Annotation: RepeatMasker vs. MITETracker

July 31

Introduction

Miniature inverted-repeat Transposable Elements (MITEs) are non-autonomous class II transposons found in high copy numbers of eukaryotes and some bacteria. MITE subfamilies can arise from autonomous elements, and based on sequences similarities of TIRs with known TE families, can be classified into these families.

Here, MITETracker was used to retrieve inverted repeat sequences from the typhina genomes. Following this, the detected MITEs in each genome were blasted against consensus sequences of TE families generated using RepeatMasker. The blastn generated file will be used for R analyses to determine the percentage of RepeatMaskers 'Unknown' TE family that can be classified as MITEs.

MITETracker

example of code used to run *Mite tracker*, where data used is our epichloe genomes and `Epo_MITETracker` is the assigned file name

```
python3 -m MITETracker -g raw_data/Etp76_Epichloe_typhina_var_poae_NFe76_38327242_v1.fna
-w 10 -j Epo_MITETracker
```

Blastn

example of code used to make *Exx_MITES_vs_RM.tsv* files: data used is families.nr.fasta from the MITETracker results, and the consensus family data generated from repeatMasker

```
blastn -query Mite_Tracker_results/Epo_MITETracker/families_nr.fasta
-subject consensi/Epo_TE_fam.fa -outfmt 6 > results/Epo_MITES_vs_RM.tsv
```

\section{R analyses using Blastn output)

```
Epo_MT_RM <- read.table("../../results/Epo_MITES_vs_RM.tsv", comment.char = "?", sep = "\t")
Ecl_MT_RM <- read.table("../../results/Ecl_MITES_vs_RM.tsv", comment.char = "?", sep = "\t")
Ety_MT_RM <- read.table("../../results/Ety_MITES_vs_RM.tsv", comment.char = "?", sep = "\t")
head(Ecl_MT_RM)
```

```
##                                                V1
## 1   MITE_T_10|Ecl_1605_22_1|871598|872054|GC|71|F3
## 2   MITE_T_10|Ecl_1605_22_1|871598|872054|GC|71|F3
## 3   MITE_T_10|Ecl_1605_22_1|871598|872054|GC|71|F3
## 4   MITE_T_10|Ecl_1605_22_1|871598|872054|GC|71|F3
## 5 MITE_T_16|Ecl_1605_22_2|7369169|7369617|TA|16|F4
## 6 MITE_T_16|Ecl_1605_22_2|7369169|7369617|TA|16|F4
##                             V2     V3  V4 V5 V6  V7  V8    V9   V10      V11
## 1 rnd-1_family-40#DNA/MuLE-MuDR 89.823 226 23  0   5 230     1   226 5.53e-81
## 2 rnd-1_family-40#DNA/MuLE-MuDR 89.500 200 18  2 255 452  2735  2933 2.02e-70
## 3 rnd-1_family-40#DNA/MuLE-MuDR 96.875  64  2  0 390 453    64     1 7.73e-25
```

```
## 4       rnd-1_family-89#Unknown 80.263 228 33  7  11 236     1   218 4.52e-42
## 5      rnd-4_family-161#Unknown 95.420 393 17  1  28 419     1   393 0.00e+00
## 6    rnd-4_family-109#LTR/Gypsy 79.926 269 43  9 111 373 17058 17321 2.04e-50
##   V12
## 1 294
## 2 259
## 3 108
## 4 165
## 5 627
## 6 193
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(stringr)
library(tidyr)
```

```
Epo_MT_RM$species <- "Epo"
Ecl_MT_RM$species <- "Ecl"
Ety_MT_RM$species <- "Ety"
```

then dyplr mutate() used to assign logical column for elements that share >= 90% identity with MITEs identified by MITEtracker. separate() will split V2 columns. V2 is currently 'familyname#TEclass' so we will split this column by the '#' to make it into two new columns called family and TE_class

```
typhina_MTRM <-rbind.data.frame(Epo_MT_RM, Ecl_MT_RM, Ety_MT_RM)
typhina_MTRM <- typhina_MTRM %>%  mutate(MITEs = V3 >= 90) %>%
  separate(V2, c("family", "TE_class"), "#")

head(typhina_MTRM)
```

```
##                                          V1           family          TE_class
## 1  MITE_T_1|Etp76_2|7275840|7276285|TA|37|F1  rnd-1_family-49           Unknown
## 2  MITE_T_1|Etp76_2|7275840|7276285|TA|37|F1  rnd-4_family-88               LTR
## 3  MITE_T_1|Etp76_2|7275840|7276285|TA|37|F1 rnd-4_family-298         LTR/Gypsy
## 4 MITE_T_19|Etp76_4|4525828|4526267|TA|32|F2  rnd-1_family-28 DNA/PIF-Harbinger
## 5 MITE_T_19|Etp76_4|4525828|4526267|TA|32|F2  rnd-1_family-28 DNA/PIF-Harbinger
## 6 MITE_T_19|Etp76_4|4525828|4526267|TA|32|F2  rnd-1_family-73           Unknown
##       V3  V4 V5 V6  V7  V8    V9   V10      V11   V12 species MITEs
## 1 96.078 408 16  0  20 427     1   408 0.00e+00 673.0     Epo  TRUE
## 2 91.685 445 33  4   1 445 32077 32517 0.00e+00 628.0     Epo  TRUE
## 3 81.726 394 66  6  20 409 14561 14952 4.49e-90 324.0     Epo FALSE
## 4 87.793 213 26  0 176 388   213     1 7.63e-68 250.0     Epo FALSE
## 5 97.059  34  1  0 355 388  3037  3070 4.96e-10  58.4     Epo  TRUE
## 6 97.059  34  1  0 355 388  2718  2751 4.96e-10  58.4     Epo  TRUE
```

Proportion of MITEs per annotated TE class

```
ggplot(typhina_MTRM, aes(x = TE_class, fill = MITEs)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 50)) +
  coord_flip() + scale_fill_brewer(palette = "Paired") +
  xlab("TE Class")
```