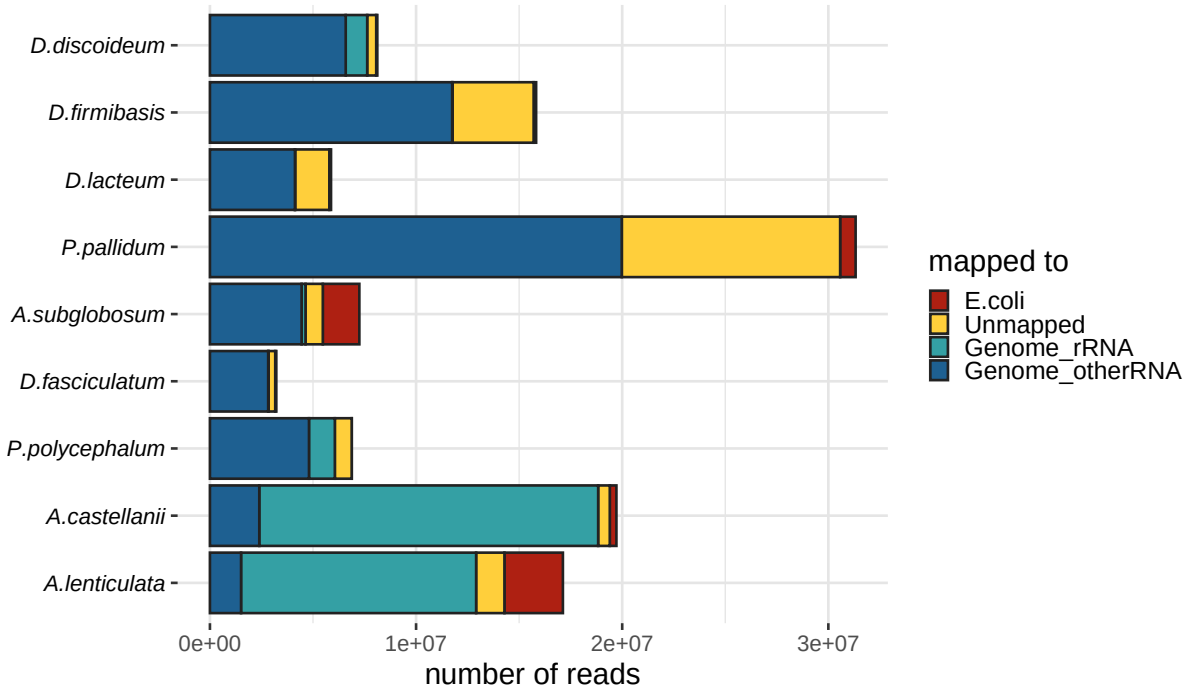
Analysis of microRNAs identified in Amoebzoa

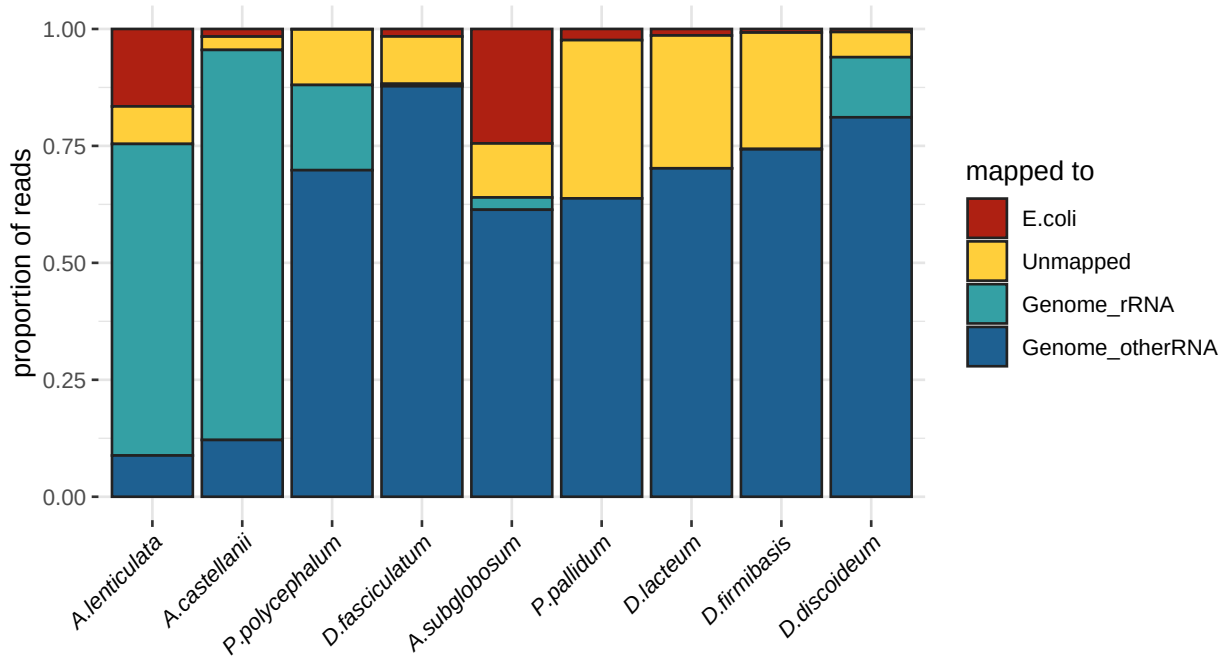
Bart Edelbroek

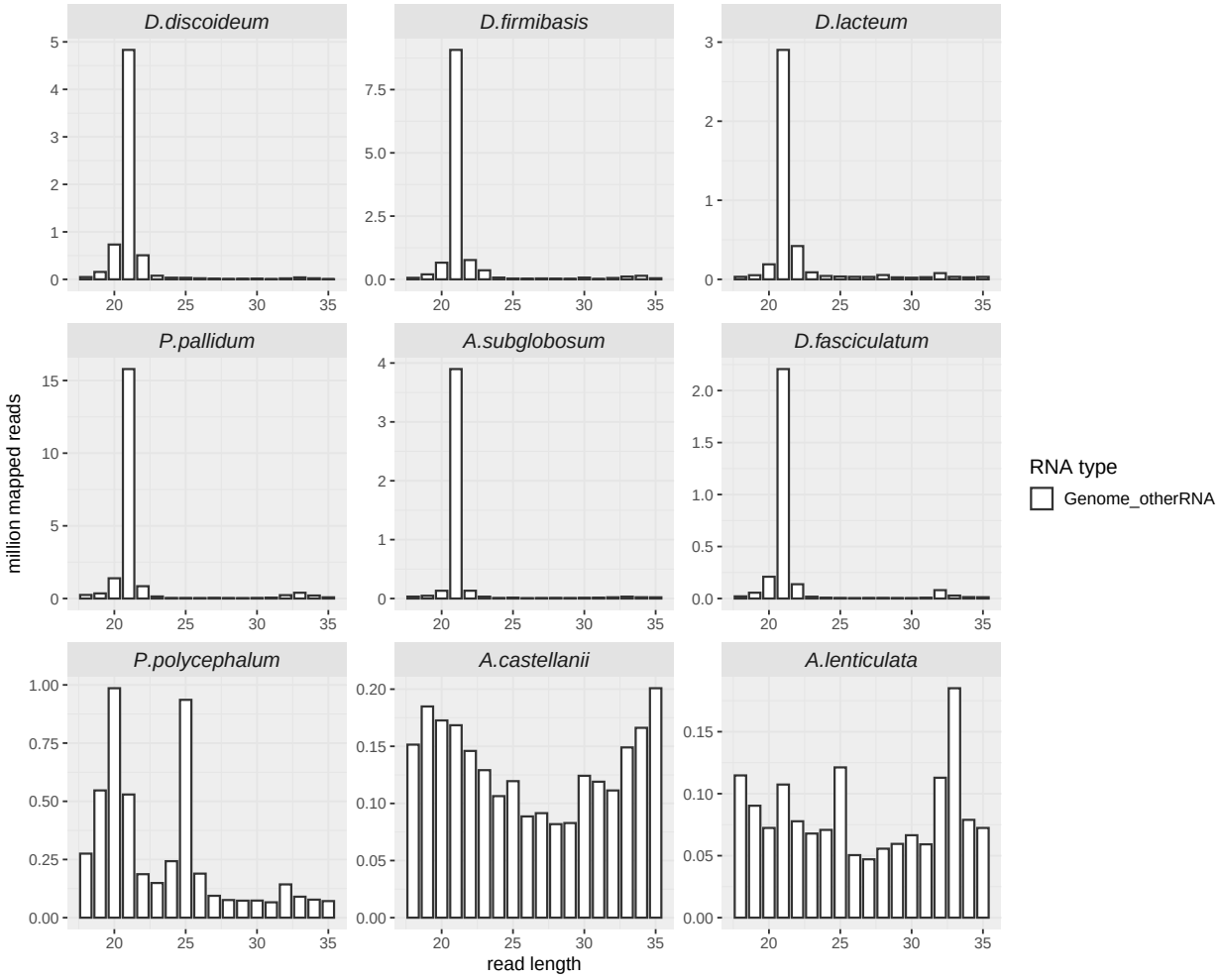
11/24/2021

miRNA identification

Quantified the number of reads mapping to different parts of the amoeba genome, E. coli, or unmapped. Plotted the length profile of the reads mapping to the genome but not rRNA. The data_in was generated using mapping_percentages.sh and extract_length_profiles.sh helper scripts.




proportion of reads
1.00
0.75
0.50
0.25
0.00
A.lenticulata
A.castellanii
P.polycephalum
D.fasciculatum
A.subglobosum
P.pallidum
D.lacteum
D.firmibasis
D.discoideum
mapped to
E.coli
Unmapped
Genome_rRNA
Genome_otherRNA

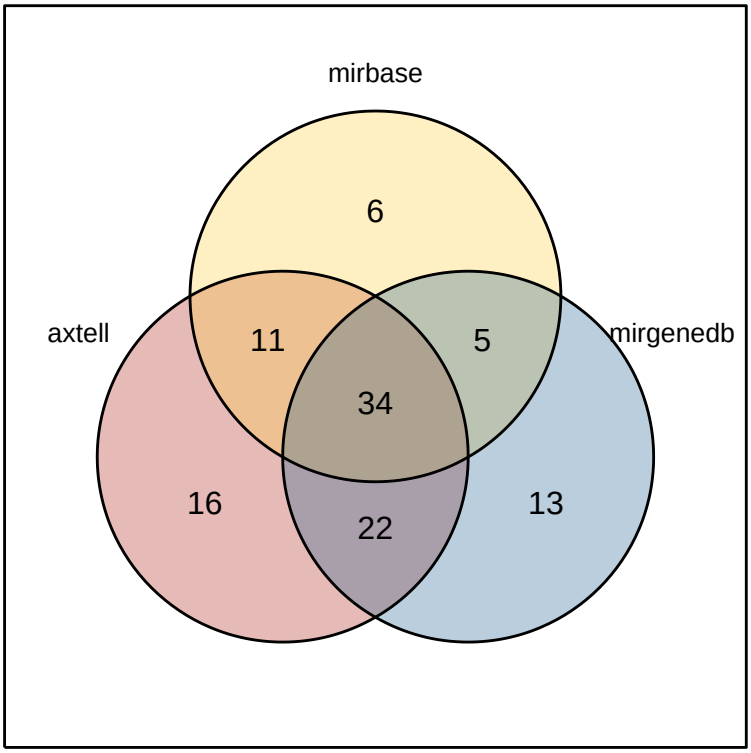

D.discoideum
D.firmibasis
D.lacteum
P.pallidum
A.subglobosum
D.fasciculatum
P.polycephalum
A.castellanii
A.lenticulata
million mapped reads
read length
RNA type
Genome_otherRNA

After running the miRNA curation python script, the characteristics of each of the miRNA candidates is summarized in data_in/species/combined_analysis.tsv. Each of the miRNA candidates is analyzed by the criteria put forth in Axtell & Meyers (2018), Fromm et al. (2022), and Kozomara et al. (2014 & 2019), as summarized in figure S3. Following manual curation, a final list of high-confidence miRNAs is identified.

```
## # A tibble: 70 x 5
##    miRNA.ID             cluster          axtell_pass mirbase_pass mirgenedb_pass
##    <chr>                <chr>            <chr>       <chr>        <chr>
##  1 acas_Cluster_5138(+) acas_Cluster_51~ TRUE        TRUE         TRUE
##  2 alen_Cluster_556(+)  alen_Cluster_55~ 50% precis~ TRUE         TRUE
##  3 alen_Cluster_1355(+) alen_Cluster_13~ TRUE        3 miR* reads TRUE
##  4 alen_Cluster_2130(+) alen_Cluster_21~ 50% precis~ TRUE         TRUE
##  5 alen_Cluster_2233(-) alen_Cluster_22~ TRUE        TRUE         TRUE
##  6 alen_Cluster_4026(-) alen_Cluster_40~ TRUE        TRUE         TRUE
##  7 asub_Cluster_2821(-) asub_Cluster_28~ TRUE        4 miR* reads TRUE
##  8 asub_Cluster_3219(+) asub_Cluster_32~ TRUE        6 miR* reads TRUE
##  9 asub_Cluster_3317(-) asub_Cluster_33~ TRUE        5 miR* reads TRUE
## 10 asub_Cluster_3339(-) asub_Cluster_33~ TRUE        7 miR* reads TRUE
## # i 60 more rows
```

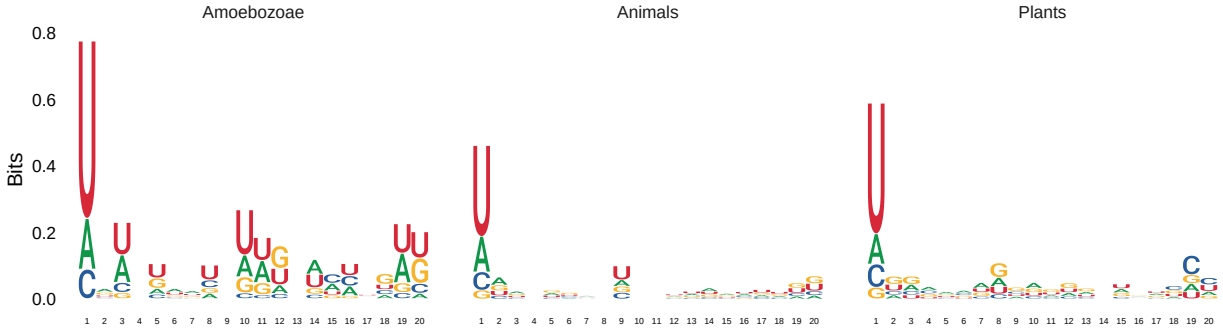
The miRNA sequences of the confirmed candidates are output in fasta files, as well as in a table. The genomic locations of the miRNAs are added and used to generate a gff file to annotated the genomes.

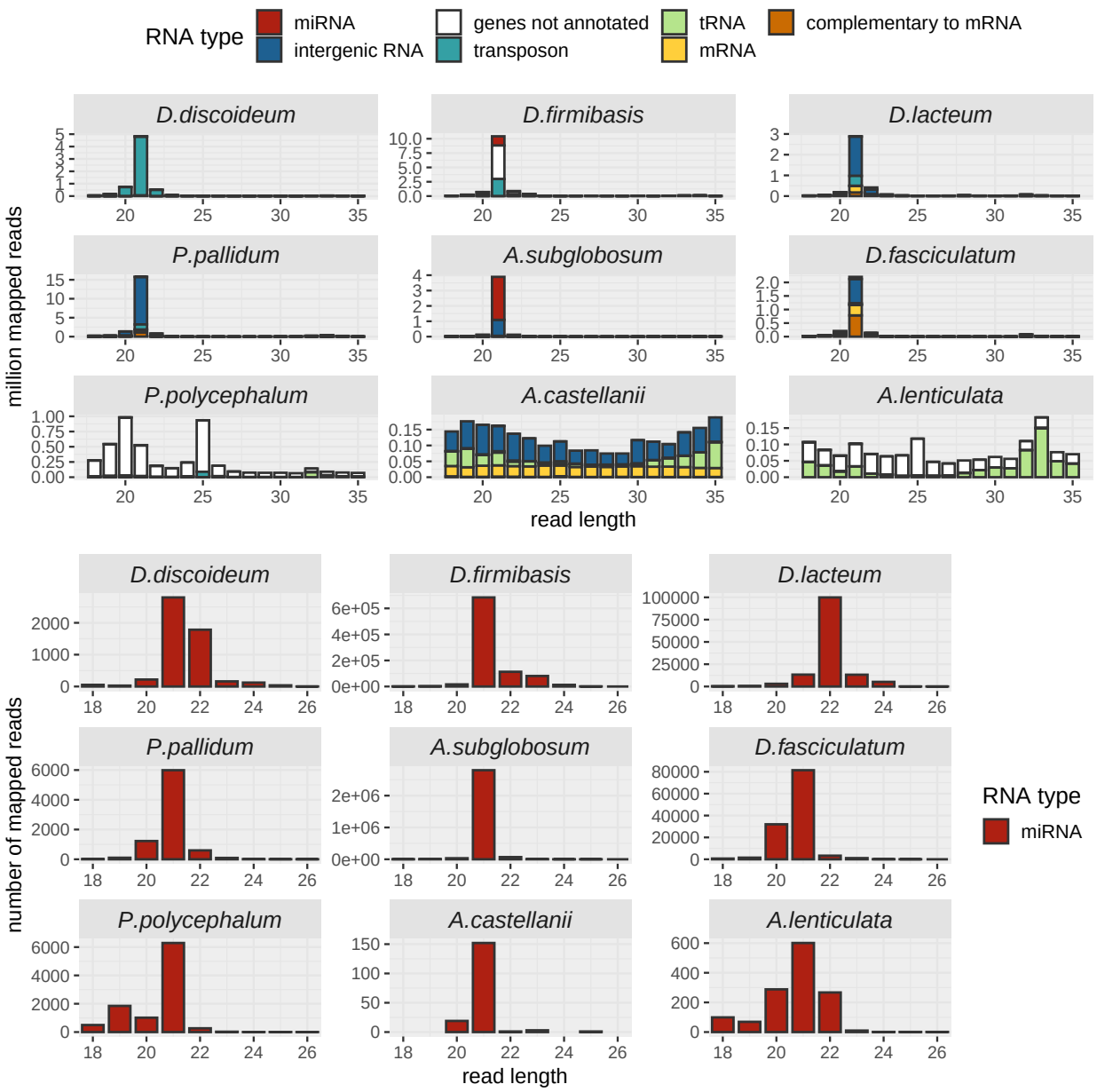
```
## # A tibble: 10 x 5
##    species  combined axtell_pass mirbase_pass mirgenedb_pass
##    <chr>       <int>       <int>        <int>          <int>
##  1 acas            1           1            1              1
##  2 alen            5           4            5              5
##  3 asub           22          23           17             24
##  4 ddis            8          10            8              6
##  5 dfas            4           9            4              4
##  6 dfir            5           9            5              7
##  7 dfir_new        9          13            9             11
##  8 dlac            2           2            1              2
##  9 ppal            4           7            3              5
## 10 ppol            7          10            6             11

## pdf
##   2
```

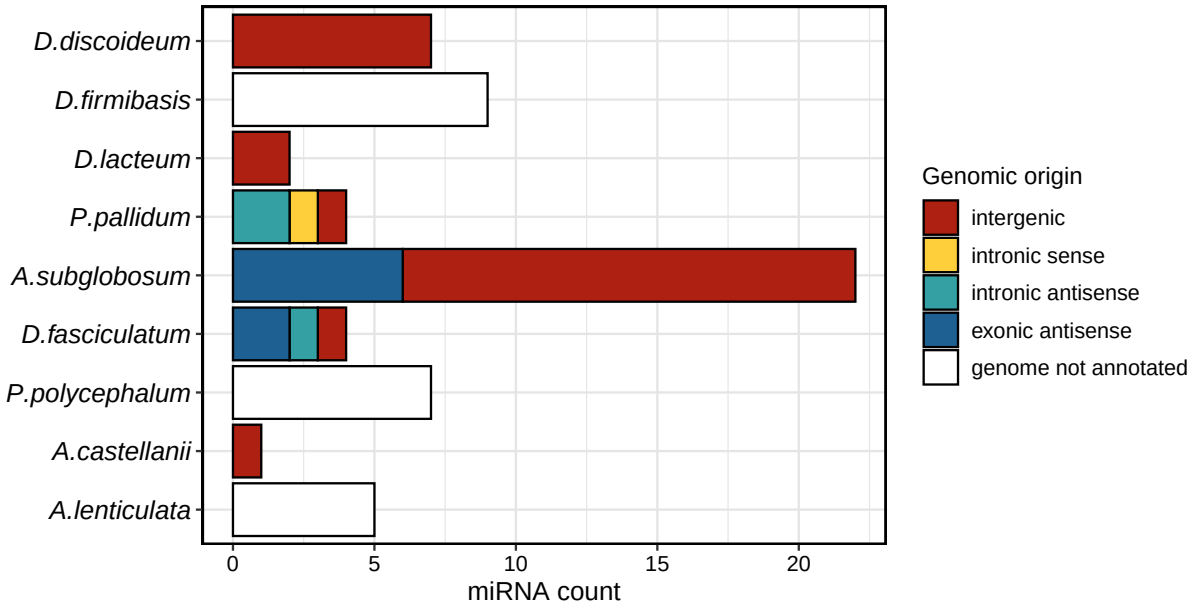


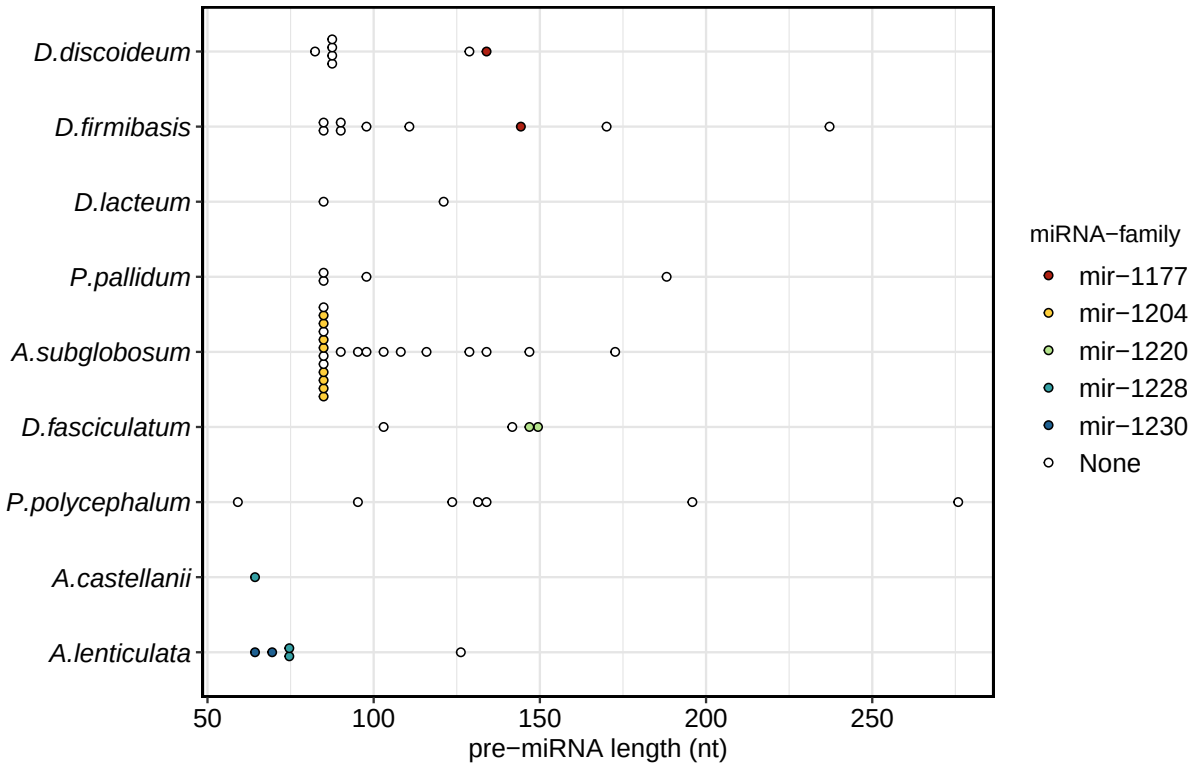
miRNA characteristics

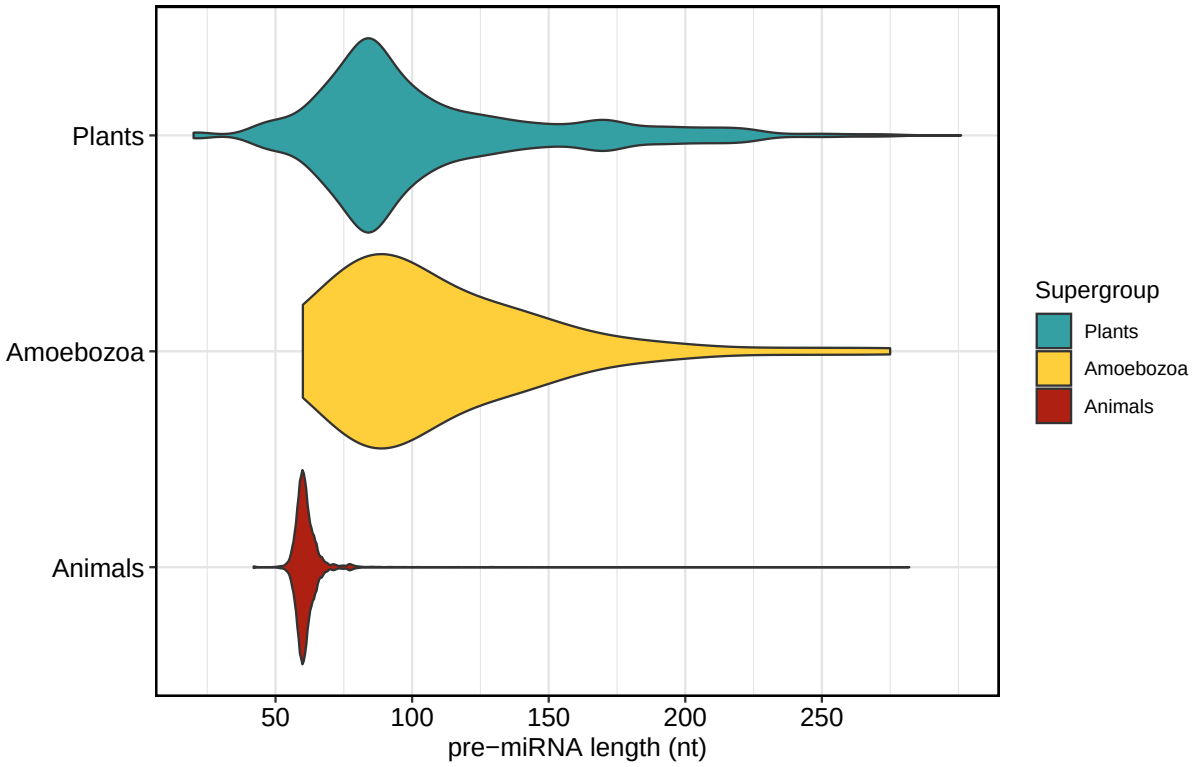
Following identification of the miRNAs, the characteristics of the miRNAs are analyzed. For those characteristics where it is relevant, the comparison is made with Plants and Animals. Plant and Animal miRNA sequences were accesses from PmiREN 2.0 and MirGeneDB 2.1 respectively.

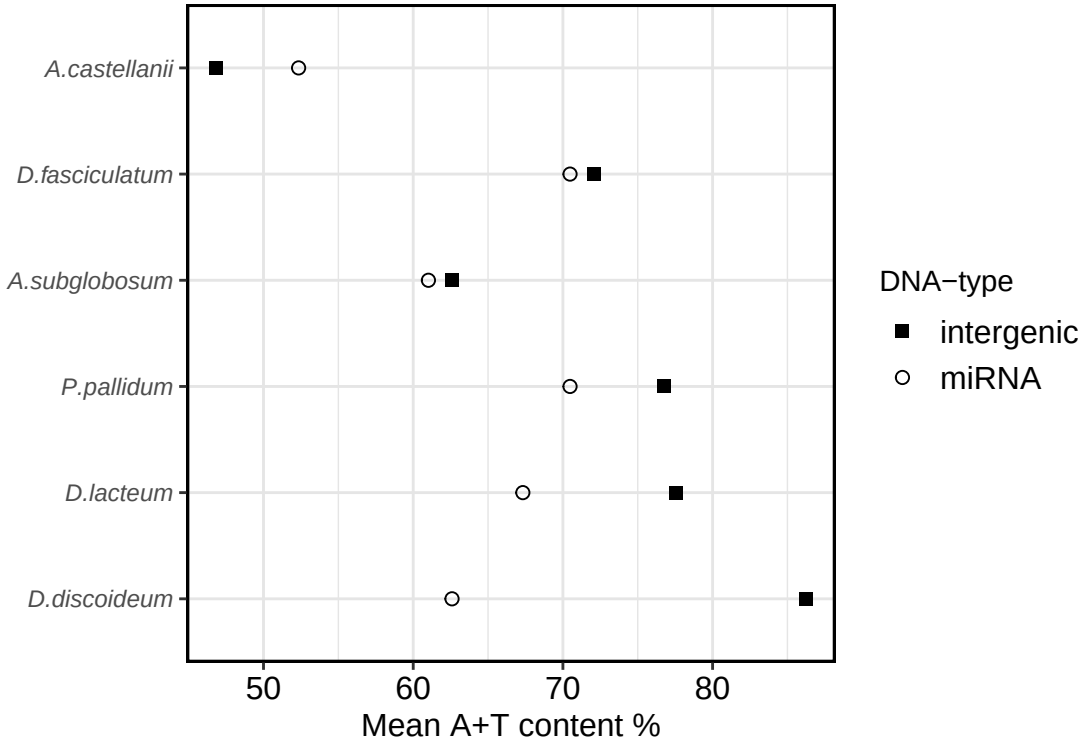


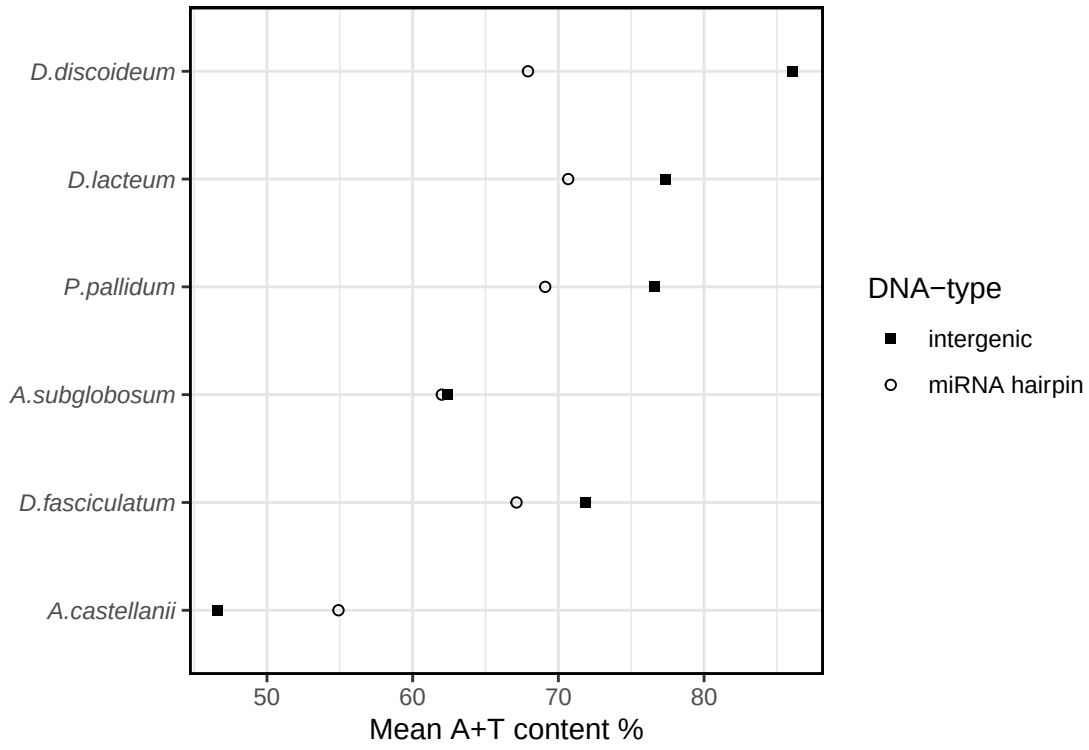

RNA type
miRNA
intergenic RNA
genes not annotated
transposon
tRNA
mRNA
complementary to mRNA
D.discoideum
D.firmibasis
D.lacteum
P.pallidum
A.subglobosum
D.fasciculatum
P.polycephalum
A.castellanii
A.lenticulata
million mapped reads
read length
number of mapped reads
RNA type
miRNA


D.discoideum
D.firmibasis
D.lacteum
P.pallidum
A.subglobosum
D.fasciculatum
P.polycephalum
A.castellanii
A.lenticulata
0
5
10
15
20
miRNA count
Genomic origin
intergenic
intronic sense
intronic antisense
exonic antisense
genome not annotated


D.discoideum
D.firmibasis
D.lacteum
P.pallidum
A.subglobosum
D.fasciculatum
P.polycephalum
A.castellanii
A.lenticulata
50
100
150
200
250
pre−miRNA length (nt)
miRNA−family
mir−1177
mir−1204
mir−1220
mir−1228
mir−1230
None

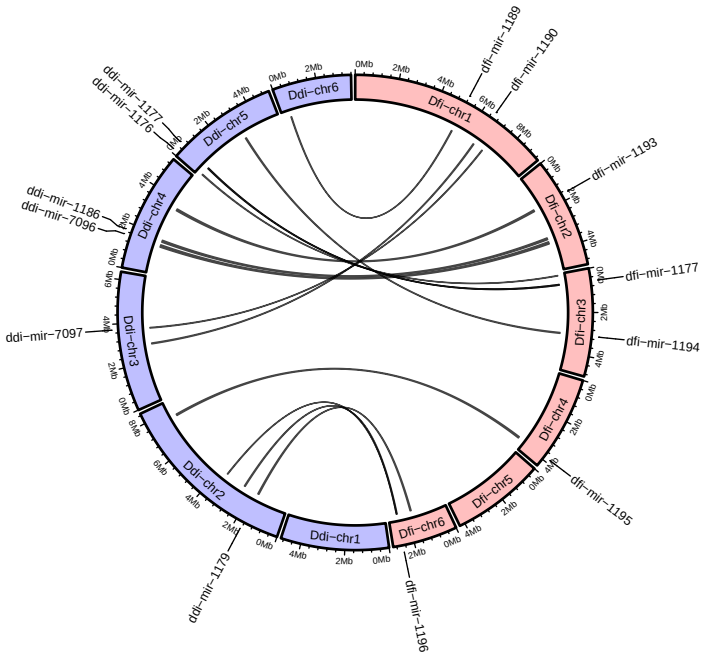

Plants
Amoebozoa
Animals
50
100
150
200
250
pre−miRNA length (nt)
Supergroup
Plants
Amoebozoa
Animals


A.castellanii
D.fasciculatum
A.subglobosum
P.pallidum
D.lacteum
D.discoideum
50
60
70
80
Mean A+T content %
DNA−type
intergenic
miRNA



Synteny of miRNAs in D. firmibasis and D. discoideum

Using Satsuma2, the synteny blocks between D. firmibasis and D. discoideum were identified, with the results summarised in data_in/satsuma_synteny.out. Here, the synteny blocks are linked together if they are within 5000nt of eachother. The first circos plot shows all the synteny that was identified between the two genomes; the second plot shows only those synteny blocks that contain a miRNA on either genome, with the label identifying which miRNA is on the region.




```
## null device
##           1
```

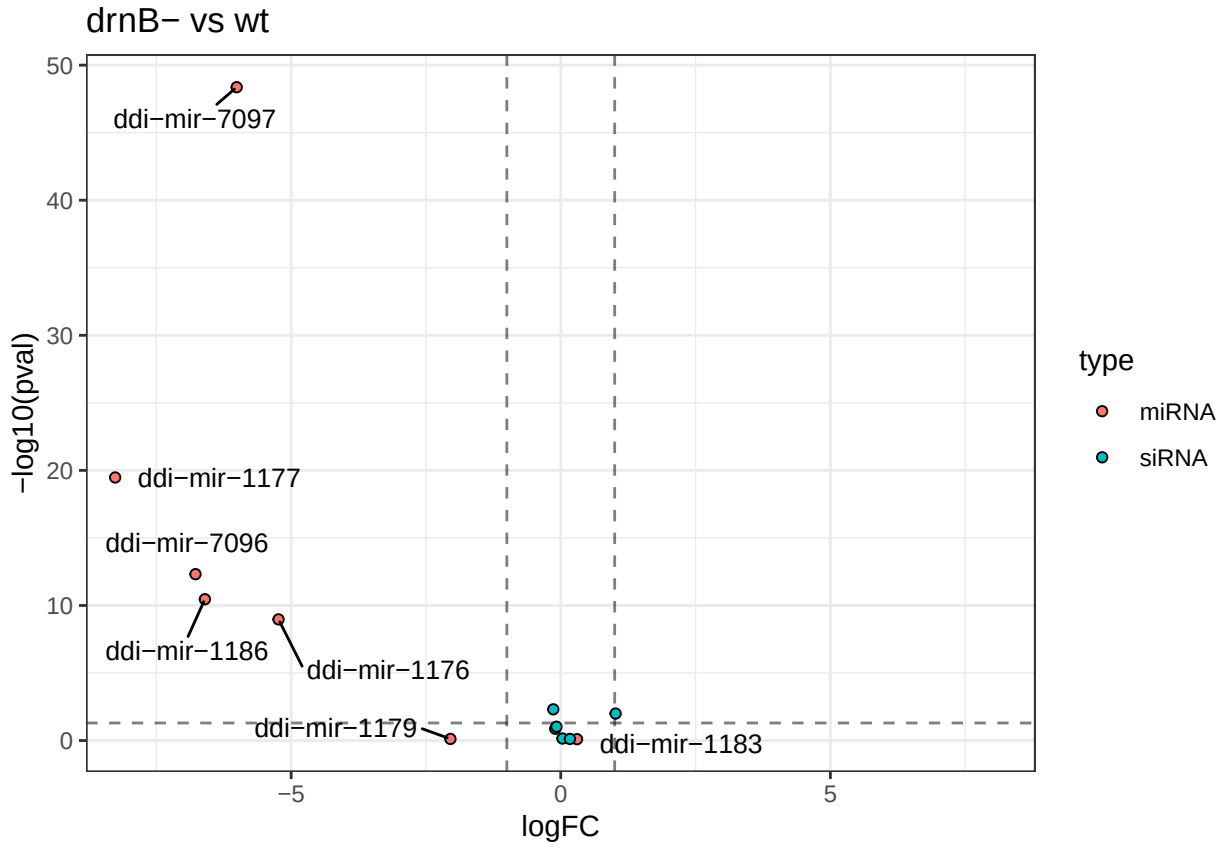
Effects of dicer on miRNA abundance and growth speed

Small RNA sequencing was performed from wildtype cells at vegetative stage (BioSample SAMN35084095), wildtype cells at slug stage (BioSample SAMN35084096), drnB- knockout cells at vegetative stage (BioSample SAMN35084097), and drnB- knockout cells at slug stage (BioSample SAMN35084098). Following mapping of sRNAs, the counts mapping to miRNAs and siRNAs were quantified (data_in/miRNA_counts.txt, data_in/siRNA_counts.txt). Here, DGE of the miRNAs and siRNAs is performed.

```
##  [1] "ddi-mir-1176"   "ddi-mir-1177"   "ddi-mir-1186"   "ddi-mir-7096"
##  [5] "ddi-mir-7097"   "TDD-5"          "EnSpm-1N1_DDi"  "DIRS1"
```

```
##  [9] "TDD3"           "TRE3C"          "Dl_TRE3-A"      "RandI-1_ACas"
## [13] "Gypsy-3_PPP-I"  "TDD4"           "DGLT-A1_I"      "piggyBacA-1_DD"
```

```
## pdf
##   2
```



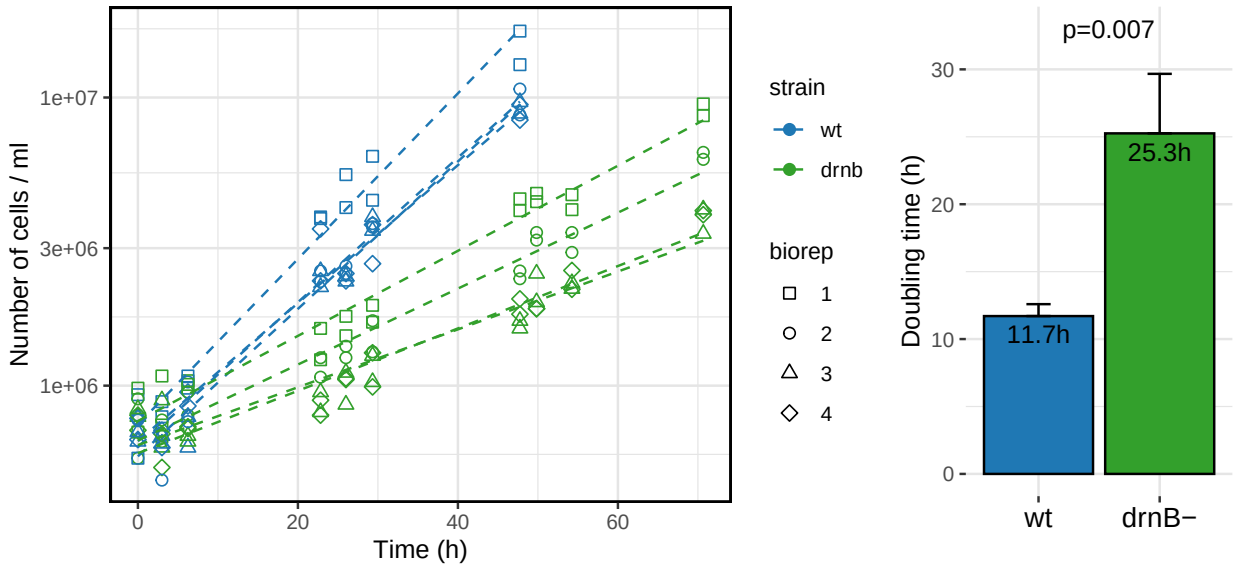
Growth of drnB- and wildtype strains was quantified and is plotted as growthcurves, and barplots with doubling time in hours.

```
##   doubling_time strain
## 1         20.39   drnb
## 2         22.84   drnb
## 3         29.99   drnb
## 4          27.8   drnb
## 5         10.48     wt
## 6            12     wt
## 7         11.77     wt
## 8         12.57     wt
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 24 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 24 rows containing missing values (`geom_point()`).
```


Number of cells / ml
1e+07
3e+06
1e+06
0
20
40
60
Time (h)
strain
wt
drnb
biorep
1
2
3
4
p=0.007
Doubling time (h)
30
20
10
0
11.7h
25.3h
wt
drnB−