R implementation

Sleiman Bassim, PhD

April 8, 2016

Loaded functions:

```
#source("/media/Data/Dropbox/humanR/01funcs.R")
rm(list=ls())
#setwd("/media/Data/Dropbox/humanR/PD/")
#setwd("~/Dropbox/humanR/PD/")
###load("PD.Rdata", .GlobalEnv)
#lsos(pat="")
```

Load packages.

```
pkgs <- c('xlsx','lattice','ggplot2',
          'latticeExtra', 'dplyr', 'tidyr')
lapply(pkgs, require, character.only = TRUE)
```

1 Virus analysis

↰ Describe the sample preparation protocol

1.1 Classification of viruses and other phyla

Kraken is used for **read classification**. Deploying Kraken is documented in Github. Database used for classification are those of NCBI (date: 3/30/2016).

↰ Add github link after creating repo

Formatting the output of classified/unclassified reads with kraken

```
df <- read.table("./data/summary_all_outputs")
colnames(df)=gl(12,2,12,labels=c("A","B","F","P","V","I"))
df <- data.frame(samples = gl(26,2,26,labels=c("Ctrl",seq(1:12))),
                 kraken=gl(2,1,26,labels=c("C","U")),
                 df)
df1 <- gather(df[, c(1:2,seq(3,14,by=2))],"Phyla","Sequence",3:8)
df2 <- gather(df[, c(1:2,seq(4,14,by=2))],"Phyla","Percentage",3:8)
dff <- data.frame(df1,Percentage=df2[,4])
```

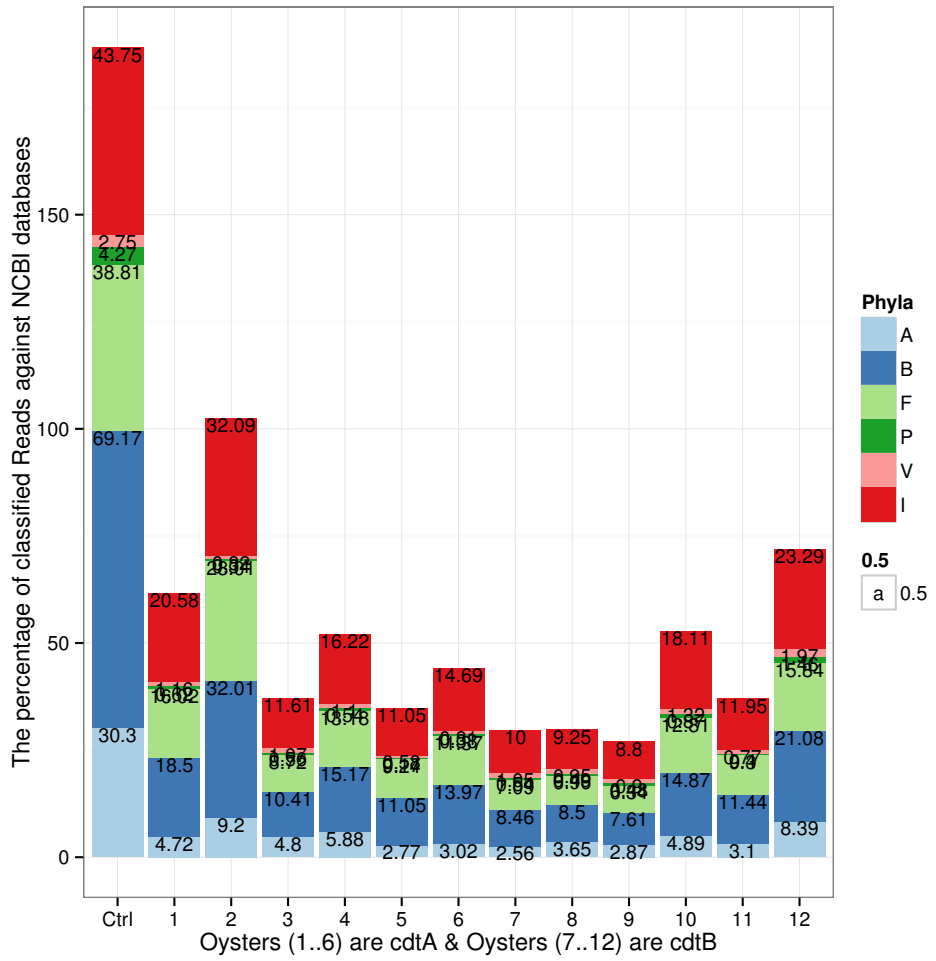
Plot the proportions of shotgun sequences being classified by sample. NCBI databases are gathered since 03/31/2016. The *control* shows aberrant classification. This can be do to bad sequencing reads due to aberrant nucleic acids in the control itself. No other sample exhibit this kind of over classification.

- Bacteria database (**B**)
- Archaea database (**A**)
- Plasmid database (**P**)
- Fungi database (**F**)
- Virus database (**V**)
- Invertebrate database (**I**)

```
summary(dff)
```

```
   samples   kraken Phyla     Sequence          Percentage
 Ctrl   :12   C:78   A:26   Min.   :    7651   Min.   : 0.2
 1      :12   U:78   B:26   1st Qu.:  924643   1st Qu.: 7.5
 2      :12          F:26   Median : 5773700   Median :50.0
 3      :12          P:26   Mean   :13241826   Mean   :50.0
 4      :12          V:26   3rd Qu.:25971749   3rd Qu.:92.5
 5      :12          I:26   Max.   :32280953   Max.   :99.8
 (Other):84
```

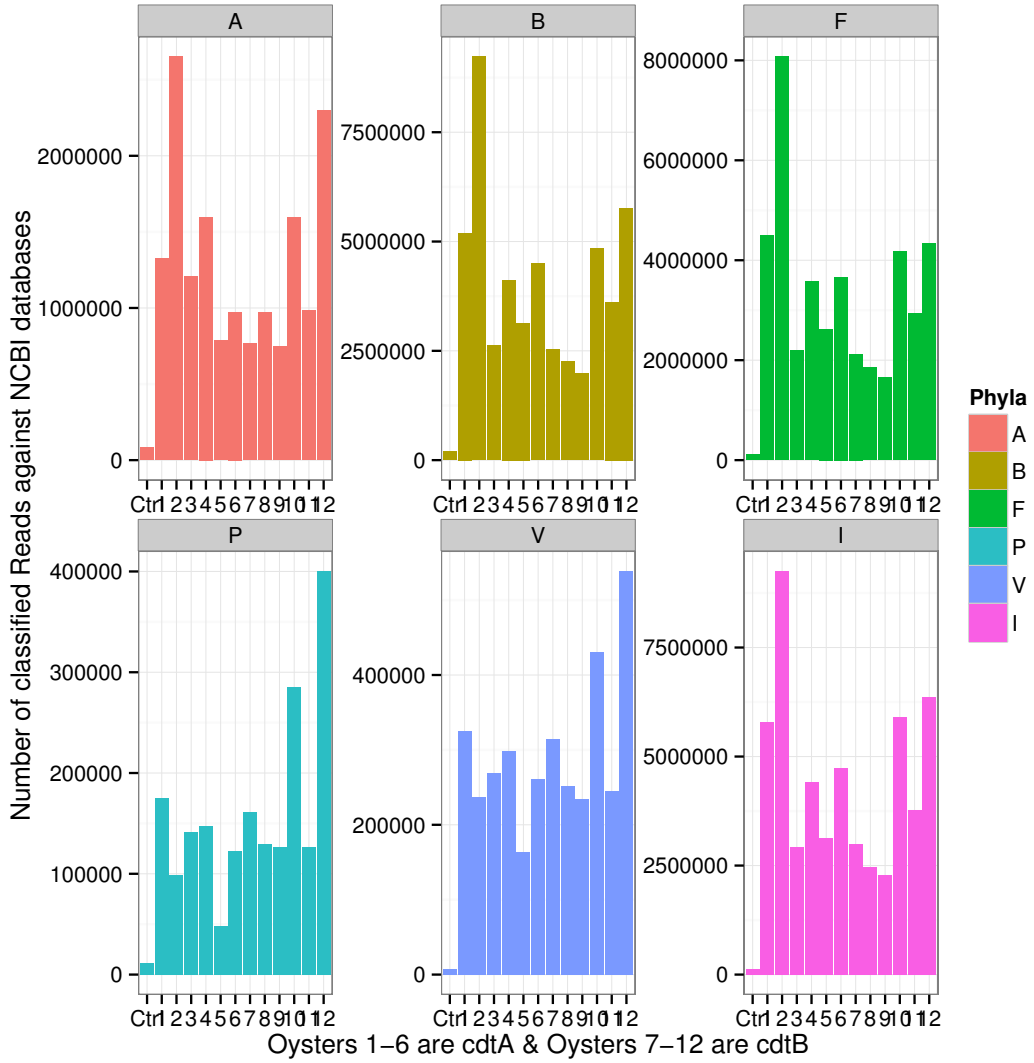
```
filter(dff, kraken == "C") %>%
    ggplot(aes(x = samples,
               y = Percentage,
               fill = Phyla)) +
    geom_bar(stat = "identity",
             position = "stack") +
    geom_text(aes(x = samples,
                  y = Percentage,
                  ymax = Percentage,
                  label = Percentage,
                  hjust = .5,
                  vjust = 1,
                  size = .5),
              position = "stack") +
    theme_bw() +
    scale_fill_brewer(type = "qual",
                      palette = 3) +
    labs(x = "Oysters (1..6) are cdtA & Oysters (7..12) are cdtB",
         y = "The percentage of classified Reads against NCBI databases")
```



17

What is the number of sequences.

```
filter(dff, kraken == "C") %>%
    ggplot(aes(x = samples,
               y = Sequence,
               fill = Phyla)) +
    geom_bar(stat = "identity") +
    theme_bw() +
    facet_wrap( ~ Phyla,
               ncol = 3,
               scales = "free") +
    labs(x = "Oysters 1-6 are cdtA & Oysters 7-12 are cdtB",
         y = "Number of classified Reads against NCBI databases")
```



2 System Information

The version number of R and packages loaded for generating the vignette were:

```
###save(list=ls(pattern=".*|.*"),file="PD.Rdata")
```

```
sessionInfo()

R version 3.2.1 (2015-06-18)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: elementary OS Luna

locale:
 [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8           LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8          LC_NAME=en_US.UTF-8
 [9] LC_ADDRESS=en_US.UTF-8        LC_TELEPHONE=en_US.UTF-8
[11] LC_MEASUREMENT=en_US.UTF-8    LC_IDENTIFICATION=en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
 [1] ggplot2_1.0.1       tidyr_0.2.0         dplyr_0.4.2
 [4] latticeExtra_0.6-26 RColorBrewer_1.1-2  lattice_0.20-31
 [7] xlsx_0.5.7          xlsxjars_0.6.1      rJava_0.9-6
[10] knitr_1.10.5        RevoUtilsMath_3.2.1

loaded via a namespace (and not attached):
 [1] Rcpp_0.11.6      magrittr_1.5     MASS_7.3-41
 [4] munsell_0.4.2    colorspace_1.2-6 R6_2.0.1
 [7] stringr_1.0.0    highr_0.5        plyr_1.8.3
[10] tools_3.2.1      parallel_3.2.1   grid_3.2.1
[13] gtable_0.1.2     DBI_0.3.1        digest_0.6.8
[16] lazyeval_0.1.10  assertthat_0.1   reshape2_1.4.1
[19] formatR_1.2      evaluate_0.7     labeling_0.3
[22] stringi_0.5-5    compiler_3.2.1   scales_0.2.5
[25] proto_0.3-10
```