TE composition and dating

21 August 2020

```
library(repeatR)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
epo_rm <- read_rm("~/analyses/RepeatR/raw_data/Epo/Epo_TE_output.tsv")
ecl_rm <- read_rm("~/analyses/RepeatR/raw_data/Ecl/Ecl_TE.out")
ety_rm <- read_rm("~/analyses/RepeatR/raw_data/Ety/Ety_TE.out")

epo_rm$species <- "epo"
ecl_rm$species <- "ecl"
ety_rm$species <- "ety"
rm_df <- rbind.data.frame(epo_rm, ecl_rm, ety_rm)
head(rm_df)
```

```
##   score p_sub p_del p_ins   qname qstart  qend  qextend complement
## 1   370  12.2   1.4   9.0 Etp76_1   5229  5300 10974872          +
## 2    13  27.2   0.0   4.0 Etp76_1  15968 16019 10964153          +
## 3   459   6.5   1.6   1.6 Etp76_1  17891 17953 10962219          +
## 4    13  33.5   0.0   0.0 Etp76_1  32915 32966 10947206          +
## 5    25   7.5   2.2   2.2 Etp76_1  33620 33664 10946508          +
## 6   257   8.1   0.0   0.0 Etp76_1  39831 39867 10940305          +
##             tname         tclass tstart  tend textend ID ali_type species
## 1 rnd-1_family-77        Unknown      3    69       0  3  primary     epo
## 2         GA-rich Low_complexity      1    50       0  4  primary     epo
## 3 rnd-4_family-53        Unknown      1    63     198  5  primary     epo
## 4          (TTT)n  Simple_repeat      1    52       0  6  primary     epo
## 5      (TCGGCTA)n  Simple_repeat      1    45       0  7  primary     epo
## 6 rnd-1_family-51      LTR/Gypsy  10634 10670       0  8  primary     epo
```

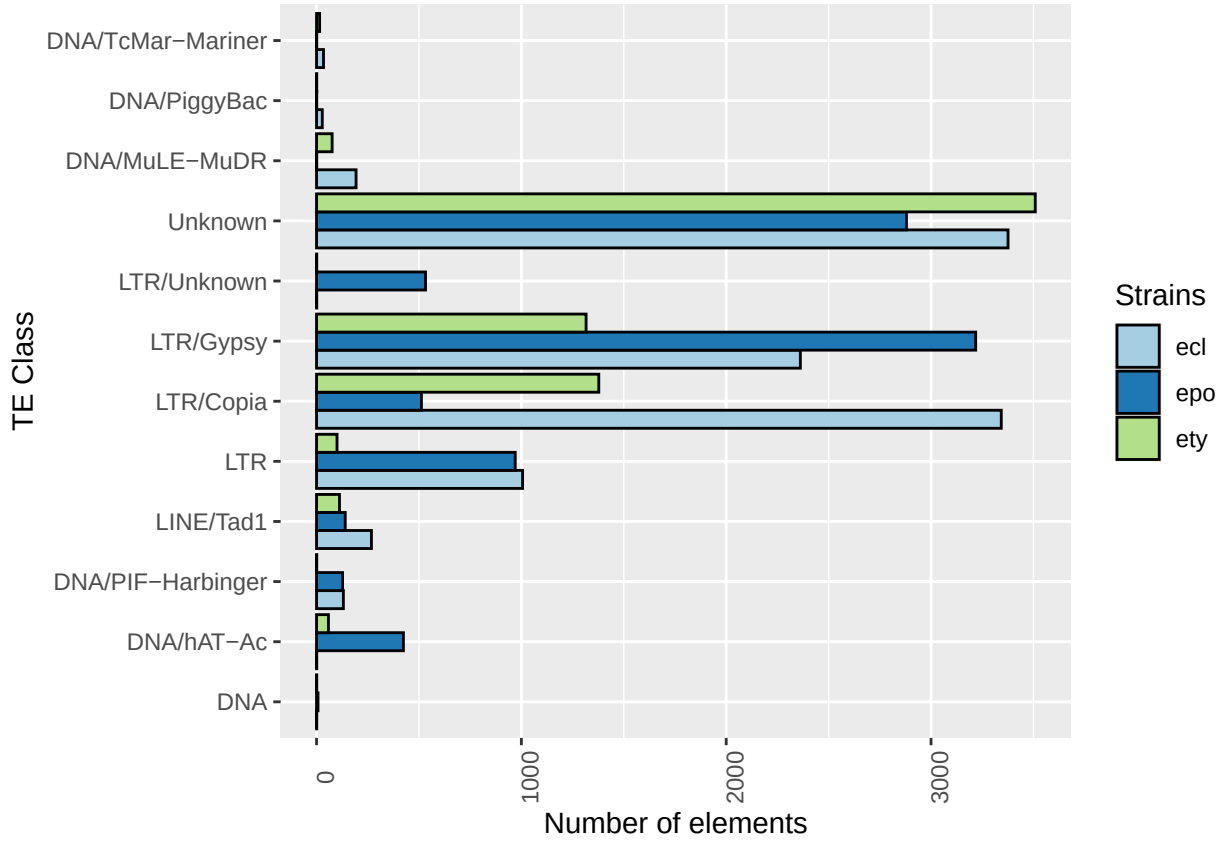
```
rm_df$ID <- paste(rm_df$species, rm_df$ID, sep = "_")
head(rm_df)
```

```
##   score p_sub p_del p_ins   qname qstart  qend  qextend complement
## 1   370  12.2   1.4   9.0 Etp76_1   5229  5300 10974872          +
## 2    13  27.2   0.0   4.0 Etp76_1  15968 16019 10964153          +
## 3   459   6.5   1.6   1.6 Etp76_1  17891 17953 10962219          +
## 4    13  33.5   0.0   0.0 Etp76_1  32915 32966 10947206          +
## 5    25   7.5   2.2   2.2 Etp76_1  33620 33664 10946508          +
```

```
## 6   257   8.1   0.0   0.0 Etp76_1  39831 39867 10940305          +
##             tname         tclass tstart  tend textend    ID ali_type species
## 1 rnd-1_family-77        Unknown      3    69       0 epo_3  primary     epo
## 2         GA-rich Low_complexity      1    50       0 epo_4  primary     epo
## 3 rnd-4_family-53        Unknown      1    63     198 epo_5  primary     epo
## 4          (TTT)n  Simple_repeat      1    52       0 epo_6  primary     epo
## 5      (TCGGCTA)n  Simple_repeat      1    45       0 epo_7  primary     epo
## 6 rnd-1_family-51      LTR/Gypsy  10634 10670       0 epo_8  primary     epo
```

Data Visualisation

```
c_rm_df <- complete(rm_df, species, tclass, fill = list(mean = 0))
ggplot(subset(c_rm_df, !(tclass %in% c("Simple_repeat", "Low_complexity", "rRNA",
  "Satellite"))), aes(x = tclass, fill = species)) +
  geom_bar(color = "black", position = position_dodge()) +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_brewer("Strains", palette = "Paired") +
  ylab("Number of elements") + xlab("TE Class") + coord_flip()
```




```
knitr::kable(table(c_rm_df$tclass, c_rm_df$species))
```

	ecl	epo	ety
DNA	1	7	1
DNA/hAT-Ac	1	424	58
DNA/PIF-Harbinger	131	127	1

	ecl	epo	ety
LINE/Tad1	268	140	112
Low_complexity	832	610	761
LTR	1006	970	100
LTR/Copia	3343	512	1378
LTR/Gypsy	2362	3218	1316
LTR/Unknown	1	532	1
rRNA	1	16	1
Satellite	1	66	1
Simple_repeat	14962	9188	12235
Unknown	3376	2880	3509
DNA/MuLE-MuDR	193	1	76
DNA/PiggyBac	28	1	1
DNA/TcMar-Mariner	34	1	15

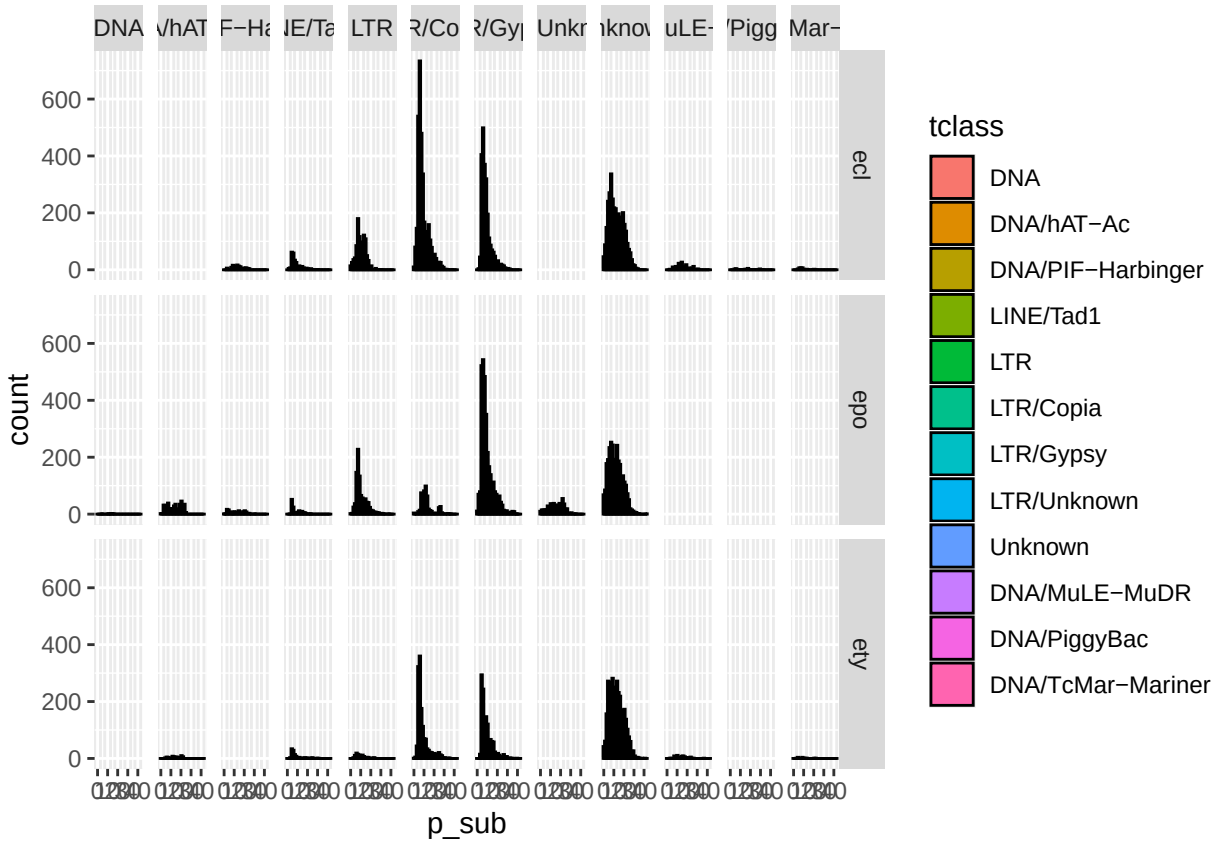
TE Dating

a variable of interest if `p_sub` which is the percentage difference between a given copy of a TE compared to the consensus sequence for that repeat. Recently active repeats will be centered close to zero, with older ones on much larger values.

```
ggplot(subset(c_rm_df, !(tclass %in% c("Simple_repeat", "Low_complexity", "rRNA",
  "Satellite"))), aes(x = p_sub, fill = tclass)) +
    geom_histogram(color = "black") +
    facet_grid(species ~ tclass)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

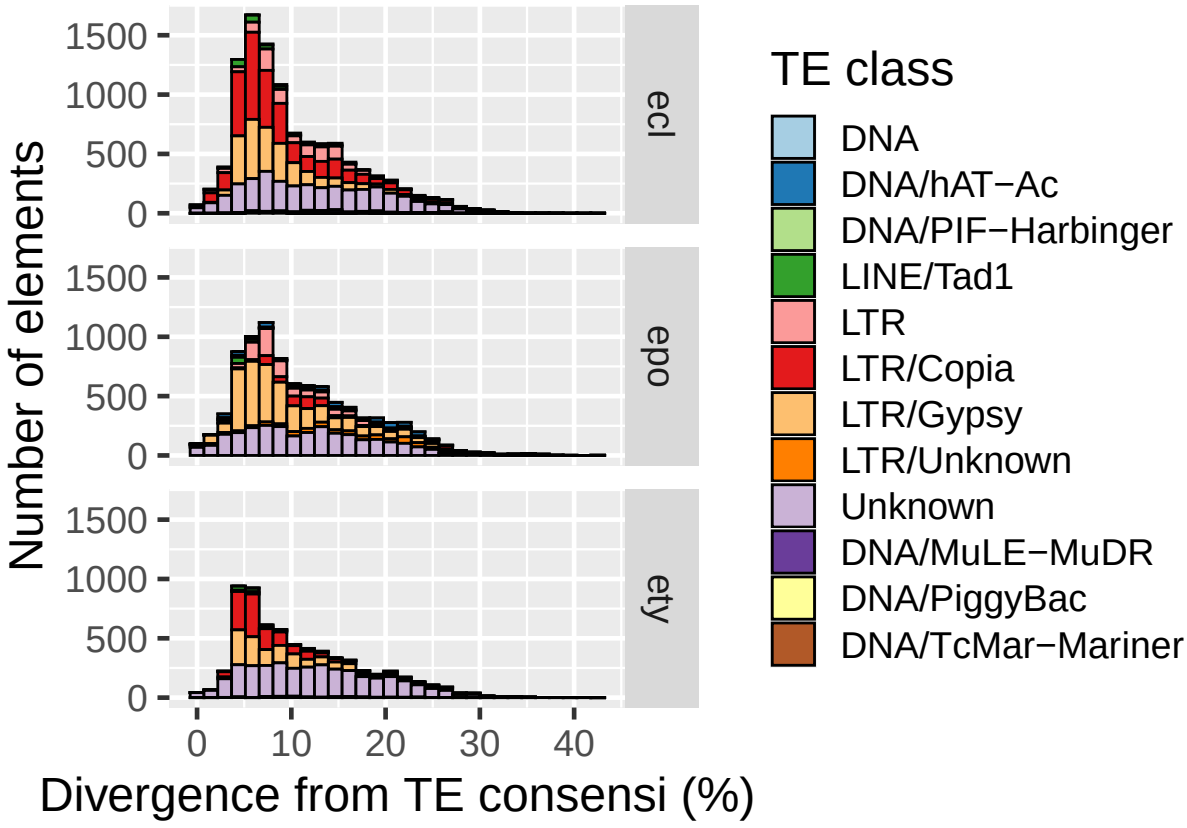
```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```




```
library(RColorBrewer)
ggplot(subset(c_rm_df, !(tclass %in% c("Simple_repeat", "Low_complexity", "rRNA","Satellite"))), aes(x =
  geom_histogram(color = "black") +
  facet_grid(species ~ .) +
  scale_fill_brewer("TE class", palette = "Paired") + ylab("Number of elements") +
  xlab("Divergence from TE consensi (%)") + theme_gray(base_size=18)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

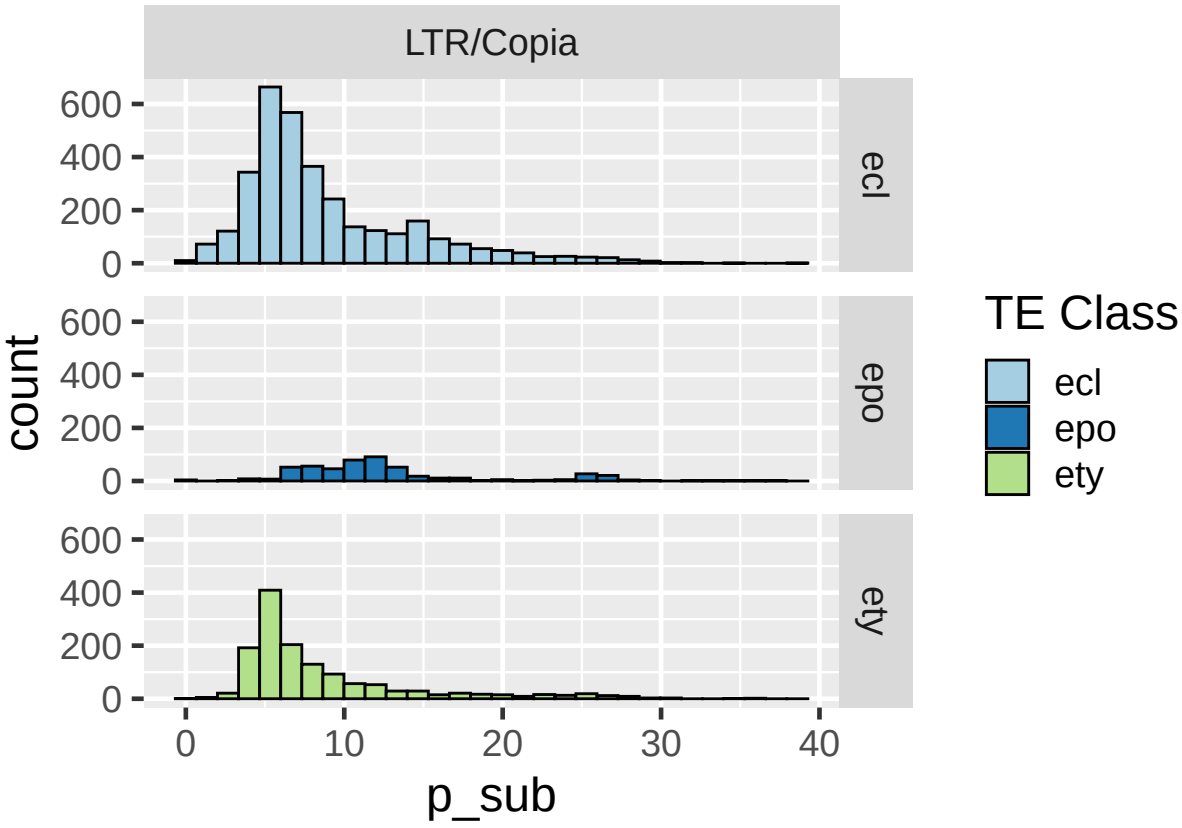
```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```



LTR copia by species

```
ggplot(subset(c_rm_df, tclass == "LTR/Copia"), aes(p_sub, fill = species)) +
  geom_histogram(color = "black") +
  facet_grid(species ~ tclass) + scale_fill_brewer("TE Class", palette = "Paired") + theme_gray(base_si
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


LTR/Copia
ecl
epo
ety
count
600
400
200
0
0
10
20
30
40
p_sub
TE Class
ecl
epo
ety