HW3

Roman Smirnov

2022-06-09

```
setwd('~/ITMO_EDUCATION/second_term/R/classes/HW3')
library(RIdeogram)
library(dplyr)
```

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

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

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

```
library(tidyr)
library(ggplot2)

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

1.Filter only interesting chromosomes (for ZANU)

```
gene_mapping <- filter(gene_mapping, contig==c('2', '3', 'X'))
gene_mapping %>% head()
```

```
##   contig middle.position strand ord      name ref.genes
## 1      2           31135     -1   0 gene_3542         1
## 2      2           46243     -1   3 gene_3544         1
## 3      2           97823      1   6   gene_82         1
## 4      2          115544      1   9   gene_83         1
## 5      2          133864     -1  12 gene_3549         1
## 6      2          187619     -1  15 gene_3551         1
##                                                    DONG
## 1  NC_053517.1,111908344,1,6540,DONG_gene-LOC120894913
## 2  NC_053517.1,111891588,1,6537,DONG_gene-LOC120904096
## 3 NC_053517.1,111846259,-1,6533,DONG_gene-LOC120908763
```

```
## 4 NC_053517.1,111822834,-1,6530,DONG_gene-LOC120893196
## 5  NC_053517.1,111797207,1,6527,DONG_gene-LOC120896719
## 6  NC_053517.1,111743563,1,6524,DONG_gene-LOC120908675
```

```
unique(gene_mapping$contig)
```

```
## [1] 2 3 X
## 84 Levels: 2 3 HiC_scaffold_10 HiC_scaffold_104 ... X
```

2.Reformating the gene_mapping dataframe (DONG column):

```
DONG_mapping <- data.frame(do.call('rbind', strsplit(as.character(gene_mapping$DONG),',',fixed=TRUE)))
colnames(DONG_mapping) <-  c('sequence_id', 'middle_coor_gene', 'strand', 'gene_length', 'dongo_name')
head(DONG_mapping, 3)
```

```
##   sequence_id middle_coor_gene strand gene_length             dongo_name
## 1 NC_053517.1        111908344      1        6540 DONG_gene-LOC120894913
## 2 NC_053517.1        111891588      1        6537 DONG_gene-LOC120904096
## 3 NC_053517.1        111846259     -1        6533 DONG_gene-LOC120908763
```

```
gene_mapping <- cbind(gene_mapping[0:6], DONG_mapping)
head(gene_mapping, 3)
```

```
##   contig middle.position strand ord      name ref.genes sequence_id
## 1      2           31135     -1   0 gene_3542         1 NC_053517.1
## 2      2           46243     -1   3 gene_3544         1 NC_053517.1
## 3      2           97823      1   6   gene_82         1 NC_053517.1
##   middle_coor_gene strand gene_length             dongo_name
## 1        111908344      1        6540 DONG_gene-LOC120894913
## 2        111891588      1        6537 DONG_gene-LOC120904096
## 3        111846259     -1        6533 DONG_gene-LOC120908763
```

3.Based on NCBI data let's make mapping of chromosomes by sequence_id

```
gene_mapping$sequence_id <- as.character(gene_mapping$sequence_id)
gene_mapping$sequence_id[gene_mapping$sequence_id == 'NC_053517.1'] <- '2'
gene_mapping$sequence_id[gene_mapping$sequence_id == 'NC_053518.1'] <- '3'
gene_mapping$sequence_id[gene_mapping$sequence_id == 'NC_053519.1'] <- 'X'
head(gene_mapping, 3)
```

```
##   contig middle.position strand ord      name ref.genes sequence_id
## 1      2           31135     -1   0 gene_3542         1           2
## 2      2           46243     -1   3 gene_3544         1           2
## 3      2           97823      1   6   gene_82         1           2
##   middle_coor_gene strand gene_length             dongo_name
## 1        111908344      1        6540 DONG_gene-LOC120894913
## 2        111891588      1        6537 DONG_gene-LOC120904096
## 3        111846259     -1        6533 DONG_gene-LOC120908763
```

4.Changing the gene name column:

```
gene_mapping$dongo_name <- gsub("DONG_", "", gene_mapping$dongo_name)
head(gene_mapping, 3)
```

```
##   contig middle.position strand ord      name ref.genes sequence_id
## 1      2           31135     -1   0 gene_3542         1           2
## 2      2           46243     -1   3 gene_3544         1           2
## 3      2           97823      1   6   gene_82         1           2
##   middle_coor_gene strand gene_length        dongo_name
## 1        111908344      1        6540 gene-LOC120894913
## 2        111891588      1        6537 gene-LOC120904096
## 3        111846259     -1        6533 gene-LOC120908763
```

5.Calculate the distance between genes:

```
gene_mapping$distance <- abs(gene_mapping$middle.position - as.integer(gene_mapping$middle_coor_gene))
```

6.Keep only genes which are shared between species

```
gene_mapping <- subset(gene_mapping, as.character(contig) == as.character(sequence_id))
```

7. Drop duplicated genes based on Distance

```
gene_mapping_drop <- gene_mapping[order(gene_mapping["distance", ])]
gene_mapping_drop <-  gene_mapping[!duplicated(gene_mapping$name), ]
gene_mapping_drop %>% head(3)
```

```
##   contig middle.position strand ord      name ref.genes sequence_id
## 1      2           31135     -1   0 gene_3542         1           2
## 2      2           46243     -1   3 gene_3544         1           2
## 3      2           97823      1   6   gene_82         1           2
##   middle_coor_gene strand.1 gene_length        dongo_name distance
## 1        111908344        1        6540 gene-LOC120894913    30711
## 2        111891588        1        6537 gene-LOC120904096    45820
## 3        111846259       -1        6533 gene-LOC120908763    97401
```

8. Build dataframes for ideogram

```
karyotype_df <- data.frame(matrix(ncol = 7, nrow = 0))
colnames(karyotype_df) <- c("Chr", "Start", "End", "fill", "species", "size", "color")
karyotype_df
```

8.1 karyotype df:

```
## [1] Chr     Start   End     fill    species size    color
## <0 rows> (or 0-length row.names)
```

```
karyotype_df <- rbind(karyotype_df, data.frame(Chr= c('X', '2', '3'),
                                               Start = c(1, 1, 1),
                                               End = c(26910000, 111990000, 95710000),
                                               fill='969696',
                                               species='Dongola',
                                               size=12,
                                               color='252525'))
karyotype_df %>% head(3)
```

8.2 add Dongo data

```
##   Chr Start       End   fill species size  color
## 1   X     1  26910000 969696 Dongola   12 252525
## 2   2     1 111990000 969696 Dongola   12 252525
## 3   3     1  95710000 969696 Dongola   12 252525
```

9. Synteny df

```
colnames(DONGOLA_genes) <- c('ID_1', 'Start_1', 'End_1', 'Strand_1')
colnames(ZANU_genes) <- c('ID_2', 'Start_2', 'End_2', 'Strand_2')

synteny_df <- gene_mapping_drop
synteny_df <- synteny_df %>% rename(Species_dongo_1 = sequence_id)
synteny_df <- synteny_df %>% rename(Species_zanu_2 = contig)
synteny_df <- merge(synteny_df, DONGOLA_genes, by.x='dongo_name', by.y='ID_1')
synteny_df <- merge(synteny_df, ZANU_genes, by.x='name', by.y='ID_2')

synteny_df$Species_dongo_1 <- as.character(synteny_df$Species_dongo_1)
synteny_df$Species_zanu_2 <- as.character(synteny_df$Species_zanu_2)
synteny_df$Species_dongo_1[synteny_df$Species_dongo_1 == 'X'] <- 1
synteny_df$Species_zanu_2[synteny_df$Species_zanu_2 == 'X'] <- 1

fill_blue <- '0000FF'
fill_red <- 'CC3300'

synteny_df %>% head(3)
```

9.1 rename columns

```
##         name        dongo_name Species_zanu_2 middle.position strand ord
## 1 gene_10000 gene-LOC120904181              3         8381575     -1 611
## 2 gene_10004 gene-LOC120901631              3         8394637     -1 617
## 3 gene_10007 gene-LOC120901889              3         8438078     -1 626
##   ref.genes Species_dongo_1 middle_coor_gene strand.1 gene_length distance
```

```
## 1         1               3         87299012        1        4110  8377392
## 2         1               3         87285889        1        4104  8390456
## 3         1               3         87244847        1        4095  8433902
##    Start_1    End_1 Strand_1 Start_2   End_2 Strand_2
## 1 87298621 87299449        1 8381256 8381894       -1
## 2 87285437 87286207        1 8394319 8394955       -1
## 3 87243963 87245566        1 8437423 8438733       -1
```

9.2 color the same genes by red. otherwise, by blue:

```
synteny_df$fill <- ifelse(synteny_df$Strand_1 == synteny_df$Strand_2, fill_red, fill_blue)
synteny_df$fill %>% head(3)
```

```
## [1] "0000FF" "0000FF" "0000FF"
```

```
synteny_df_filter <- synteny_df[c('Species_dongo_1', 'Start_1', 'End_1', 'Species_zanu_2', 'Start_2', '
synteny_df_filter %>% head(3)
```

```
##   Species_dongo_1  Start_1    End_1 Species_zanu_2 Start_2   End_2   fill
## 1               3 87298621 87299449              3 8381256 8381894 0000FF
## 2               3 87285437 87286207              3 8394319 8394955 0000FF
## 3               3 87243963 87245566              3 8437423 8438733 0000FF
```

```
synteny_df_filter$Species_dongo_1 <-  as.numeric(synteny_df_filter$Species_dongo_1)
synteny_df_filter$Species_zanu_2 <-  as.numeric(synteny_df_filter$Species_zanu_2)
karyotype_df$Chr <- as.character(karyotype_df$Chr)
```

10.Plot the ideogram

```
# ideogram(karyotype=karyotype_df, synteny=synteny_df_filter)
# convertSVG("chromosome.svg", device="png")
```

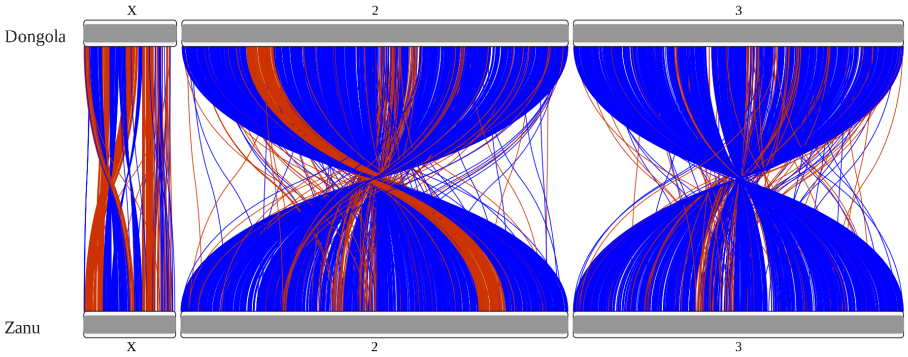

Figure 1: A caption