HW3 statistics

Anna Rusnak

23 06 2022

```
Sys.setenv(LANG = "en")
```

```
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
#install.packages("RIdeogram")
library(RIdeogram)
#install.packages("devtools")
library(tidyr) # for separate function
```

```
#setwd("D:/Downloads")
#getwd()
```

```
dongola <- read.csv('DONGOLA_genes.tsv', sep = '\t' )
zanu <- read.csv('ZANU_genes.tsv', sep = '\t' )
mapping <- read.csv('gene_mapping.tsv', sep = '\t' )
```

Selecting required chromosomes in mapping data for gene mapping ZANU

```
mapping <- mapping[mapping$contig %in%  c('X', '2', '3'),]
unique(mapping$contig)
```

```
## [1] "2" "3" "X"
```

Separate DONG column

```
mapping <- separate(data=mapping, col=DONG, into=c("seq_id_dg", "middle_dg", 'strand_dg', 'length_dg',
```

Match seq_id of DONGOLA to chrososome names and filter

```
seq_id_to_chr = data.frame(id=c('2',"3","X"),
                           val=c('NC_053517.1', 'NC_053518.1', 'NC_053519.1'))
mapping$seq_id_dg <- with(seq_id_to_chr, id[match(mapping$seq_id_dg, val)])

mapping$name_dg <- gsub("DONG_", "", mapping$name_dg)



mapping <- mapping[mapping$seq_id_dg %in% c('2',"3","X"),]
unique(mapping$seq_id_dg)
```

```
## [1] "2" "X" "3"
```

Remove duplicated genes

```
mapping <- mapping[!duplicated(mapping$name),]
```

Karyotype table

```
karyotype_table <- setNames(data.frame(matrix(ncol=7, nrow=0)), c("Chr", "Start", "End", "fill", "speci

karyotype_table <- rbind(karyotype_table, data.frame(Chr=c('X','2','3'), Start=c(1, 1, 1), End=c(272380

karyotype_table <- rbind(karyotype_table, data.frame(Chr=c('X','2','3'), Start=c(1, 1, 1), End=c(269100
karyotype_table
```

```
##   Chr Start       End   fill species size  color
## 1   X     1  27238055 969696    ZANU   12 252525
## 2   2     1 114783175 969696    ZANU   12 252525
## 3   3     1  97973315 969696    ZANU   12 252525
## 4   X     1  26910000 969696 DONGOLA   12 252525
## 5   2     1 111990000 969696 DONGOLA   12 252525
## 6   3     1  95710000 969696 DONGOLA   12 252525
```

Synteny table

```
colnames(zanu) <- c('ID_1', 'Start_1', 'End_1', 'Strand_1')
colnames(dongola) <- c('ID_2', 'Start_2', 'End_2', 'Strand_2')
synteny_table <- merge(mapping, zanu, by.x='name', by.y='ID_1')
synteny_table <- merge(synteny_table, dongola, by.x='name_dg', by.y='ID_2')

names(synteny_table)[names(synteny_table) == 'contig'] <- 'Species_1'
names(synteny_table)[names(synteny_table) == 'seq_id_dg'] <- 'Species_2'
synteny_table$Species_1 <-mapvalues(synteny_table$Species_1,
                                    from=c('X', '2', '3'),
                                    to=c(1, 2, 3))
synteny_table$Species_2 <-mapvalues(synteny_table$Species_2,
                                    from=c('X', '2', '3'),
                                    to=c(1, 2, 3))

synteny_table$Species_1 <- as.integer(synteny_table$Species_1)
synteny_table$Species_2 <- as.integer(synteny_table$Species_2)
head(synteny_table)
```

```
##             name_dg      name Species_1 middle.position strand  ord ref.genes
## 1 gene-LOC120893177 gene_5019         2        48531603     -1 2862         1
## 2 gene-LOC120893178 gene_6182         2        86040949     -1 5204         1
## 3 gene-LOC120893179 gene_2643         2        86040395      1 5203         1
## 4 gene-LOC120893180 gene_5313         2        58398932     -1 3461         1
## 5 gene-LOC120893183 gene_2537         2        82790246      1 4995         1
## 6 gene-LOC120893185 gene_6082         2        82797727     -1 4998         1
##   Species_2 middle_dg strand_dg length_dg  Start_1    End_1 Strand_1  Start_2
## 1         2  65514822         1      3925 48528403 48534803       -1 65511152
## 2         2  28681053         1      1788 86040710 86041188       -1 28680597
## 3         2  28681607        -1      1789 86040192 86040598        1 28681316
## 4         2  55921684         1      3534 58381587 58416277       -1 55853085
## 5         2  31941591        -1      1998 82789431 82791062        1 31940683
## 6         2  31934112         1      1995 82796508 82798947       -1 31932898
##      End_2 Strand_2
## 1 65519724        1
## 2 28681368        1
## 3 28681908       -1
## 4 55941166        1
## 5 31942410       -1
## 6 31935462        1
```

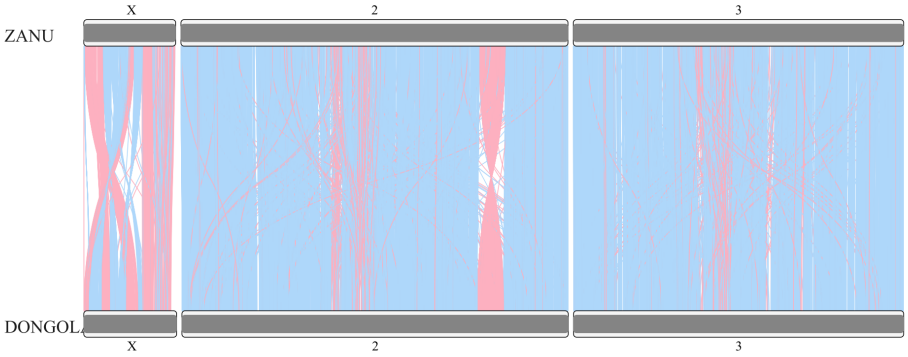
```
pink <- 'FFC0CB'
blue <- 'bbdffb'
dong_max_2 <- 111990000
dong_max_3 <- 95710000

color <- function(strand1, strand2, pink, blue){
  if (strand1 == strand2)
    return(pink)
  else
    return(blue)
```

```
}

synteny_table$fill <- mapply(color,
                             synteny_table$Strand_1,
                             synteny_table$Strand_2,
                             pink,
                             blue)
# inverse forc hr 2 ad chr3

two_to_three_color <- function(chr1, strand1, strand2, prev_fill, pink, blue){
  if (chr1 == 2 || chr1 == 3){
    if (strand1 == strand2)
      return(pink)
    else
      return(blue)
  }
  return(prev_fill)
}

synteny_table$fill <- mapply(two_to_three_color,
                             synteny_table$Species_1,
                             synteny_table$Strand_1,
                             synteny_table$Strand_2,
                             synteny_table$fill,
                             pink,
                             blue)

two_to_three <- function(chr1, pos2, dong_max_2, dong_max_3){
  if (chr1 == 2 || chr1 == 3){
    if (chr1 == 2)
      return(dong_max_2 - pos2 + 1)
    else
      return(dong_max_3 - pos2 + 1)
  }
  return(pos2)
}

synteny_table$Start_2 <- mapply(two_to_three,
                                synteny_table$Species_1,
                                synteny_table$Start_2,
                                dong_max_2,
                                dong_max_3)

synteny_table$End_2 <- mapply(two_to_three,
                              synteny_table$Species_1,
                              synteny_table$End_2,
                              dong_max_2,
                              dong_max_3)

synteny_table <- synteny_table[c('Species_1', 'Start_1', 'End_1', 'Species_2', 'Start_2', 'End_2', 'fil
synteny_table <- synteny_table[synteny_table$Species_1==synteny_table$Species_2, ]
head(synteny_table)
```

```
##   Species_1  Start_1    End_1 Species_2  Start_2    End_2   fill
## 1         2 48528403 48534803         2 46478849 46470277 bbdffb
## 2         2 86040710 86041188         2 83309404 83308633 bbdffb
## 3         2 86040192 86040598         2 83308685 83308093 bbdffb
## 4         2 58381587 58416277         2 56136916 56048835 bbdffb
## 5         2 82789431 82791062         2 80049318 80047591 bbdffb
## 6         2 82796508 82798947         2 80057103 80054539 bbdffb
```

Plot (converted svg to png online)

```
ideogram(karyotype=karyotype_table, synteny=synteny_table)
convertSVG("chromosome.svg", device="png")
```


X
2
3
ZANU
DONGOL
X
2
3