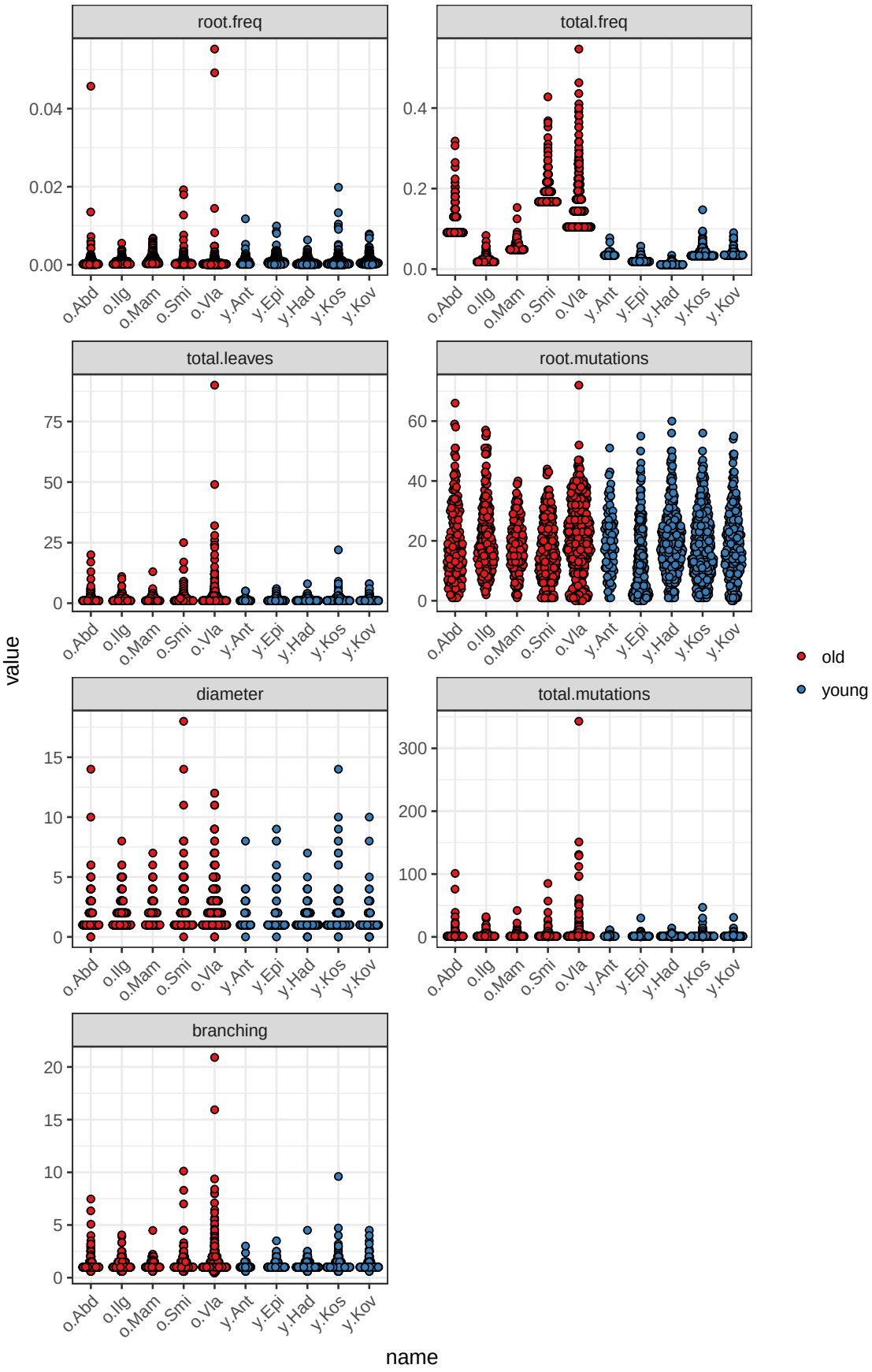
Phylogenetic analysis of clonal trees

Anna Obraztsova

2/9/2017

Diameter is the largest number of mutations between root and leave. Branching is a number of leaves divided by mean length of path from root to leave. Total.freq is a sum of frequencies of all vertices of tree.

```
library(ggplot2)
library(stringr)
library(reshape2)
library(dplyr)
library(ggbeeswarm)

old_rna = c("Abdulain", "Ilgen", "Mamaev", "Smirnov", "Vlasov")
young_rna = c("Antipyat", "Epifancev", "Hadjibekov", "Koshkin", "Kovalchuk")

gs <- data.frame()

for (sample in old_rna){
  .df <- read.table(paste('~/yf/trees/stat/yf_old_RNA/', sample, ".txt", sep = ""), header=T, sep="\t")
  names(.df) <- c('ndn', 'root.freq', 'total.freq', 'total.leaves', 'total.nodes', 'root.mutations', 'di
    .df$proj <- "old"
    .df$sample <- sample
    gs <- rbind(gs, .df)
}

for (sample in young_rna){
    .df <- read.table(paste('~/yf/trees/stat/yf_young_RNA/', sample, ".txt", sep = ""), header=T, sep="'
    names(.df) <- c('ndn', 'root.freq', 'total.freq', 'total.leaves', 'total.nodes', 'root.mutations',
    .df$proj <- "young"
    .df$sample <- sample
    gs <- rbind(gs, .df)
}

gs2 <- gs %>% melt(id=c('ndn', 'proj', 'sample')) %>%
  mutate(name=paste(str_sub(proj, 1, 1), str_sub(sample,1, 3), sep='.')) %>%
  filter(variable != 'total.nodes' & variable != 'mean.path')


ggplot(gs2) + geom_quasirandom(aes(x = name, y = value, group=proj, fill = proj), varwidth = TRUE, shape
facet_wrap(~variable, nrow=4, scales='free') +
scale_fill_brewer("", palette = "Set1") +
theme_bw() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


root.freq
total.freq
total.leaves
root.mutations
diameter
total.mutations
branching
value
name
old
young
o.Abd
o.Ilg
o.Mam
o.Smi
o.Vla
y.Ant
y.Epi
y.Had
y.Kos
y.Kov

```
gs3 <- gs2 %>% dplyr::group_by(sample, proj) %>% summarise(n = n())
ggplot(gs3) + geom_boxplot(aes(x = proj, y = n)) + ggtitle('Total number of clonal families')
```