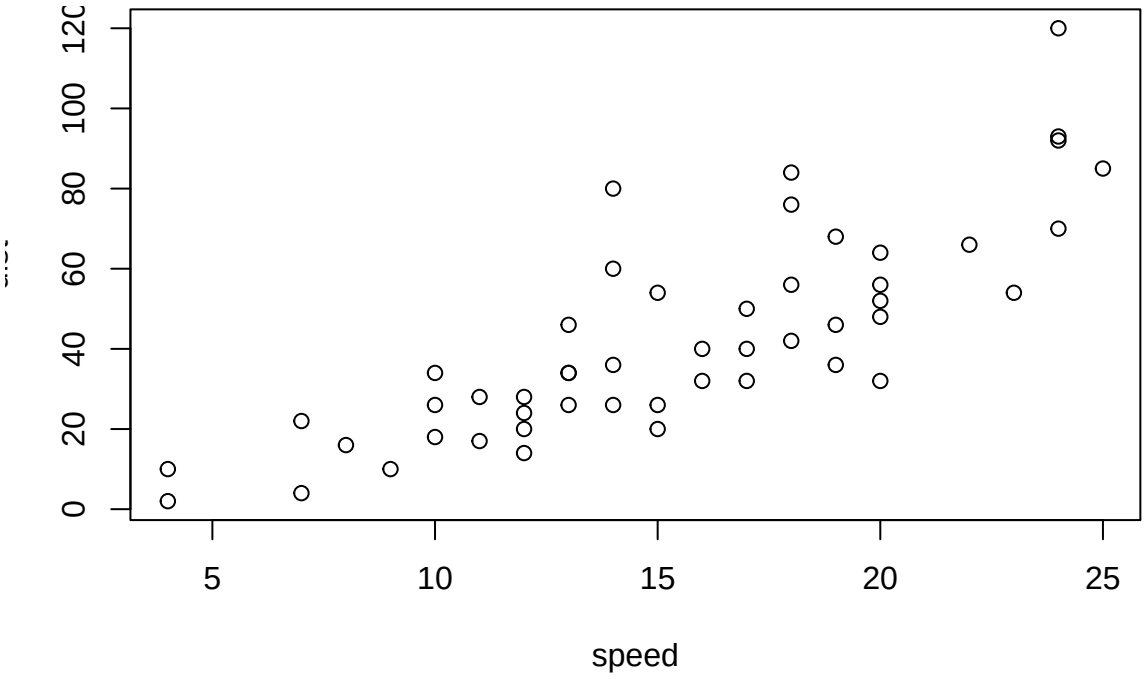
R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
plot(cars)
```



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed. Identity distribution: Figure 2 03/25/2024 filter by identity and length inspect biggest identity number and shortest length, in EDTA they excluded anything shorter than 80 maybe do that for my analysis too filter the whole table whose sum length is shorter than 80 I need class type and identity change names,combine DNA and MITE reorder class add copy number?

```
library(readr)
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(ggsci)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggplot2)
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
```

```
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
```

```
##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
mcic <- read_tsv("~/bigdata/TE_composition-EDTA/tables/McicTEtableV3.tsv")
```

```
## Rows: 2393218 Columns: 17
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (8): query, leftq, strand, family, class, beginr, leftr, overlap
## dbl (9): score, div., del., ins., beginq, endq, endr, ID, length
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
zrad <- read_tsv("~/bigdata/TE_composition-EDTA/tables/ZradTEtableV3.tsv")
```

```
## Rows: 748260 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (8): query, leftq, strand, family, class, beginr, leftr, overlap
## dbl (9): score, div., del., ins., beginq, endq, endr, ID, length
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
zradUn<-zrad %>% filter(class != "Unknown")
emus <- read_table("~/bigdata/EDTA/RepeatLandscape2-EDTA/Entomophthora_muscae_UCB.Nanopore10X_v2.rmblas
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_character(),
##   X6 = col_double(),
```

```
##   X7 = col_double(),
##   X8 = col_character(),
##   X9 = col_character(),
##   X10 = col_character(),
##   X11 = col_character(),
##   X12 = col_character(),
##   X13 = col_double(),
##   X14 = col_character(),
##   X15 = col_double(),
##   X16 = col_character()
## )
```

```
colnames(emus) <-c("score","div.","del.","ins.","query","beginq","endq","leftq","strand","family","class
emus <- emus %>% mutate(length=endq-beginq +1)
emus1 <- subset(emus,class != "Low_complexity")
emus2 <- subset(emus1,class != "Simple_repeat")
```

```
emai <- read_table("~/bigdata/EDTA/RepeatLandscape2-EDTA/Entomophaga_maimaiga_var_ARSEF_7190.rmblast1/E
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_character(),
##   X6 = col_double(),
##   X7 = col_double(),
##   X8 = col_character(),
##   X9 = col_character(),
##   X10 = col_character(),
##   X11 = col_character(),
##   X12 = col_character(),
##   X13 = col_double(),
##   X14 = col_character(),
##   X15 = col_double(),
##   X16 = col_character()
## )
```

```
colnames(emai) <-c("score","div.","del.","ins.","query","beginq","endq","leftq","strand","family","class
emai <- emai %>% mutate(length=endq-beginq +1)
emai1 <- subset(emai,class != "Low_complexity")
emai2 <- subset(emai1,class != "Simple_repeat")
mcic$overlap[is.na(mcic$overlap)] <- 1
mcicNO<-mcic %>% filter(overlap != "*")
zrad$overlap[is.na(zrad$overlap)] <- 1
zradNO<-zrad %>% filter(overlap != "*")
emus2$overlap[is.na(emus2$overlap)] <- 1
emusNO<-emus2 %>% filter(overlap != "*")
emai2$overlap[is.na(emai2$overlap)] <- 1
emaiNO<-emai2 %>% filter(overlap != "*")
```

```
mcicNO<-mcicNO %>% mutate(Identity=100-div.)
mcicNO$family<-sub("_INT","",mcicNO$family)
```

```
mcicNO$family<-sub("_LTR","",mcicNO$family)
mcicNO <- mcicNO %>% mutate(familyname = paste("Mcic",mcicNO$family,mcicNO$class))
zradNO<-zradNO %>% mutate(Identity=100-div.)
zradNO$family<-sub("_INT","",zradNO$family)
zradNO$family<-sub("_LTR","",zradNO$family)
zradNO <- zradNO %>% mutate(familyname = paste("Zrad",zradNO$family,zradNO$class))
emusNO<-emusNO %>% mutate(Identity=100-div.)
emusNO$family<-sub("_INT","",emusNO$family)
emusNO$family<-sub("_LTR","",emusNO$family)
emusNO <- emusNO %>% mutate(familyname = paste("Emus",emusNO$family,emusNO$class))
emaiNO<-emaiNO %>% mutate(Identity=100-div.)
emaiNO$family<-sub("_INT","",emaiNO$family)
emaiNO$family<-sub("_LTR","",emaiNO$family)
emaiNO <- emaiNO %>% mutate(familyname = paste("Emai",emaiNO$family,emaiNO$class))
```

```
emaiNO1<-emaiNO %>% group_by(ID,familyname,class) %>% summarise(Identity=mean(Identity),Len=sum(length)
```

```
## `summarise()` has grouped output by 'ID', 'familyname'. You can override using
## the `.groups` argument.
```

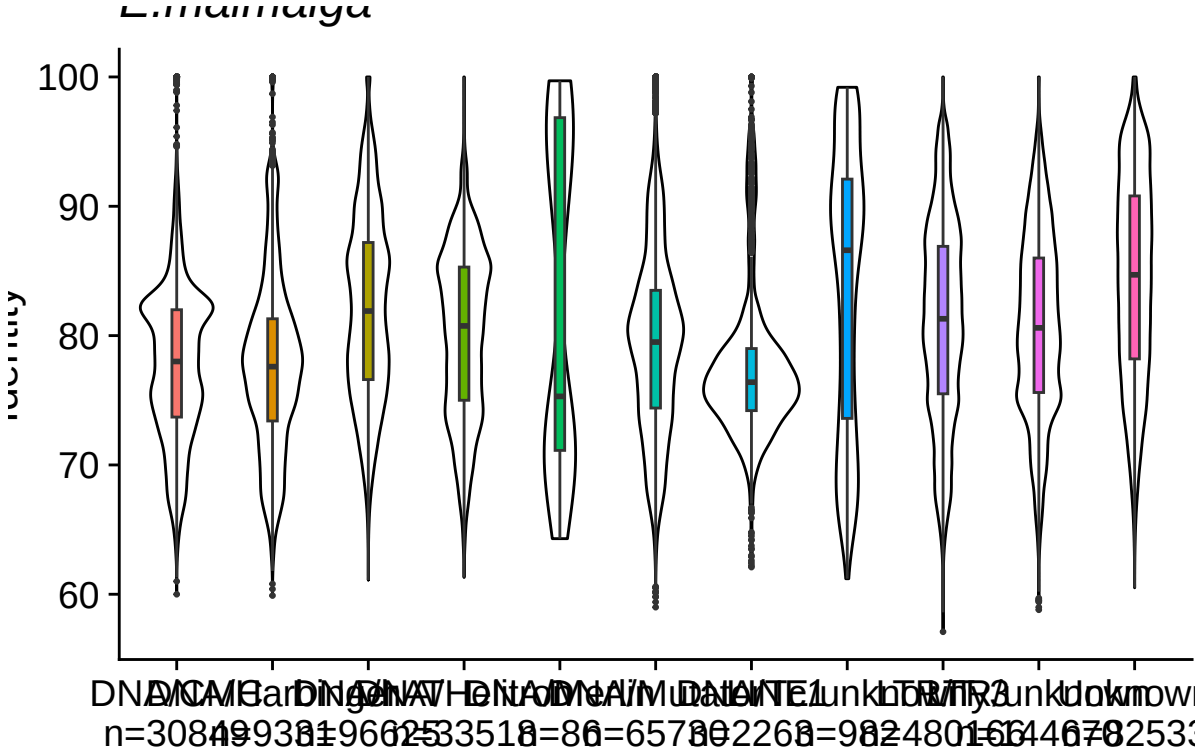
```
emaiNO2<-subset(emaiNO1, Len > 80)
```

```
emaiNO2$class <- gsub("TIR/Tc1_Mariner", "DNA/Tc1",emaiNO2$class )
emaiNO2$class <- gsub("TIR/Merlin", "DNA/Merlin",emaiNO2$class )
emaiNO2$class <- gsub("DNA/DTH", "DNA/Harbinger",emaiNO2$class )
emaiNO2$class <- gsub("DNA/DTT", "DNA/Tc1",emaiNO2$class )
emaiNO2$class <- gsub("MITE/DTM", "DNA/Mutator",emaiNO2$class )
emaiNO2$class <- gsub("DNA/DTM", "DNA/Mutator",emaiNO2$class )
emaiNO2$class <- gsub("DNA/DTA", "DNA/hAT",emaiNO2$class )
emaiNO2$class <- gsub("MITE/DTC", "DNA/CMC",emaiNO2$class )
emaiNO2$class <- gsub("DNA/DTC", "DNA/CMC",emaiNO2$class )
emaiNO2$class <- gsub("MITE/DTA", "DNA/hAT",emaiNO2$class )
emaiNO2$class <- gsub("MITE/DTH", "DNA/Harbinger",emaiNO2$class )
emaiNO2$class <- gsub("MITE/DTT", "DNA/Tc1",emaiNO2$class)
emaiNO2$class <- gsub("LTR/Gypsy", "LTR/Ty3",emaiNO2$class )
samplesizeemai<-emaiNO2 %>% group_by(class) %>% summarise(num=n())

library(cowplot)
fig2a<-emaiNO2 %>%
  left_join(samplesizeemai) %>%
  mutate(myaxis = paste0(class, "\n", "n=", num)) %>%
  ggplot( aes(x=myaxis, y=Identity, fill=class)) +
  geom_violin(width=1,color="black",fill="white")+
  geom_boxplot(width=0.1, outlier.size = 0.5)+
  theme_cowplot(16)+
  labs(title="E.maimaiga") +
  theme(legend.position="none",plot.title = element_text(face = "italic")) +
  xlab("")
```

```
## Joining with `by = join_by(class)`
```

```
fig2a
```




```
emusNO1<-emusNO %>% group_by(ID,familyname,class) %>% summarise(Identity=mean(Identity),Len=sum(length))
```

```
## `summarise()` has grouped output by 'ID', 'familyname'. You can override using
## the `.groups` argument.
```

```
emusNO2<-subset(emusNO1, Len > 80)
```

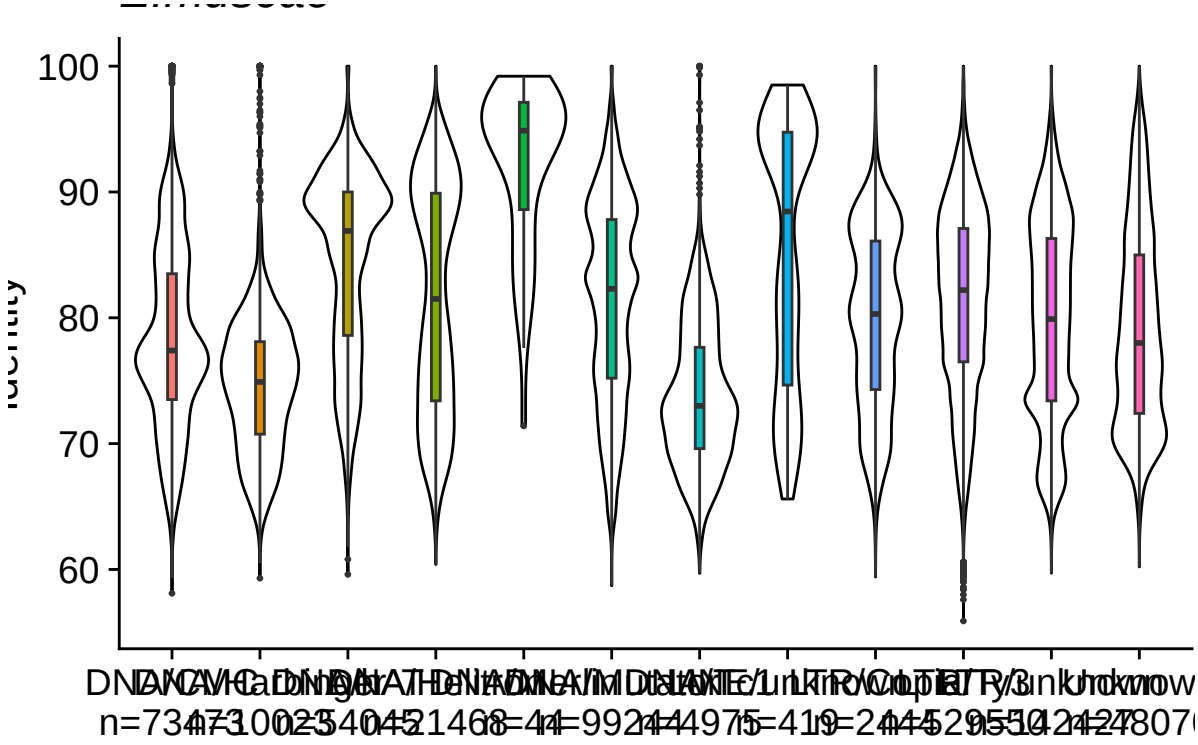
```
emusNO2$class <- gsub("TIR/Merlin", "DNA/Merlin",emusNO2$class )
emusNO2$class <- gsub("DNA/DTH", "DNA/Harbinger",emusNO2$class )
emusNO2$class <- gsub("DNA/DTT", "DNA/Tc1",emusNO2$class )
emusNO2$class <- gsub("MITE/DTM", "DNA/Mutator",emusNO2$class )
emusNO2$class <- gsub("DNA/DTM", "DNA/Mutator",emusNO2$class )
emusNO2$class <- gsub("DNA/DTA", "DNA/hAT",emusNO2$class )
emusNO2$class <- gsub("MITE/DTC", "DNA/CMC",emusNO2$class )
emusNO2$class <- gsub("DNA/DTC", "DNA/CMC",emusNO2$class )
emusNO2$class <- gsub("MITE/DTA", "DNA/hAT",emusNO2$class )
emusNO2$class <- gsub("MITE/DTH", "DNA/Harbinger",emusNO2$class )
emusNO2$class <- gsub("MITE/DTT", "DNA/Tc1",emusNO2$class)
emusNO2$class <- gsub("LTR/Gypsy", "LTR/Ty3",emusNO2$class )
samplesizeemus<-emusNO2 %>% group_by(class) %>% summarise(num=n())


fig2b<-emusNO2 %>%
  left_join(samplesizeemus) %>%
  mutate(myaxis = paste0(class, "\n", "n=", num)) %>%
  ggplot( aes(x=myaxis, y=Identity, fill=class)) +
  geom_violin(width=1,color="black",fill="white")+
  geom_boxplot(width=0.1, outlier.size = 0.5)+
  theme_cowplot(16)+
  labs(title="E.muscae") +
```

```
  theme(legend.position="none",plot.title = element_text(face = "italic")) +
  xlab("")
```

```
## Joining with `by = join_by(class)`
```

```
fig2b
```




```
mcicNO1<-mcicNO %>% group_by(ID,familyname,class) %>% summarise(Identity=mean(Identity),Len=sum(length))
```

```
## `summarise()` has grouped output by 'ID', 'familyname'. You can override using
## the `.groups` argument.
```

```
mcicNO2<-subset(mcicNO1, Len > 80)
```

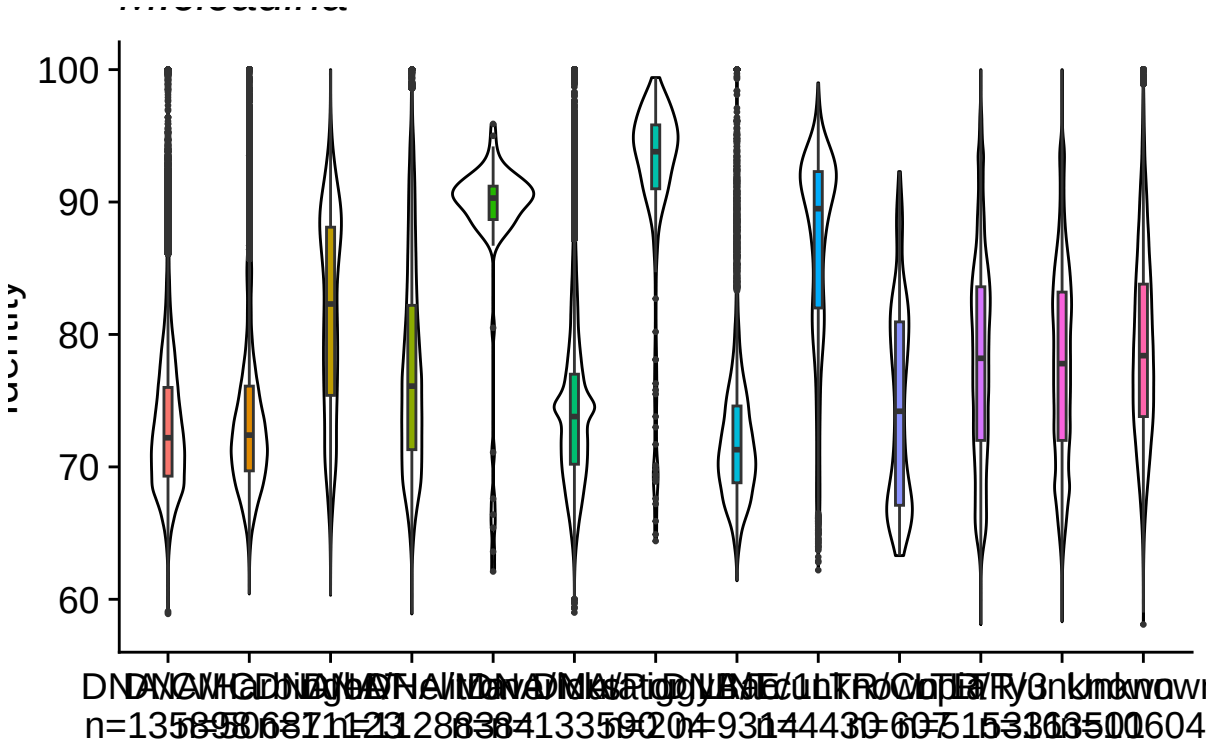
```
mcicNO2$class <- gsub("TIR/Tc1_Mariner", "DNA/Tc1",mcicNO2$class )
mcicNO2$class <- gsub("TIR/PiggyBac", "DNA/PiggyBac",mcicNO2$class )
mcicNO2$class <- gsub("DNA/DTH", "DNA/Harbinger",mcicNO2$class )
mcicNO2$class <- gsub("DNA/DTT", "DNA/Tc1",mcicNO2$class )
mcicNO2$class <- gsub("MITE/DTM", "DNA/Mutator",mcicNO2$class )
mcicNO2$class <- gsub("DNA/DTM", "DNA/Mutator",mcicNO2$class )
mcicNO2$class <- gsub("DNA/DTA", "DNA/hAT",mcicNO2$class )
mcicNO2$class <- gsub("MITE/DTC", "DNA/CMC",mcicNO2$class )
mcicNO2$class <- gsub("DNA/DTC", "DNA/CMC",mcicNO2$class )
mcicNO2$class <- gsub("MITE/DTA", "DNA/hAT",mcicNO2$class )
mcicNO2$class <- gsub("MITE/DTH", "DNA/Harbinger",mcicNO2$class )
mcicNO2$class <- gsub("MITE/DTT", "DNA/Tc1",mcicNO2$class)
mcicNO2$class <- gsub("polinton", "DNA/Mavericks",mcicNO2$class )
mcicNO2$class <- gsub("LTR/Gypsy", "LTR/Ty3",mcicNO2$class )
samplesizemcic<-mcicNO2 %>% group_by(class) %>% summarise(num=n())
#classname<-c("LTR/Ty3","LTR/unknown","LTR/Copia","LINE/unknown","DNA/Mutator","DNA/CMC","DNA/hAT","DNA
#classname<-rev(classname)
```

```
#mutate(class=fct_relevel(class,classname))
library(NatParksPalettes)
library(cowplot)
library(RColorBrewer)

fig2c<-mcicNO2 %>%
  left_join(samplesizemcic) %>%
  mutate(myaxis = paste0(class, "\n", "n=", num)) %>%
  ggplot( aes(x=myaxis, y=Identity, fill=class)) +
  geom_violin(width=1,color="black", fill="white")+
  geom_boxplot(width=0.1, outlier.size = 0.5)+
  theme_cowplot(16)+
  labs(title="M.cicadina") +
  theme(legend.position="none",plot.title = element_text(face = "italic")) +
  xlab("")
```

```
## Joining with `by = join_by(class)`
```

```
fig2c
```




```
zradNO1<-zradNO %>% group_by(ID,familyname,class) %>% summarise(Identity=mean(Identity),Len=sum(length))
```

```
## `summarise()` has grouped output by 'ID', 'familyname'. You can override using
## the `.groups` argument.
```

```
zradNO2<-subset(zradNO1, Len > 80)
```

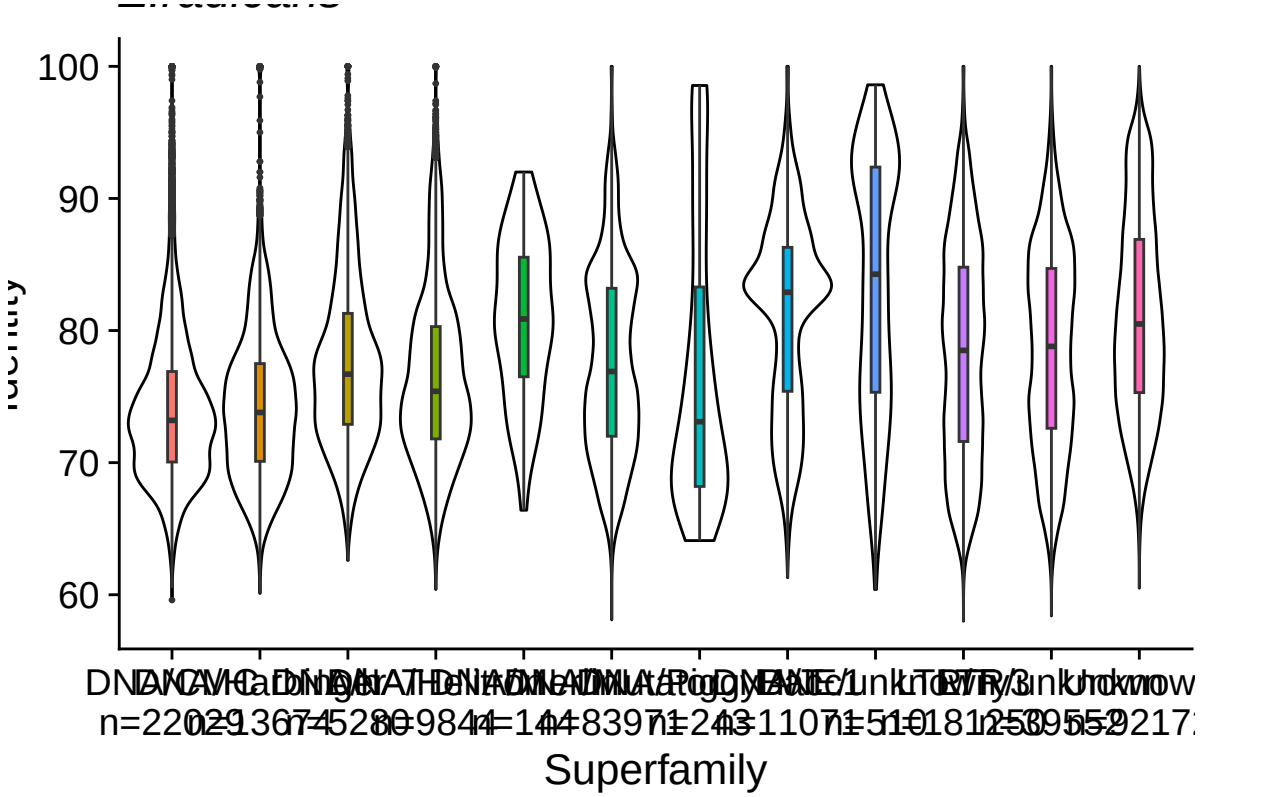
```
zradNO2$class <- gsub("TIR/Merlin", "DNA/Merlin",zradNO2$class )
zradNO2$class <- gsub("DNA/DTH", "DNA/Harbinger",zradNO2$class )
zradNO2$class <- gsub("DNA/DTT", "DNA/Tc1",zradNO2$class )
zradNO2$class <- gsub("MITE/DTM", "DNA/Mutator",zradNO2$class )
```

```
zradNO2$class <- gsub("DNA/DTM", "DNA/Mutator",zradNO2$class )
zradNO2$class <- gsub("DNA/DTA", "DNA/hAT",zradNO2$class )
zradNO2$class <- gsub("MITE/DTC", "DNA/CMC",zradNO2$class )
zradNO2$class <- gsub("DNA/DTC", "DNA/CMC",zradNO2$class )
zradNO2$class <- gsub("MITE/DTA", "DNA/hAT",zradNO2$class )
zradNO2$class <- gsub("MITE/DTH", "DNA/Harbinger",zradNO2$class )
zradNO2$class <- gsub("MITE/DTT", "DNA/Tc1",zradNO2$class)
zradNO2$class <- gsub("LTR/Gypsy", "LTR/Ty3",zradNO2$class )
zradNO2$class <- gsub("TIR/Tc1_Mariner", "DNA/Tc1",zradNO2$class )
zradNO2$class <- gsub("TIR/PiggyBac", "DNA/PiggyBac",zradNO2$class )
samplesizezrad<-zradNO2 %>% group_by(class) %>% summarise(num=n())


fig2d<-zradNO2 %>%
  left_join(samplesizezrad) %>%
  mutate(myaxis = paste0(class, "\n", "n=", num)) %>%
  ggplot( aes(x=myaxis, y=Identity, fill=class)) +
  geom_violin(width=1,color="black",fill="white")+
  geom_boxplot(width=0.1,outlier.size = 0.5)+
  theme_cowplot(16)+
  labs(title="Z.radicans") +
  theme(legend.position="none",plot.title = element_text(face = "italic")) +
  xlab("Superfamily")
```

```
## Joining with `by = join_by(class)`
```

```
fig2d
```



```
fig2<-plot_grid(fig2a, fig2b,fig2c,fig2d,ncol = 1)

ggsave(plot = fig2,
       path = "~/bigdata/TE_composition-EDTA/plots/",
       filename = "Fig2.pdf",
       width=18,
       height =25,
       units = "in",
       dpi = 500)
```