R implementation

Sleiman Bassim, PhD

July 17, 2015

Loaded functions:

↰ Project started July 10 2015

```
#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="")
```

1 Data preprocessing

Load packages.

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

```
[[1]]
[1] TRUE

[[2]]
[1] TRUE

[[3]]
[1] TRUE

[[4]]
[1] TRUE

[[5]]
[1] TRUE

[[6]]
[1] TRUE

[[7]]
[1] TRUE
```

Load file with read counts per sample.

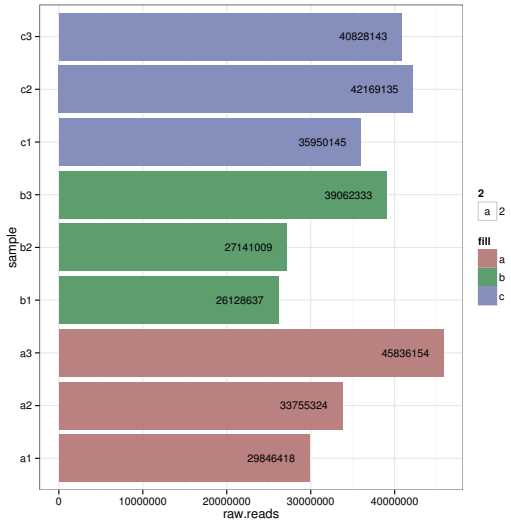
- A samples as nodule tissue
- B samples as non nodule diseased tissue
- C samples as non nodule non diseased tissue

Raw reads is the number of reads that have been trimmed, mapped to reference genome (Steve Roberts v15 with 21280 contigs), sorted by position on the genome, and cleaned from duplicated reads.

↰ Hypothetically these raw reads includes specific QPX reads

```
reads.counts <- read.xlsx("./data/mappedNodules.xlsx", sheetIndex = 1)
```

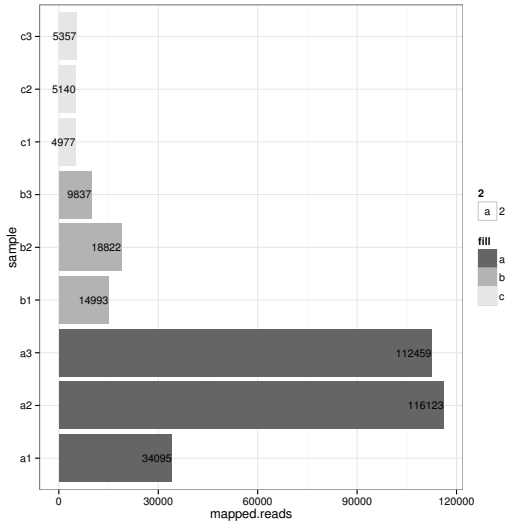
```
reads.counts$fill <- gl(3, 3, 9, labels = c("a", "b", "c"))
ggplot(reads.counts,
       aes(x = sample,
           y = raw.reads,
           fill = fill)) +
    coord_flip() +
        theme_bw() +
            geom_bar(stat = "identity") +
                geom_text(aes(x = sample,
                              y = raw.reads,
                              ymax = raw.reads,
                              label = raw.reads,
                              size = 2,
                              hjust = 1.3)) +
                    scale_fill_hue(c = 40, l = 60)
```



Number of reads that mapped to the reference genome of QPX.

↰ These reads are probably those of QPXs'

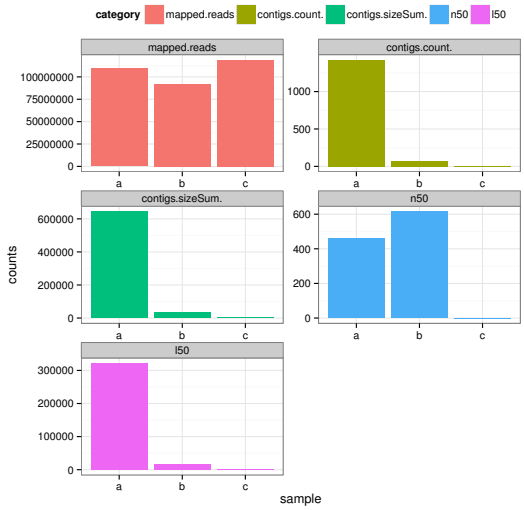
```
ggplot(reads.counts,
       aes(x = sample,
           y = mapped.reads,
           fill = fill)) +
    geom_bar(stat = "identity") +
        geom_text(aes(x = sample,
                      y = mapped.reads,
                      ymax = mapped.reads,
                      label = mapped.reads,
                      hjust = 1,
                      size = 2)) +
            coord_flip() +
                theme_bw() +
                    scale_fill_grey(start = .4, end = .9)
```



Mapped reads to the QPX reference where than assembled into contigs (ie, the reads showing in the chart above).

↰ These contigs must be specific transcripts to QPX

```
reads.counts <- read.xlsx("./data/mappedNodules.xlsx", sheetIndex = 2)
reads.counts <- gather(reads.counts, "category", "counts", 3:7)
ggplot(reads.counts,
       aes(x = sample,
           y = counts,
           fill = category)) +
    geom_bar(stat = "identity") +
       theme_bw() +
           facet_wrap( ~ category,
                       ncol = 2,
                       scales = "free") +
               theme(legend.position = "top")
```

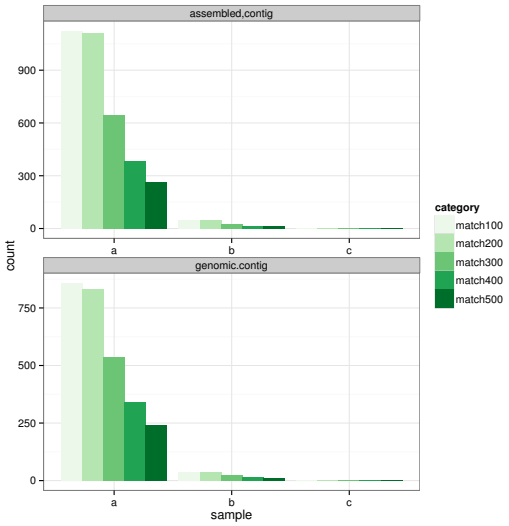


All contigs were than aligned to the reference genome QPX. The chart shows the number of contigs (both genomic and mRNA sequenced) that align with an increasing length of 100>200>300>400>500.

↰ Helps discard misassembled contigs or non QPX ones

```
reads.counts <- read.xlsx("./data/mappedNodules.xlsx", sheetIndex = 4)
```

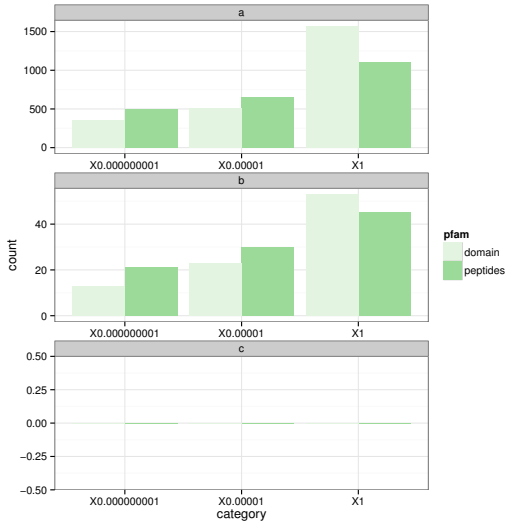
```
reads.counts <- gather(reads.counts, "category", "count", 3:7)
ggplot(reads.counts,
       aes(x = sample,
           y = count,
           fill = category)) +
    theme_bw() +
        geom_bar(stat = "identity",
                 position = "dodge") +
            facet_wrap( ~ blat, ncol = 1, scale = "free") +
                scale_fill_manual(values = brewer.pal(5, "Greens"))
```



All contigs were than translated to peptides in 6 frames. Each frame peptide was than aligned to the whole PFAM library (v28, date: Jul 14 2015).

```
reads.counts <- read.xlsx("./data/mappedNodules.xlsx", sheetIndex = 3)
reads.counts <- gather(reads.counts, "category", "count", 3:5)
ggplot(reads.counts,
       aes(x = category,
           y = count,
           fill = pfam)) +
    geom_bar(stat = "identity",
             position = "dodge") +
        theme_bw() +
            facet_wrap( ~ sample, ncol = 1,
                        scales = "free") +
                scale_fill_manual(values = brewer.pal(2, "Greens"))
```

```
Warning in brewer.pal(2, "Greens"): minimal value for n is 3, returning requested
palette with 3 different levels
```



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 3.1.2 (2014-10-31)
Platform: x86_64-unknown-linux-gnu (64-bit)

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] grid      stats     graphics  grDevices utils     datasets
[7] methods   base

other attached packages:
 [1] glmnet_1.9-8        Matrix_1.1-4        leaps_2.9
 [4] caret_6.0-37        ggbiplot_0.55       scales_0.2.4
 [7] plyr_1.8.1          tidyr_0.1           vegan_2.2-0
[10] permute_0.8-3       dplyr_0.3.0.2       ggplot2_1.0.0
[13] latticeExtra_0.6-26 RColorBrewer_1.0-5  lattice_0.20-29
[16] xlsx_0.5.7          xlsxjars_0.6.1      rJava_0.9-6
[19] knitr_1.8

loaded via a namespace (and not attached):
 [1] assertthat_0.1      BradleyTerry2_1.0-5 brglm_0.5-9
 [4] car_2.0-22          cluster_1.15.3      codetools_0.2-9
 [7] colorspace_1.2-4    compiler_3.1.2      DBI_0.3.1
[10] digest_0.6.4        evaluate_0.5.5      foreach_1.4.2
[13] formatR_1.0         gtable_0.1.2        gtools_3.4.1
[16] highr_0.4           iterators_1.0.7     labeling_0.3
[19] lazyeval_0.1.9      lme4_1.1-7          magrittr_1.5
[22] MASS_7.3-35         mgcv_1.8-4          minqa_1.2.4
[25] munsell_0.4.2       nlme_3.1-118        nloptr_1.0.4
[28] nnet_7.3-8          parallel_3.1.2      proto_0.3-10
[31] Rcpp_0.11.3         reshape2_1.4        splines_3.1.2
[34] stringr_0.6.2       tcltk_3.1.2         tools_3.1.2
```