Broliden_5325

04 Maj, 2021

Contents

#Load libraries and other scripts

```
##      checking for file ‘/private/var/folders/0c/vbtq1b6d1y10hd7zqb4wffpr1ldg_w/T/RtmpfwSeR0/remotes4
##   -  preparing ‘niceRplots’:
##      checking DESCRIPTION meta-information ...  v  checking DESCRIPTION meta-information
##   -  excluding invalid files
##      Subdirectory ’man’ contains invalid file names:
##      ‘nCoV_PBMC_1.h5’ ‘Normal_PBMC_13.h5’ ‘vignette.Rmd’
##   -  checking for LF line-endings in source and make files and shell scripts
##   -  checking for empty or unneeded directories
##   -  building ‘niceRplots_0.1.0.tar.gz’
##      Warning: invalid uid value replaced by that for user ’nobody’
##    Warning: invalid gid value replaced by that for user ’nobody’
##
##
```

#Defining some variables for the analysis

```
# create "other" taxonomy
other_taxa.fun <- function(df) {
  other_taxa <- df %>%
  unite("taxa_other", Kingdom:Species, sep = ";", remove = F, na.rm = T) %>%
  mutate(taxa_other = ifelse(is.na(.$Species), paste0(.$taxa_other, ";other"), .$taxa_other)) %>%
  #mutate(taxa_other = str_replace(.$taxa_other, ";NA.+|;NA", ";other")) %>%
  mutate(taxa_other = str_extract(.$taxa_other, "([^;]+);([^;]+)$")) %>% # get the two last taxa levels
```

```
  mutate(taxa_other = sub('^(.*);(other)', '\\2;\\1', .$taxa_other)) # place "other" first
return(other_taxa)
}

# create genus lacto taxonomy
genus_lacto.fun <- function(df){
  genus_lacto <- df %>%
    mutate(Genus_lacto = .$Genus, .after=Seq_ID) %>% #ifelse(grepl("Lactobacillus",.$Genus), paste(.$Ge
    mutate(Genus_lacto = ifelse(grepl("crispatus",.$Species), paste0(.$Genus_lacto, "crispatus/acidophil
                                ifelse(grepl("Lactobacillus",.$Genus), paste(.$Genus, .$Species),
                                       as.character(.$Genus_lacto))) )
  return(genus_lacto)
}

aggregate.fun <- function(df, level){
  level <- enquo(level)
  aggregated <- df %>%
    dplyr::rename(Taxonomy = !!level) %>%
    unite(., "Full_taxonomy", Kingdom:Species, sep = ";") %>%
    dplyr::select(-any_of(c("Sequence", "Seq_ID", "Genus_lacto", "taxa_other", "Full_taxonomy"))) %>%
    group_by(Taxonomy) %>%
    summarise(across(where(is.numeric), sum)) %>%
    ungroup()
return(aggregated)
}
```

Loading data and metadata

Taxonomy agglomeration

```
## $ASV_tissue_B2.csv
## [1] 409  95
##
## $ASV_tissue_B1.csv
## [1] 409   1
##
## $ASV_CVL_V2_B1.csv
## [1] 409  27
##
## $ASV_CVL_V2_B2.csv
## [1] 409  49
##
## $ASV_CVL_V2_C.csv
## [1] 409  62
##
## $ASV_CVL_V3.csv
## [1] 409 111
```

Loading data and metadata Paulos code

```
## $ASV_tissue_B2.csv
## [1] 506  95
##
## $ASV_tissue_B1.csv
## [1] 506   1
##
## $ASV_CVL_V2_B1.csv
## [1] 506  27
##
## $ASV_CVL_V2_B2.csv
## [1] 506  49
##
## $ASV_CVL_V2_C.csv
## [1] 389  62
##
## $ASV_CVL_V3.csv
## [1] 506 111
##
## $ASV_tissue_B2.csv
##  [1] "P089" "P094" "P100" "P051" "P118" "P098" "P085" "P114" "P108" "P121"
## [11] "P112" "P106" "P105" "P111" "P102" "P110" "P032" "P107" "P103" "P063"
## [21] "P120" "P119" "P116" "P117" "P109" "P115" "P113" "P021" "P036" "P045"
## [31] "P044" "P014" "P002" "P019" "P034" "P066" "P046" "P024" "P029" "P031"
## [41] "P050" "P041" "P003" "P008" "P059" "P018" "P052" "P039" "P042" "P057"
## [51] "P043" "P053" "P064" "P068" "P081" "P078" "P071" "P065" "P069" "P040"
## [61] "P061" "P027" "P086" "P087" "P080" "P073" "P083" "P079" "P074" "P082"
## [71] "P070" "P077" "P025" "P075" "P062" "P006" "P038" "P023" "P009" "P007"
## [81] "P047" "P016" "P011" "P026" "P048" "P028" "P060" "P010" "P049" "P037"
## [91] "P013" "P099" "P093" "P091" "P020"
##
## $ASV_tissue_B1.csv
## [1] "P001"
##
## $ASV_CVL_V2_B1.csv
##  [1] "P059" "P081" "P075" "P044" "P043" "P084" "P051" "P061" "P040" "P065"
## [11] "P078" "P052" "P068" "P094" "P089" "P100" "P098" "P054" "P008" "P045"
## [21] "P057" "P006" "P062" "P027" "P036" "P003" "P046"
##
## $ASV_CVL_V2_B2.csv
##  [1] "P007" "P009" "P010" "P011" "P012" "P013" "P014" "P015" "P018" "P023"
## [11] "P028" "P029" "P032" "P033" "P035" "P037" "P038" "P047" "P049" "P055"
## [21] "P060" "P063" "P064" "P067" "P082" "P085" "P087" "P088" "P091" "P093"
## [31] "P097" "P103" "P104" "P105" "P106" "P107" "P108" "P110" "P111" "P112"
## [41] "P113" "P114" "P115" "P116" "P117" "P118" "P119" "P120" "P121"
##
## $ASV_CVL_V2_C.csv
##  [1] "P053" "P061" "P073" "P054" "P062" "P058" "P036" "P003" "P046" "P059"
## [11] "P081" "P075" "P044" "P043" "P084" "P076" "P086" "P051" "P072" "P040"
## [21] "P065" "P080" "P078" "P052" "P056" "P079" "P068" "P099" "P094" "P089"
## [31] "P100" "P095" "P098" "P096" "P092" "P090" "P031" "P048" "P039" "P008"
## [41] "P045" "P022" "P057" "P004" "P002" "P020" "P025" "P021" "P006" "P026"
## [51] "P027" "P041" "P024" "P042" "P001" "P005" "P050" "P071" "P069" "P066"
```

```
## [61] "P074" "P034"
##
## $ASV_CVL_V3.csv
##   [1] "P001" "P002" "P003" "P004" "P005" "P006" "P007" "P008" "P009" "P010"
##  [11] "P011" "P012" "P013" "P014" "P017" "P018" "P019" "P020" "P021" "P022"
##  [21] "P023" "P024" "P025" "P026" "P027" "P028" "P029" "P030" "P031" "P032"
##  [31] "P033" "P034" "P035" "P036" "P037" "P038" "P039" "P040" "P041" "P042"
##  [41] "P043" "P044" "P045" "P047" "P048" "P049" "P050" "P051" "P052" "P053"
##  [51] "P057" "P059" "P060" "P061" "P062" "P063" "P064" "P066" "P067" "P068"
##  [61] "P069" "P070" "P071" "P072" "P073" "P074" "P075" "P076" "P077" "P078"
##  [71] "P079" "P080" "P081" "P082" "P083" "P085" "P086" "P087" "P088" "P089"
##  [81] "P090" "P091" "P092" "P093" "P094" "P095" "P096" "P097" "P098" "P099"
##  [91] "P100" "P101" "P102" "P109" "P103" "P105" "P106" "P107" "P108" "P110"
## [101] "P111" "P112" "P113" "P114" "P115" "P116" "P117" "P118" "P119" "P120"
## [111] "P121"
##
## $ASV_tissue_B2.csv
## [1] 767  95
##
## $ASV_tissue_B1.csv
## [1] 767   1
##
## $ASV_CVL_V2_B1.csv
## [1] 767  27
##
## $ASV_CVL_V2_B2.csv
## [1] 767  49
##
## $ASV_CVL_V2_C.csv
## [1] 767  62
##
## $ASV_CVL_V3.csv
## [1] 767 111
```

Merging, renaming and correcting batch effects on datasets

```
## [1] "ASV_tissue_B2.csv" "ASV_tissue_B1.csv" "ASV_CVL_V2_B1.csv"
## [4] "ASV_CVL_V2_B2.csv" "ASV_CVL_V2_C.csv"  "ASV_CVL_V3.csv"
## $ASV_CVL_V2_B1.csv
## [1] 409  27
##
## $ASV_CVL_V2_B2.csv
## [1] 409  49
##
## $ASV_CVL_V2_C.csv
## [1] 409  62
##
## $ASV_CVL_V3.csv
## [1] 409 111
##
## $ASV_tissue_V3
## [1] 409  96
```

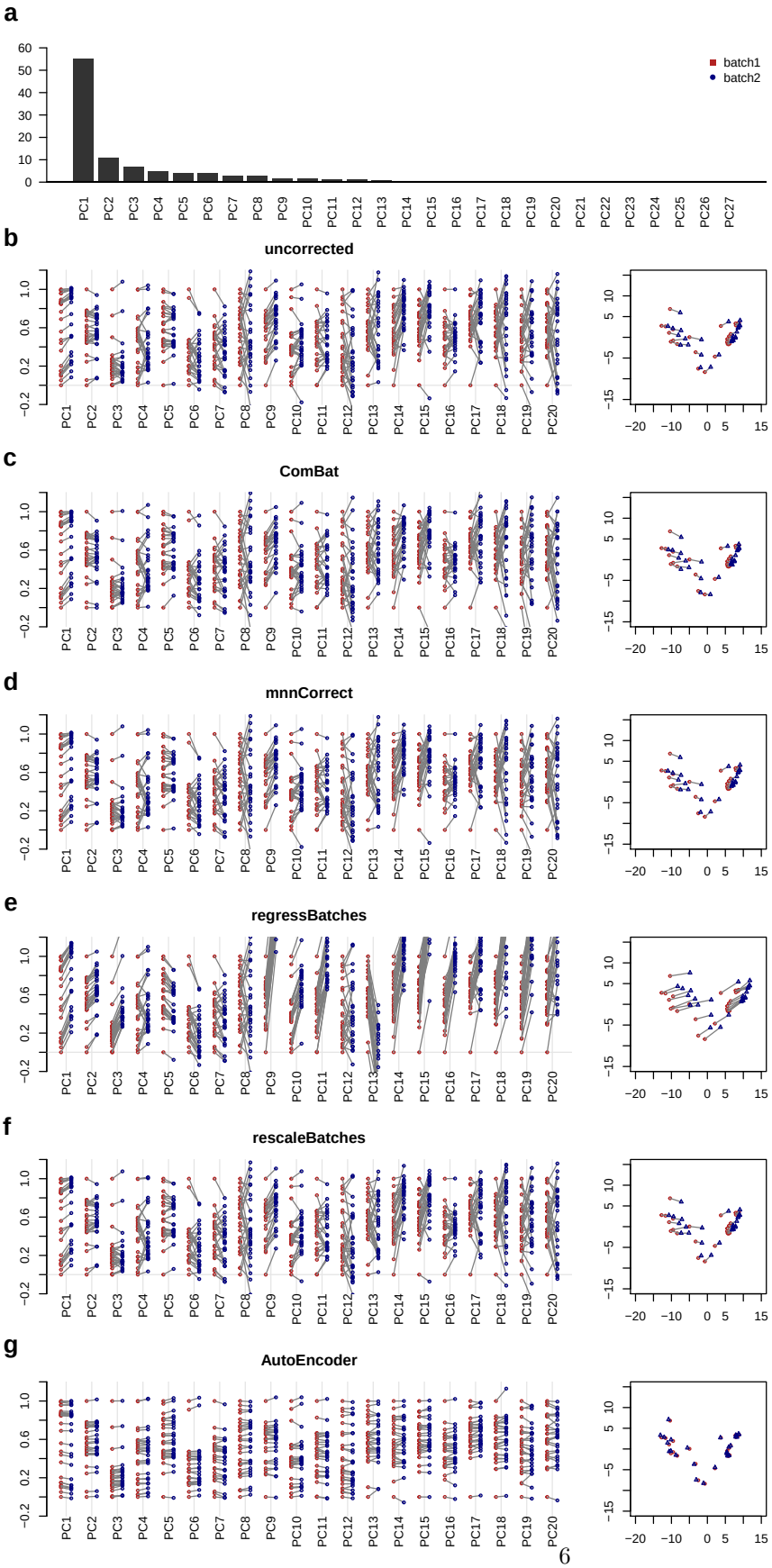
Batch correction with different approaches

```
## Standardizing Data across genes
```

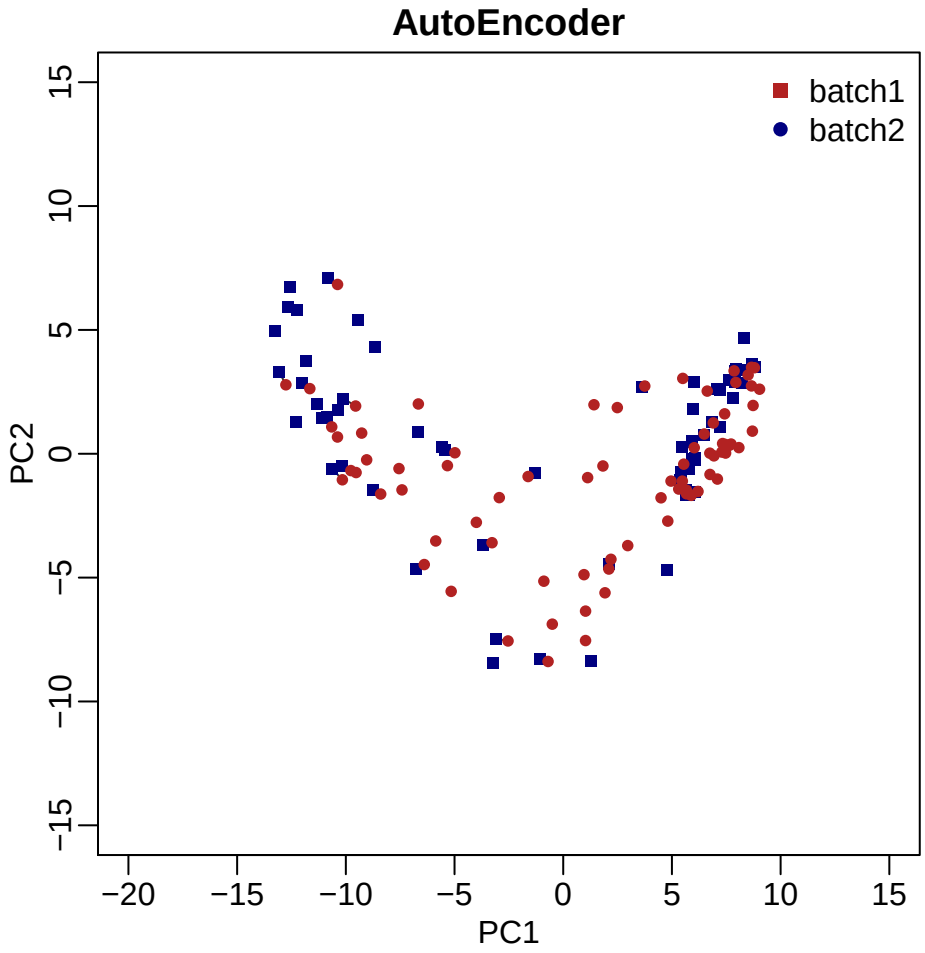
Batch correction with Autoencoder

```
## [1]  27 297
## [1] 297
## [1]  27 297
## [1]  27 297
```

Plotting the batch-corrected data for comparisson



As we can observe above, when using enough epochs (between 1000 and 2000), the autoencoder fits and learns the structure of the data.




```
## $ASV_CVL_V3.csv
## [1] 409 111
##
## $ASV_tissue_V3
## [1] 409  96
##
## $ASV_CVL_V2_raw
## [1] 409 111
##
## $ASV_CVL_V2
## [1] 297 111
##
## $ASV_CVL_V3.csv
##   [1] "P001" "P002" "P003" "P004" "P005" "P006" "P007" "P008" "P009" "P010"
##  [11] "P011" "P012" "P013" "P014" "P017" "P018" "P019" "P020" "P021" "P022"
##  [21] "P023" "P024" "P025" "P026" "P027" "P028" "P029" "P030" "P031" "P032"
##  [31] "P033" "P034" "P035" "P036" "P037" "P038" "P039" "P040" "P041" "P042"
##  [41] "P043" "P044" "P045" "P047" "P048" "P049" "P050" "P051" "P052" "P053"
##  [51] "P057" "P059" "P060" "P061" "P062" "P063" "P064" "P066" "P067" "P068"
##  [61] "P069" "P070" "P071" "P072" "P073" "P074" "P075" "P076" "P077" "P078"
```

```
##  [71] "P079" "P080" "P081" "P082" "P083" "P085" "P086" "P087" "P088" "P089"
##  [81] "P090" "P091" "P092" "P093" "P094" "P095" "P096" "P097" "P098" "P099"
##  [91] "P100" "P101" "P102" "P109" "P103" "P105" "P106" "P107" "P108" "P110"
## [101] "P111" "P112" "P113" "P114" "P115" "P116" "P117" "P118" "P119" "P120"
## [111] "P121"
##
## $ASV_tissue_V3
##  [1] "P089" "P094" "P100" "P051" "P118" "P098" "P085" "P114" "P108" "P121"
## [11] "P112" "P106" "P105" "P111" "P102" "P110" "P032" "P107" "P103" "P063"
## [21] "P120" "P119" "P116" "P117" "P109" "P115" "P113" "P021" "P036" "P045"
## [31] "P044" "P014" "P002" "P019" "P034" "P066" "P046" "P024" "P029" "P031"
## [41] "P050" "P041" "P003" "P008" "P059" "P018" "P052" "P039" "P042" "P057"
## [51] "P043" "P053" "P064" "P068" "P081" "P078" "P071" "P065" "P069" "P040"
## [61] "P061" "P027" "P086" "P087" "P080" "P073" "P083" "P079" "P074" "P082"
## [71] "P070" "P077" "P025" "P075" "P062" "P006" "P038" "P023" "P009" "P007"
## [81] "P047" "P016" "P011" "P026" "P048" "P028" "P060" "P010" "P049" "P037"
## [91] "P013" "P099" "P093" "P091" "P020" "P001"
##
## $ASV_CVL_V2_raw
##   [1] "P007" "P009" "P010" "P011" "P012" "P013" "P014" "P015" "P018" "P023"
##  [11] "P028" "P029" "P032" "P033" "P035" "P037" "P038" "P047" "P049" "P055"
##  [21] "P060" "P063" "P064" "P067" "P082" "P085" "P087" "P088" "P091" "P093"
##  [31] "P097" "P103" "P104" "P105" "P106" "P107" "P108" "P110" "P111" "P112"
##  [41] "P113" "P114" "P115" "P116" "P117" "P118" "P119" "P120" "P121" "P059"
##  [51] "P081" "P075" "P044" "P043" "P084" "P051" "P061" "P040" "P065" "P078"
##  [61] "P052" "P068" "P094" "P089" "P100" "P098" "P054" "P008" "P045" "P057"
##  [71] "P006" "P062" "P027" "P036" "P003" "P046" "P053" "P073" "P058" "P076"
##  [81] "P086" "P072" "P080" "P056" "P079" "P099" "P095" "P096" "P092" "P090"
##  [91] "P031" "P048" "P039" "P022" "P004" "P002" "P020" "P025" "P021" "P026"
## [101] "P041" "P024" "P042" "P001" "P005" "P050" "P071" "P069" "P066" "P074"
## [111] "P034"
##
## $ASV_CVL_V2
##   [1] "P059" "P081" "P075" "P044" "P043" "P084" "P051" "P061" "P040" "P065"
##  [11] "P078" "P052" "P068" "P094" "P089" "P100" "P098" "P054" "P008" "P045"
##  [21] "P057" "P006" "P062" "P027" "P036" "P003" "P046" "P007" "P009" "P010"
##  [31] "P011" "P012" "P013" "P014" "P015" "P018" "P023" "P028" "P029" "P032"
##  [41] "P033" "P035" "P037" "P038" "P047" "P049" "P055" "P060" "P063" "P064"
##  [51] "P067" "P082" "P085" "P087" "P088" "P091" "P093" "P097" "P103" "P104"
##  [61] "P105" "P106" "P107" "P108" "P110" "P111" "P112" "P113" "P114" "P115"
##  [71] "P116" "P117" "P118" "P119" "P120" "P121" "P001" "P002" "P004" "P005"
##  [81] "P020" "P021" "P022" "P024" "P025" "P026" "P031" "P034" "P039" "P041"
##  [91] "P042" "P048" "P050" "P053" "P056" "P058" "P066" "P069" "P071" "P072"
## [101] "P073" "P074" "P076" "P079" "P080" "P086" "P090" "P092" "P095" "P096"
## [111] "P099"
##
## $ASV_CVL_V3.csv
## [1] 409 111
##
## $ASV_tissue_V3
## [1] 409  96
##
## $ASV_CVL_V2_raw
## [1] 409 111
```

```
##
## $ASV_CVL_V2
## [1] 409 111
```