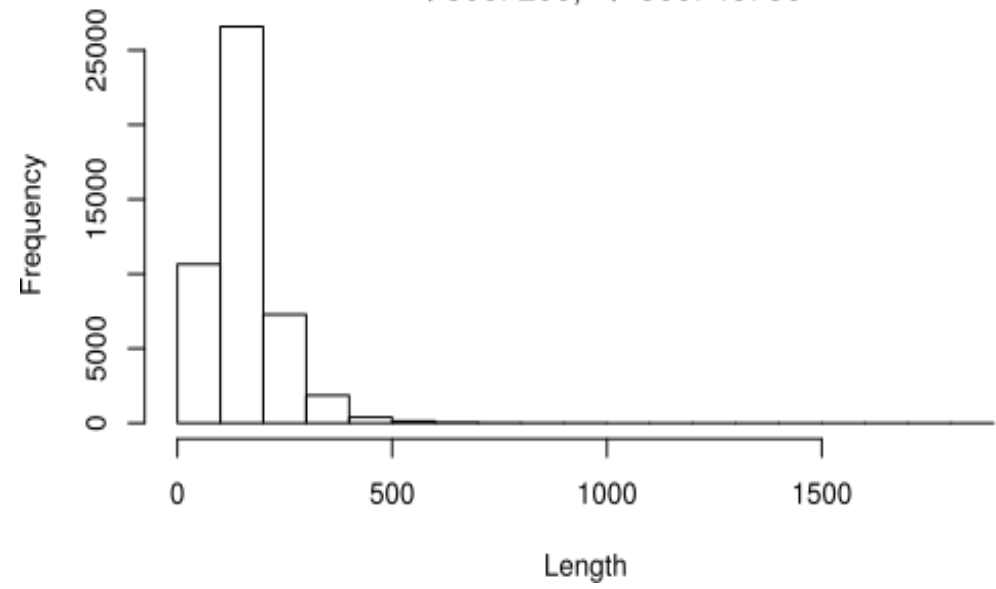
# RNAscClust Overview
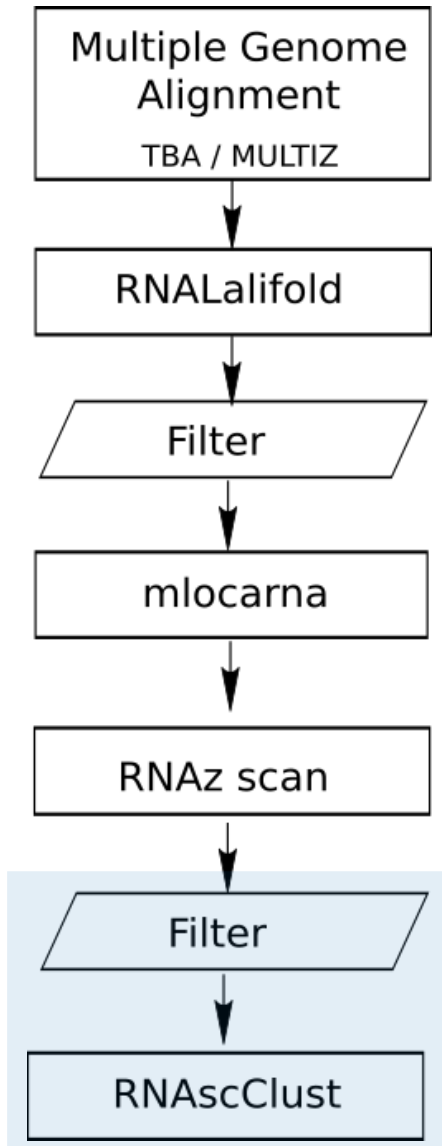


| Input | Avg block length (nt) | Iteration | no. of cluster per iteration | Cluster | MPI (avg) | SCI >0.5 |
|---|---|---|---|---|---|---|
| **98888** | **162** | **5** | **100** | **465** | **71.6** | **323** |
| 98888 | 162 | 5 | 10 | 40 | 80.8 | 33 |
| 98888 | 162 | 15 | 10 | 134 | 78.6 | 108 |
| **63842** | **160** | **5** | **100** | **402** | **38.9** | **66** |
| 63842 | 160 | 5 | 20 | 70 | 49.4 | 27 |

Milad Miladi et al; 2017  Bioinformatics, Volume 33
RNAscClust: clustering RNA sequences using structure conservation and graph based motifs

# RNAscClust Clusters

- Total no. of input alignment blocks **63842**

- Number of sequences in clusters **4364**

# RNAscClust Clusters



- Total no. of input alignment blocks **63842**

- Number of sequences in clusters **4364**

- No. of Sequences : 12

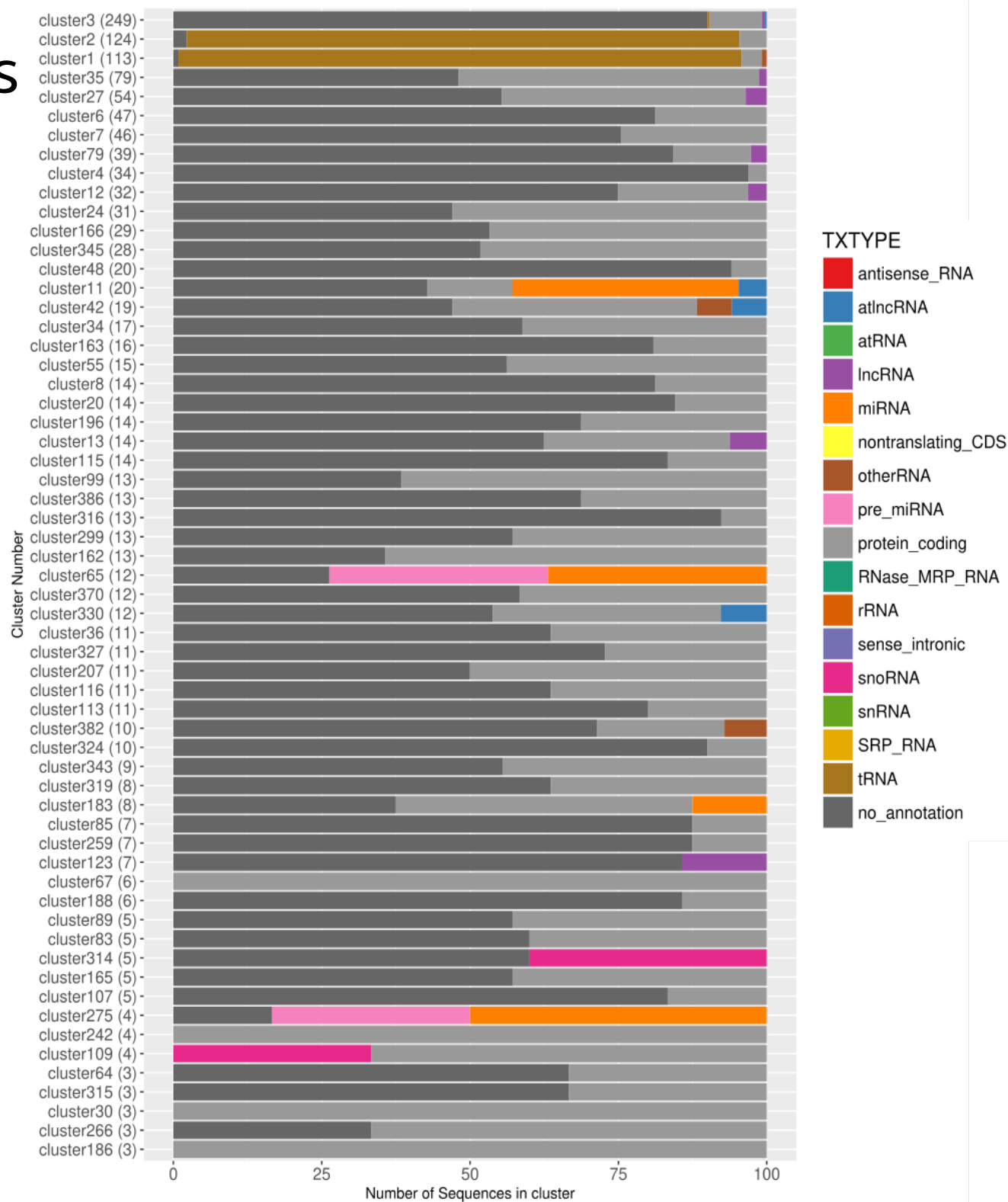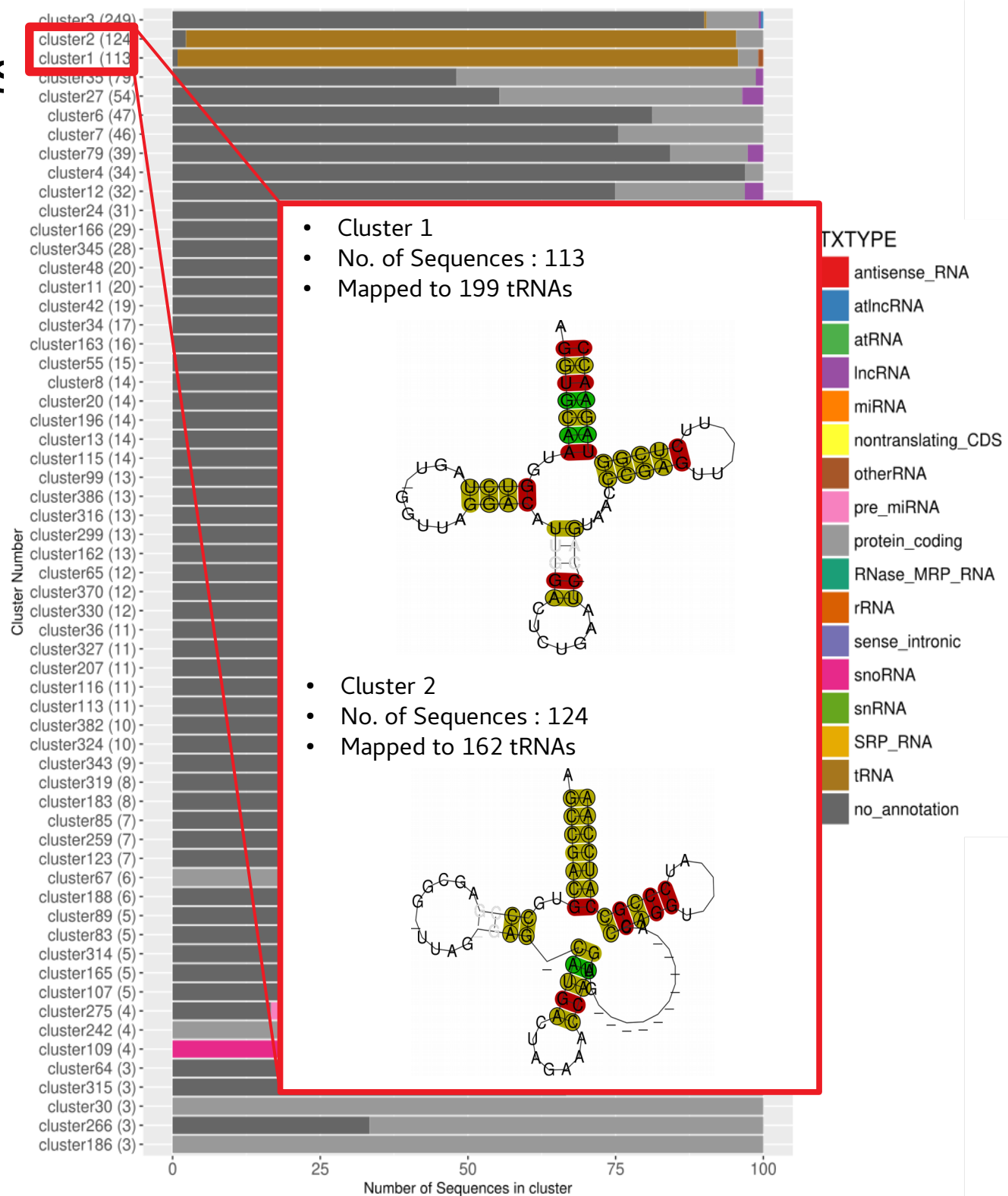- Mapped to UTR3 and CDS of 3 transcripts annotated as Protein coding

# RNAscClust Clusters

- Total no. of input alignment blocks **63842**

- Number of sequences in clusters **4364**

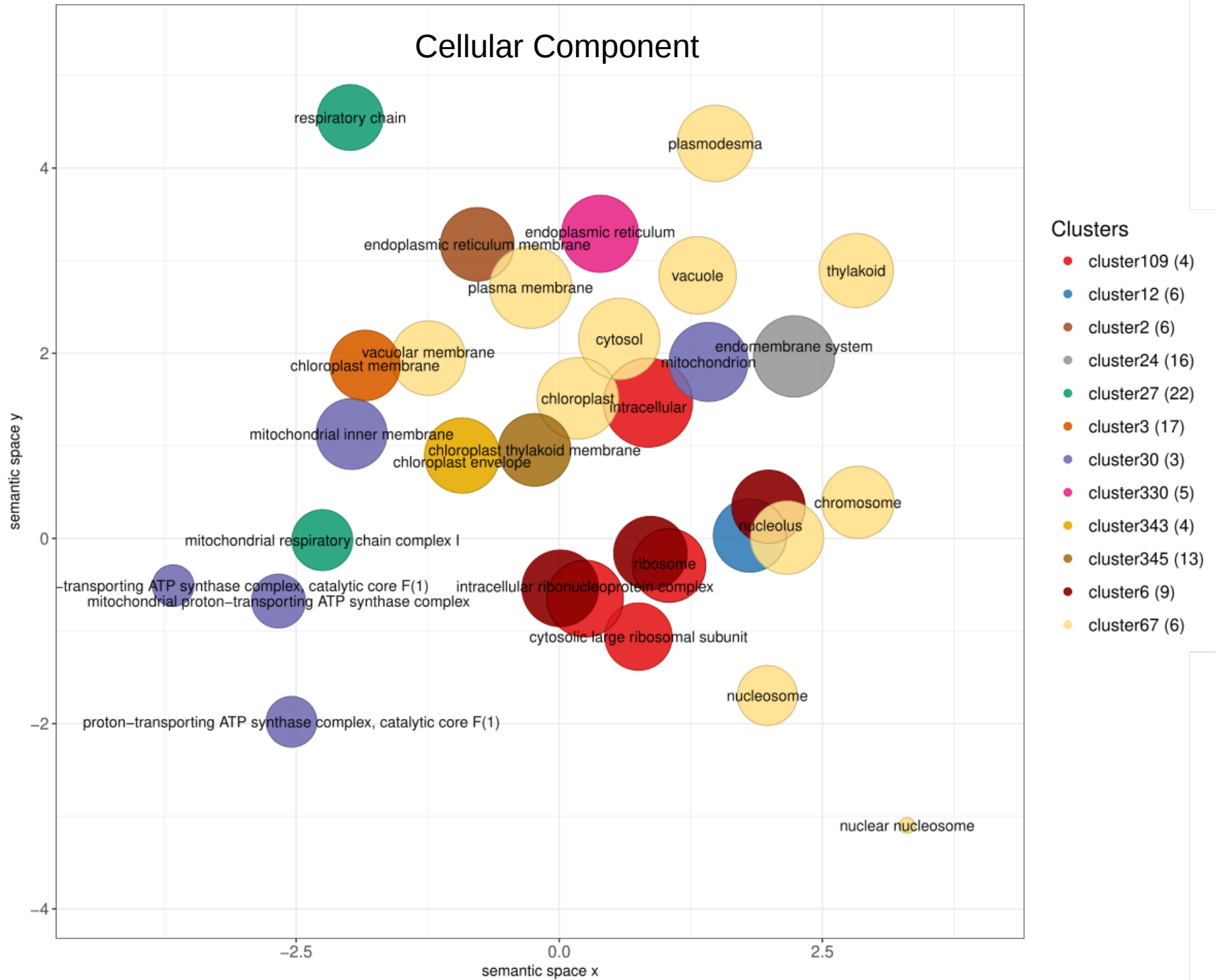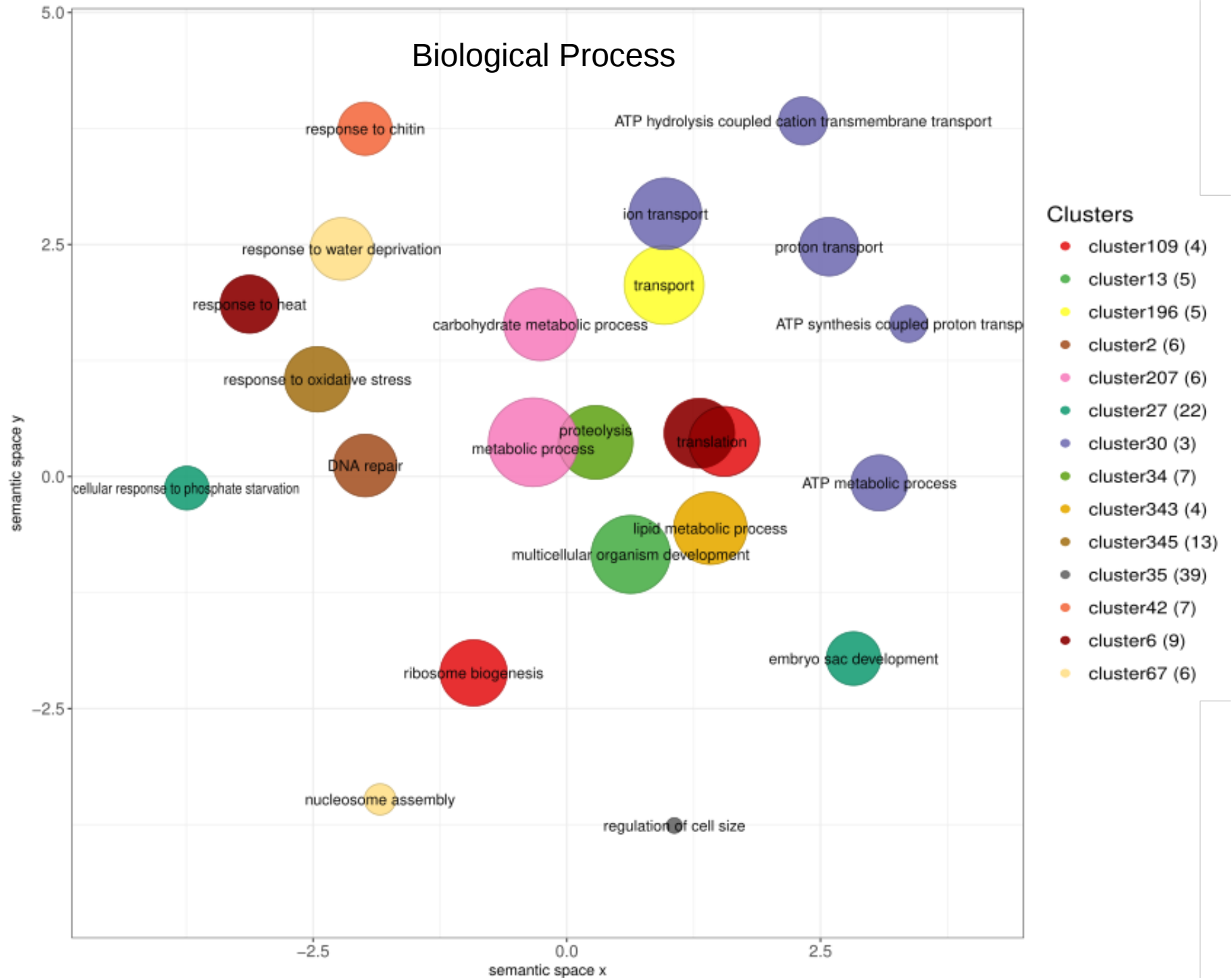- **1701** mapped to current annotations

# RNAscClust Clusters



- Total no. of input alignment blocks **63842**

- Number of sequences in clusters **4364**

- **1701** mapped to current annotations

- Cluster 1
- No. of Sequences : 113
- Mapped to 199 tRNAs

- Cluster 2
- No. of Sequences : 124
- Mapped to 162 tRNAs

TXTYPE
- antisense_RNA
- atlncRNA
- atRNA
- lncRNA
- miRNA
- nontranslating_CDS
- otherRNA
- pre_miRNA
- protein_coding
- RNase_MRP_RNA
- rRNA
- sense_intronic
- snoRNA
- snRNA
- SRP_RNA
- tRNA
- no_annotation

# Annotated transcripts mapped to alignment blocks

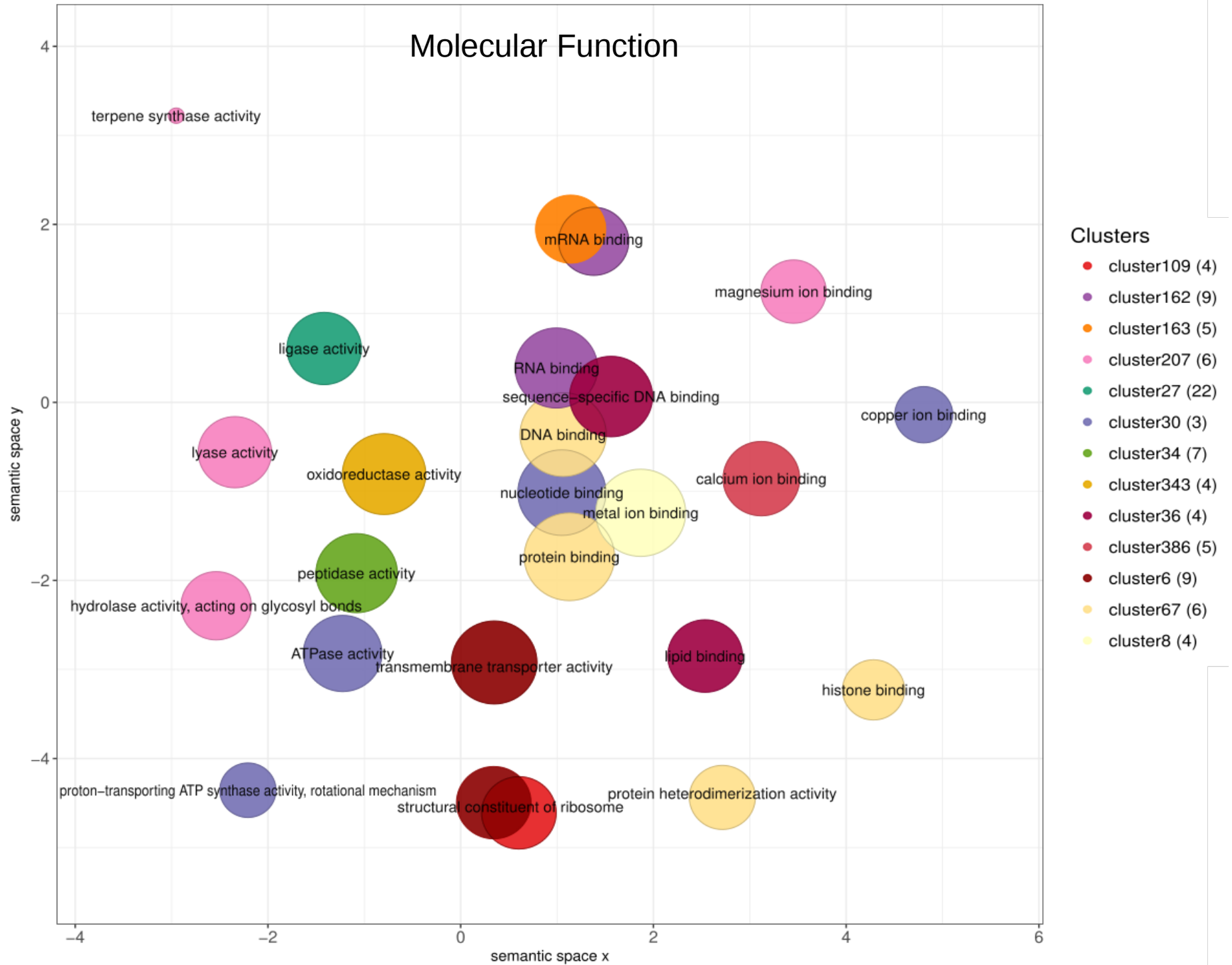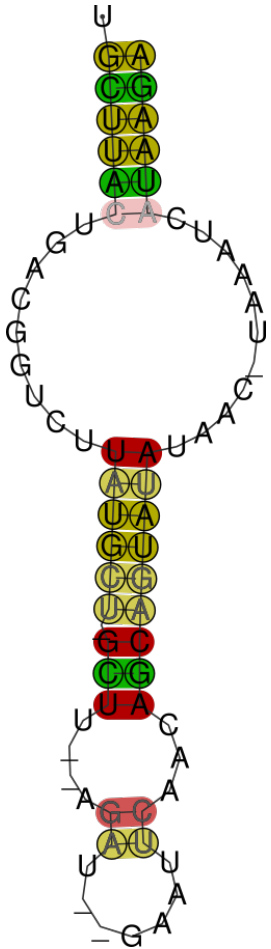| TXTYPE | Annotated TAIR10.37 | MGA | RNALalifold | RNAz | RNAz Filtered | RNAsc Clust |
|---|---|---|---|---|---|---|
| **protein_coding** | **48321** | **45196** | **42061** | **33632** | **28118** | **2563** |
| lncRNA | 2455 | 1558 | 832 | 354 | 201 | 29 |
| atlncRNA | 1424 | 1320 | 1150 | 715 | 475 | 36 |
| **tRNA** | **1033** | **762** | **659** | **552** | **370** | **358** |
| rRNA | 530 | 19 | 11 | 2 | 1 | 0 |
| snoRNA | 421 | 360 | 297 | 223 | 108 | 11 |
| **miRNA** | **325** | **194** | **126** | **108** | **62** | **24** |
| otherRNA | 286 | 177 | 105 | 66 | 36 | 8 |
| snRNA | 154 | 122 | 96 | 70 | 21 | 1 |
| atRNA | 91 | 85 | 75 | 30 | 19 | 0 |
| pre_miRNA | 39 | 33 | 31 | 29 | 15 | 10 |
| nontranslating_CDS | 38 | 33 | 32 | 23 | 23 | 2 |
| sense_intronic | 27 | 0 | 0 | 0 | 0 | 0 |
| SRP_RNA | 10 | 6 | 5 | 5 | 2 | 0 |
| RNase_MRP_RNA | 2 | 1 | 1 | 1 | 0 | 0 |
| antisense_RNA | 1 | 0 | 0 | 0 | 0 | 0 |

# Functional enrichment analysis



Cellular Component

# Functional enrichment analysis



Biological Process

# Functional enrichment analysis



Molecular Function

**Clusters**
- cluster109 (4)
- cluster162 (9)
- cluster163 (5)
- cluster207 (6)
- cluster27 (22)
- cluster30 (3)
- cluster34 (7)
- cluster343 (4)
- cluster36 (4)
- cluster386 (5)
- cluster6 (9)
- cluster67 (6)
- cluster8 (4)

terpene synthase activity

mRNA binding

magnesium ion binding

ligase activity

RNA binding

sequence-specific DNA binding

copper ion binding

DNA binding

lyase activity

oxidoreductase activity

calcium ion binding

nucleotide binding

metal ion binding

protein binding

peptidase activity

hydrolase activity, acting on glycosyl bonds

ATPase activity

transmembrane transporter activity

lipid binding

histone binding

proton-transporting ATP synthase activity, rotational mechanism

protein heterodimerization activity
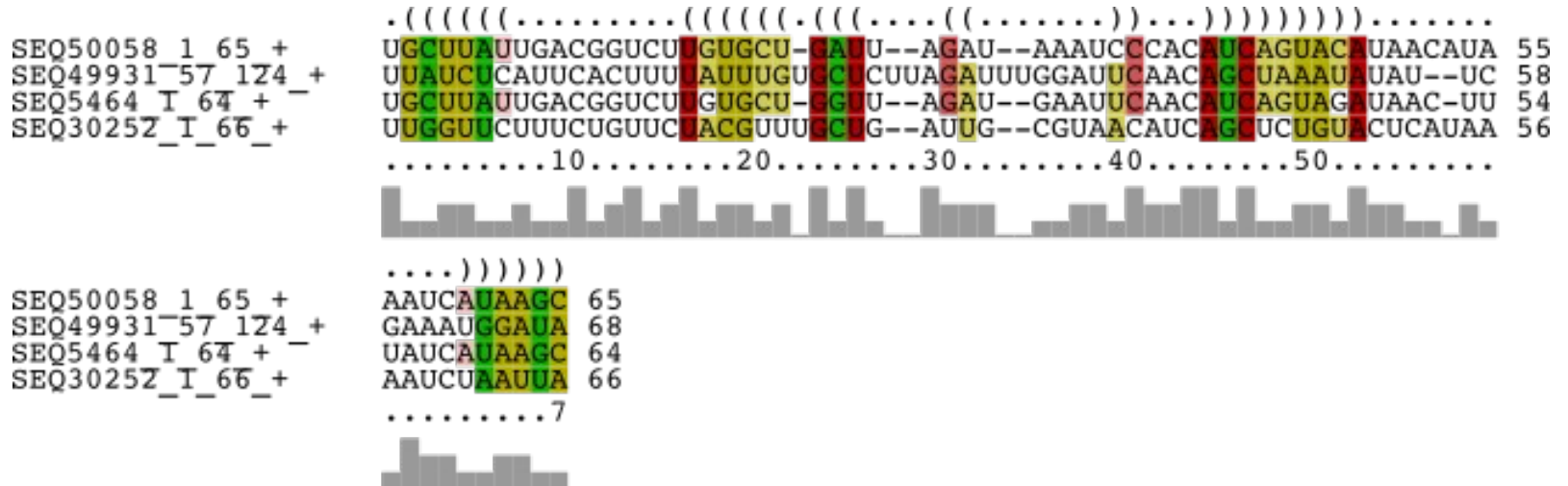
structural constituent of ribosome

semantic space y

semantic space x

# Cluster 109



- No. of Sequences: 4
- Mapped to UTR 3' and introns of four genes annotated as protein coding

```
.((((((...........(((((.(((....((........)))...))))))))).......
SEQ50058_1_65_+   UGCUUAUUGACGGUCUUGUGCU-GAUU--AGAU--AAAUCCCACAUCAGUACAUAACAUA 55
SEQ49931_57_124_+ UUAUCUCAUUCACUUUUAUUUGUGCUCUUAGAUUUGGAUUCAACAGCUAAAUAUAU--UC 58
SEQ5464_I_64_+    UGCUUAUUGACGGUCUUGUGCU-GGUU--AGAU--GAAUUCAACAUCAGUAGAUAAC-UU 54
SEQ30252_I_66_+   UUGGUUCUUUCUGUUCUACGUUUGCUG--AUUG--CGUAACAUCAGCUCUGUACUCAUAA 56
                  ..........10........20........30........40........50.........
```



```
....))))))))
SEQ50058_1_65_+   AAUCAUAAGC 65
SEQ49931_57_124_+ GAAAUGGAUA 68
SEQ5464_I_64_+    UAUCAUAAGC 64
SEQ30252_I_66_+   AAUCUAAUUA 66
                  ..........7
```
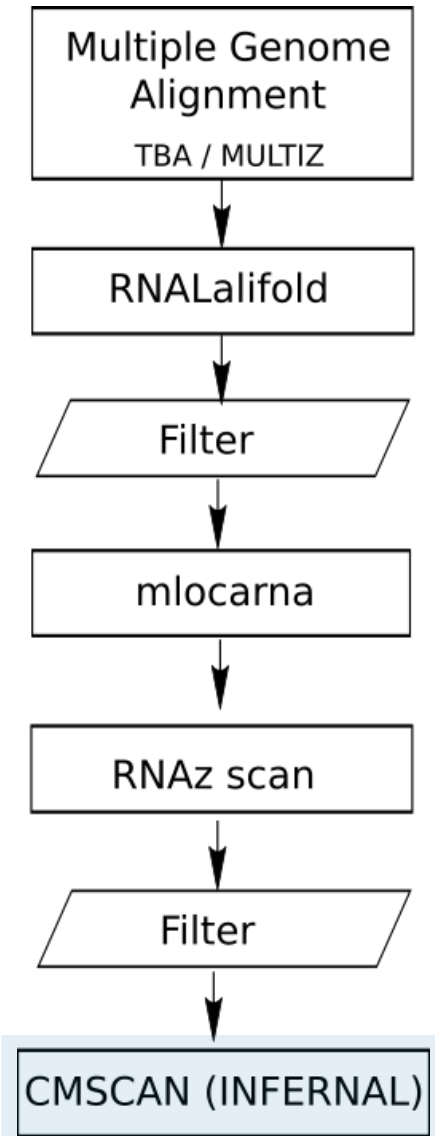
# Cluster 30

- No. of Sequences: 3
- Mapped to CDS and introns of three genes annotated as protein coding
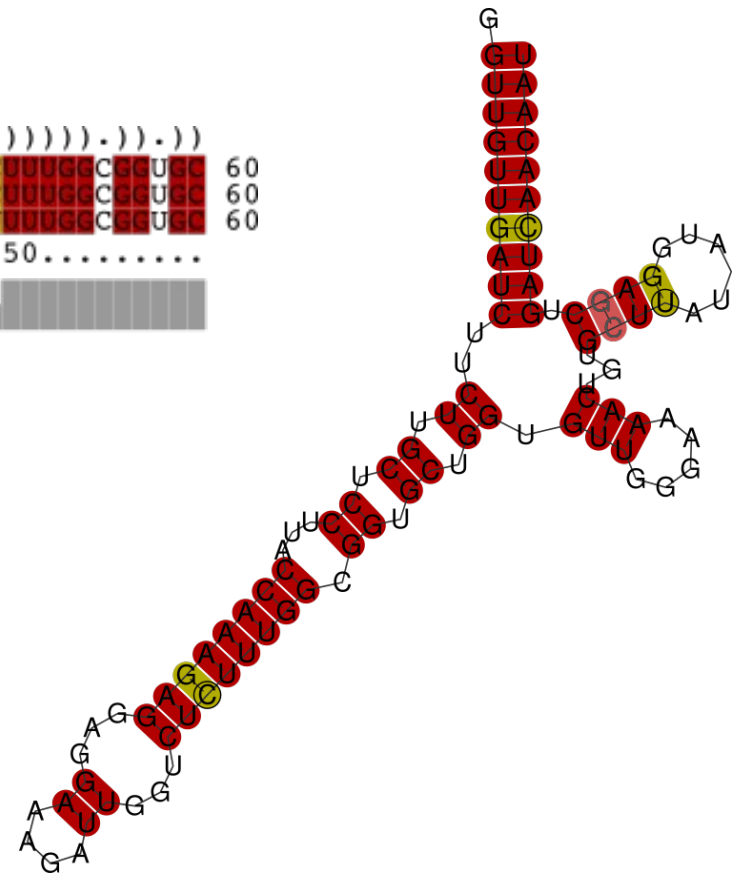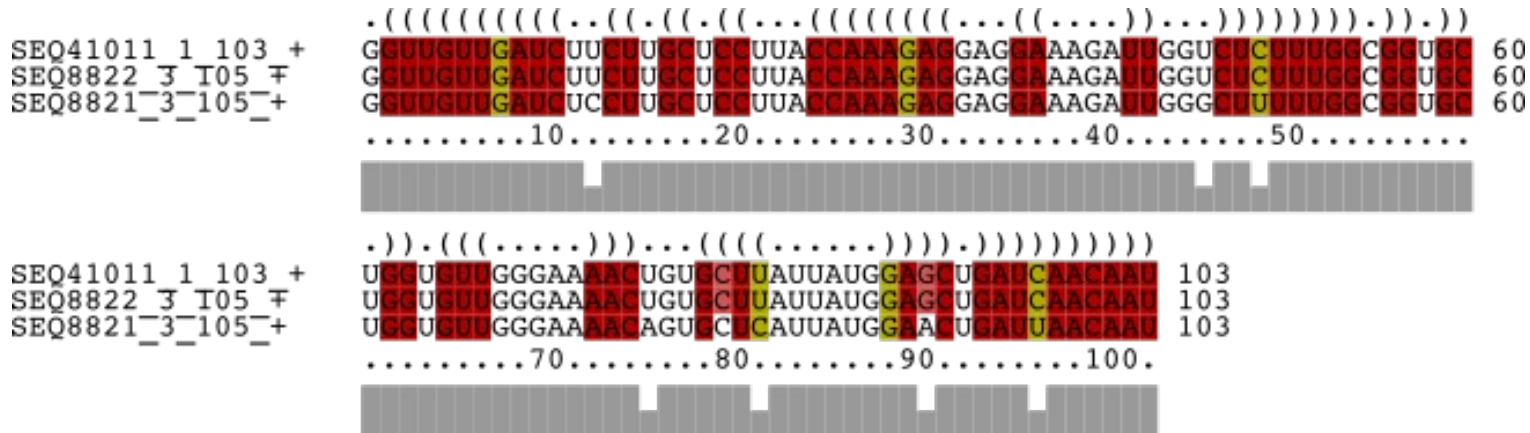
# CMSCAN



- Total no. of input alignment blocks: **63842**

- No. of sequences in clusters outside input alignments: **119788**

- No. of sequences in clusters: **7767**

- **3759** mapped to the annotations

# Annotation mapped to RNAscClust and CMSCAN

| TXTYPE | Annotated TAIR10.37 | RNAscClust | CMSCAN | CMSCAN outside input alignments |
|---|---|---|---|---|
| protein_coding | 48321 | 2563 | 2496 | 26101 |
| lncRNA | 2455 | 29 | 7 | 231 |
| atlncRNA | 1424 | 36 | 54 | 522 |
| tRNA | 1033 | 358 | 248 | 956 |
| rRNA | 530 | 0 | 19 | 476 |
| snoRNA | 421 | 11 | 15 | 127 |
| miRNA | 325 | 24 | 15 | 95 |
| otherRNA | 286 | 8 | 5 | 45 |
| snRNA | 154 | 1 | 5 | 109 |
| atRNA | 91 | 0 | 1 | 18 |
| pre_miRNA | 39 | 10 | 5 | 26 |
| nontranslating_CDS | 38 | 2 | 3 | 20 |
| sense_intronic | 27 | 0 | 0 | 0 |
| SRP_RNA | 10 | 0 | 0 | 10 |
| RNase_MRP_RNA | 2 | 0 | 0 | 0 |
| antisense_RNA | 1 | 0 | 0 | 0 |

# Next steps..

- Improvements in the filtering steps

- Tweak GraphClust parameters in RNAscClust to achieve better clustering

- Include the hits from CMSCAN for the clustering

# Acknowledgments