

Task 2

Nathan LeRoy

4/28/2022

Task 2

```
# read in files
files <- list.files(
  path="./fasta",
  pattern="*.fa",
  full.names=TRUE,
  recursive=FALSE
)

# construct matrix and replace all NA
# with 0's
counts.frame = data.frame()
for (file in files) {
  # skip all.fa
  if(grepl("all.fa", file, fixed = TRUE)) {
    next
  }
  counts.frame <- bind_rows(counts.frame, countKmers(.readFasta(file), 4))
}
counts.frame[is.na(counts.frame)] <- 0

# run PCA
pc <- prcomp(
  counts.frame,
  center=TRUE,
  scale=TRUE
)

ggplot(as.data.frame(pc$x), aes(x=PC1, y=PC2)) +
  geom_point(color="blue") +
  theme_classic() +
  theme(
    axis.text=element_text(size=14),
    axis.title=element_text(size=14,face="bold"),
    panel.border = element_rect(
      color = "black", fill = NA, size = 2
    )
  )
)
```

