HW3

Kvan Valerii

2022-05-19

```
library("RIdeogram")
```

```
## Warning: package 'RIdeogram' was built under R version 4.1.3
```

```
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## 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
```

Read all data

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

head(gene_mapping)
```

```
##   contig middle.position strand ord      name ref.genes
## 1      2           31135     -1   0 gene_3542         1
## 2      2           38868     -1   1 gene_3543         1
## 3      2           42746      1   2   gene_80         1
## 4      2           46243     -1   3 gene_3544         1
## 5      2           53442     -1   4 gene_3545         1
## 6      2           60574      1   5   gene_81         1
##                                                    DONG
## 1  NC_053517.1,111908344,1,6540,DONG_gene-LOC120894913
## 2  NC_053517.1,111899667,1,6539,DONG_gene-LOC120904110
## 3 NC_053517.1,111895084,-1,6538,DONG_gene-LOC120904105
## 4  NC_053517.1,111891588,1,6537,DONG_gene-LOC120904096
## 5  NC_053517.1,111884408,1,6536,DONG_gene-LOC120895288
## 6 NC_053517.1,111877309,-1,6535,DONG_gene-LOC120895290
```

Split DONG column and drop it

```
gene_mapping <- cbind(gene_mapping, setNames(data.frame(x = do.call('rbind', strsplit(as.character(gene_
head(gene_mapping)
```

```
##   contig middle.position strand ord      name ref.genes    seq_id_d  middle_d
## 1      2           31135     -1   0 gene_3542         1 NC_053517.1 111908344
## 2      2           38868     -1   1 gene_3543         1 NC_053517.1 111899667
## 3      2           42746      1   2   gene_80         1 NC_053517.1 111895084
## 4      2           46243     -1   3 gene_3544         1 NC_053517.1 111891588
## 5      2           53442     -1   4 gene_3545         1 NC_053517.1 111884408
## 6      2           60574      1   5   gene_81         1 NC_053517.1 111877309
##   strand_d length_d                name_d
## 1        1     6540 DONG_gene-LOC120894913
## 2        1     6539 DONG_gene-LOC120904110
## 3       -1     6538 DONG_gene-LOC120904105
## 4        1     6537 DONG_gene-LOC120904096
## 5        1     6536 DONG_gene-LOC120895288
## 6       -1     6535 DONG_gene-LOC120895290
```

Filter mapping data

Choose only 2, 3, X chr for ZANU

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

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

Transfomr Dongola sequence id to chr

```
#NC_053517.1    2
#NC_053518.1    3
#NC_053519.1    X
#http://v2.insect-genome.com/Chromosome/Anopheles%20arabiensis

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

```
##   contig middle.position strand ord      name ref.genes seq_id_d  middle_d
## 1      2           31135     -1   0 gene_3542         1        2 111908344
## 2      2           38868     -1   1 gene_3543         1        2 111899667
## 3      2           42746      1   2   gene_80         1        2 111895084
## 4      2           46243     -1   3 gene_3544         1        2 111891588
## 5      2           53442     -1   4 gene_3545         1        2 111884408
## 6      2           60574      1   5   gene_81         1        2 111877309
##   strand_d length_d                name_d
## 1        1     6540 DONG_gene-LOC120894913
## 2        1     6539 DONG_gene-LOC120904110
## 3       -1     6538 DONG_gene-LOC120904105
## 4        1     6537 DONG_gene-LOC120904096
## 5        1     6536 DONG_gene-LOC120895288
```

```
## 6       -1     6535 DONG_gene-LOC120895290
```

Choose only 2, 3, X chr for DONGOLA

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

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

Transform name of DONGOLA genes in gene mapping table to format that used in DONGOLA csv.

```
head(gene_mapping)
```

```
##   contig middle.position strand ord      name ref.genes seq_id_d  middle_d
## 1      2           31135     -1   0 gene_3542         1        2 111908344
## 2      2           38868     -1   1 gene_3543         1        2 111899667
## 3      2           42746      1   2   gene_80         1        2 111895084
## 4      2           46243     -1   3 gene_3544         1        2 111891588
## 5      2           53442     -1   4 gene_3545         1        2 111884408
## 6      2           60574      1   5   gene_81         1        2 111877309
##   strand_d length_d                 name_d
## 1        1     6540 DONG_gene-LOC120894913
## 2        1     6539 DONG_gene-LOC120904110
## 3       -1     6538 DONG_gene-LOC120904105
## 4        1     6537 DONG_gene-LOC120904096
## 5        1     6536 DONG_gene-LOC120895288
## 6       -1     6535 DONG_gene-LOC120895290
```

```
head(dongola)
```

```
##                  ID start   end strand
## 1 gene-LOC120906950 59885 60345     -1
## 2 gene-LOC120906947 61728 64249      1
## 3 gene-LOC120906949 88010 88555     -1
## 4 gene-LOC120906948 90190 90789     -1
## 5 gene-LOC120906980   657  1316     -1
## 6 gene-LOC120906964 23986 24588      1
```

We need to remove "DONG" at the beginning of the name.

```
gene_mapping$name_d <- gsub("^DONG_(\\w+)", "\\1", gene_mapping$name_d)
```

Calculate distance between genes

```
gene_mapping$middle_d <- as.numeric(gene_mapping$middle_d)
gene_mapping$distance <- abs(gene_mapping$middle.position - gene_mapping$middle_d)
```

Mapping 1:1 ZANU to DONGOLA genes

Function to choose closest not reserved dongola gene for mapping

```
choose_closest_not_used_gene <- function(final_mapping) {
```

```
  #first we will map the genes with less distance.
  #For this we will sort all possible maps by distance in ascending order
  #p.s. That is not best options, because it can be more suitable variations
  #of closest genes
  gene_mapping <<- gene_mapping[order(gene_mapping$distance),]

  #here will be present the name of Dongola genes that were already mapped with
  #some ZANU gene.
  #It is need, because we have duplicated DONGOLA genes that shared between
  #multiple ZANU genes
  dongola_name_buffer <- c()

  for (zname in unique(gene_mapping$name)){
    #choose rows with this name
    tmp_rows = gene_mapping[gene_mapping$name == zname,]

    #sort by distance to iterate from min to max
    tmp_rows <- tmp_rows[order(tmp_rows$distance),]
    for (i in 1:nrow(tmp_rows)) {
      dname <- tmp_rows[i, ]$name_d
      contig <- tmp_rows[i, ]$contig
      seq_id <- tmp_rows[i, ]$seq_id_d

      if (!(dname %in% dongola_name_buffer)) {
        if (contig != seq_id)
          next

        #add to buffer
        dongola_name_buffer <- append(dongola_name_buffer, dname)

        #add to final mapping table
        final_mapping <- rbind(final_mapping, data.frame(chrZ=contig, chrD=seq_id,
                                                         zname=zname, dname=dname))
        break
      }
    }
  }

  return(final_mapping)
}
```

make table with the most closest genes

```
final_mapping <- setNames(data.frame(matrix(ncol = 4, nrow = 0)), c("chrZ", "chrD",
                                                                    "zname", "dname"))

final_mapping <- choose_closest_not_used_gene(final_mapping)
head(final_mapping)
```

```
##   chrZ chrD      zname             dname
## 1    X    X gene_13388 gene-LOC120905991
## 2    X    X gene_13057 gene-LOC120906736
## 3    X    X gene_13164 gene-LOC120905715
```

```
## 4    X    X gene_13015 gene-LOC120905674
## 5    X    X gene_13389 gene-LOC120905990
## 6    X    X gene_13761 gene-LOC120906317
```

Make tables for plots

```
create_karyotype_table <- function(final_mapping, specie1, specie2) {

  synteny_table_dual <- setNames(data.frame(matrix(ncol = 7, nrow = 0)),
                                 c("Species_1", "Start_1", "End_1", "Species_2",
                                   "Start_2", "End_2", "fill"))

  dongola_chr_2_max = 111990000
  dongola_chr_3_max = 95710000

  #final_mapping <- final_mapping[order(final_mapping$chr),]

  j = 1

  for(i in 1:nrow(final_mapping)) {
    tmp_row <- final_mapping[i, ]


    zname = tmp_row$zname[1]
    dname = tmp_row$dname[1]
    chrZ = tmp_row$chrZ[1]
    chrD = tmp_row$chrD[1]

    specie1_row <- specie1[specie1$ID == zname,]
    specie2_row <- specie2[specie2$ID == dname,]

    specie1_chr_num <- switch(chrZ, "X" = 1, "2" = 2, "3" = 3)
    specie2_chr_num <- switch(chrD, "X" = 1, "2" = 2, "3" = 3)


    #invert for 2 and 3 chr
    if (specie1_chr_num == 2 || specie1_chr_num == 3)
    {
      #5891bf - blue
      #db4527 - red
      color_to_fill <- if (specie1_row$strand[1] == specie2_row$strand[1]) 'db4527' else '5891bf'

      start_reverse <- if(specie1_chr_num == 2) dongola_chr_2_max - specie2_row$start + 1 else dongola_
      end_reverse <- if(specie1_chr_num == 2) dongola_chr_2_max - specie2_row$end + 1 else dongola_chr_3

      synteny_table_dual <- rbind(synteny_table_dual,
                                  data.frame(Species_1=specie1_chr_num, Start_1=specie1_row$start,
                                             End_1=specie1_row$end,
                Species_2=specie2_chr_num, Start_2=start_reverse, End_2=end_reverse,
                fill=color_to_fill))
    }
    else
    {
```

```
      #5891bf - blue
      #db4527 - red
      color_to_fill <- if (specie1_row$strand[1] == specie2_row$strand[1]) '5891bf' else 'db4527'

      synteny_table_dual <- rbind(synteny_table_dual,
                                  data.frame(Species_1=specie1_chr_num, Start_1=specie1_row$start,
                                             End_1=specie1_row$end,
                Species_2=specie2_chr_num, Start_2=specie2_row$start, End_2=specie2_row$end,
                fill=color_to_fill))
    }



    j <- j + 2
  }

  return (synteny_table_dual)
}
```

Final

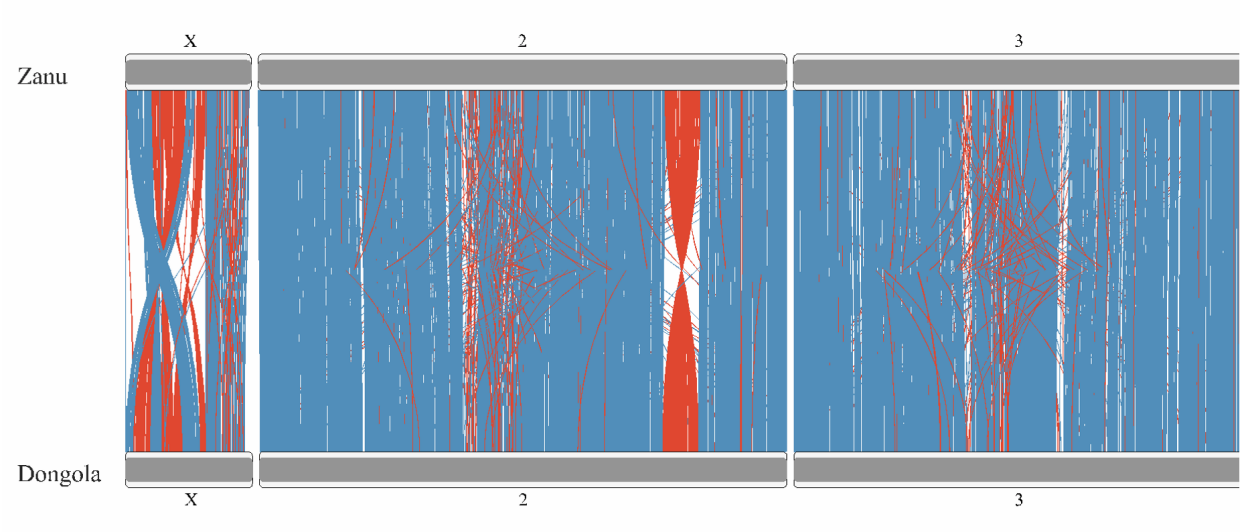
```
synteny_table_dual <- create_karyotype_table(final_mapping, zanu, dongola)

#karyotype table contains info about chromosomes
karyotype_table_dual <- setNames(data.frame(matrix(ncol = 7, nrow = 0)),
                                 c("Chr", "Start", "End", "fill",
                                   "species", "size", "color"))
#the length of ZENU chr was taken from HW3 description
karyotype_table_dual <- rbind(karyotype_table_dual,
                              data.frame(Chr=c('X','2','3'), Start=c(1, 1, 1),
                                         End=c(27238055, 114783175, 97973315),
                                    fill='969696',
                                         species='Zanu', size=12, color='252525'))

#dongola chromosomes length
#http://v2.insect-genome.com/Chromosome/Anopheles%20arabiensis
#need to convert mb to bp
karyotype_table_dual <- rbind(karyotype_table_dual,
                              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'))
```

Plot with Rideogram

```
ideogram(karyotype = karyotype_table_dual, synteny = synteny_table_dual)
convertSVG("chromosome.svg", device = "png")
```


X
2
3
Zanu
Dongola
X
2
3