# HW3

Violetta Konygina

07/06/2022

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

```
## Warning:  ’dplyr’    R  4.1.2
```

```
##
##  : ’dplyr’
```

```
##     ’package:stats’:
##
##     filter, lag
```

```
##     ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library("tidyr")
```

```
## Warning:  ’tidyr’    R  4.1.2
```

## 1. Read data

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

### 1.1. Gene mapping table

```
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
```

### 1.2. Zanu table

```
head(zanu)
```

```
##           ID  start    end strand
## 1 gene_13164   5022  23194     -1
## 2 gene_13165  40014  45938     -1
## 3 gene_13166  92876  97357     -1
## 4 gene_12497  99657 102434      1
## 5 gene_13167 106482 122413     -1
## 6 gene_13168 129453 131721     -1
```

### 1.3. Dongola table

```
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
```

## 2. Correction gene mapping table

### 2.1. Creating data frame from column DONG and then combining it with gene mapping

```
dong <- data.frame(x = do.call('rbind', strsplit(as.character(gene_mapping$DONG), ',', fixed=TRUE)))
colnames(dong) <- c('seq_id','middle_coord','strand_d','gene_length','gene_name')
```

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

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

Choose in contig column only 2, 3, X chromosomes

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

### 2.2. Perform mapping between chromosomes names and sequence IDs

From NCBI genome database: Chr 2 - NC_053517.1 Chr 3 - NC_053518.1 Chr X - NC_053519.1

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

```
##   contig middle.position strand ord      name ref.genes seq_id middle_coord
## 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 gene_length              gene_name
## 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 chromosomes in DONGOLA

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

### 2.3. Editing gene_name column

remove DONG_ in the gene_name

```
gene_mapping$gene_name <- as.character(lapply(gene_mapping$gene_name, gsub, pattern = '^DONG_', replace
head(gene_mapping)
```

```
##   contig middle.position strand ord      name ref.genes seq_id middle_coord
## 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 gene_length         gene_name
## 1        1        6540 gene-LOC120894913
## 2        1        6539 gene-LOC120904110
## 3       -1        6538 gene-LOC120904105
## 4        1        6537 gene-LOC120904096
## 5        1        6536 gene-LOC120895288
## 6       -1        6535 gene-LOC120895290
```

## 3. Distance calculation

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

Leave only same chromosomes between ZANU and DONGOLA

```
gene_mapping<-subset(gene_mapping, contig==seq_id)
```

## 4. Mapping between ZANU and DONGOLA genes

```
dong_map<-data.frame()
for (i in unique(gene_mapping$gene_name)){
  row_coll <- gene_mapping[gene_mapping$gene_name == i, ]
  min_count <- min(row_coll$distance)
  dong_map <- rbind(dong_map,row_coll[row_coll$distance == min_count, ])
}
dong_map <- dong_map[order(dong_map$distance),]
```

```
zanu_map<-data.frame()
for (i in unique(dong_map$name)){
  row_coll <- dong_map[dong_map$name == i, ]
  min_count <- min(row_coll$distance)
  zanu_map <- rbind(zanu_map,row_coll[row_coll$distance == min_count, ])
}
final_mapping <- zanu_map[order(zanu_map$distance),]
head(final_mapping)
```

```
##       contig middle.position strand  ord       name ref.genes seq_id
## 16445      X         7865798     -1  420 gene_13388         1      X
## 17420      X        22554898      1 1158 gene_13057         1      X
## 15952      X           14108     -1    0 gene_13164         1      X
## 17310      X        20658297      1 1063 gene_13015         1      X
## 16446      X         7870724     -1  421 gene_13389         1      X
## 17419      X        22549360     -1 1157 gene_13761         1      X
##       middle_coord strand_d gene_length         gene_name distance
## 16445      7858209        1         416 gene-LOC120905991     7589
## 17420     22562586       -1        1090 gene-LOC120906736     7688
## 15952        30435       -1           1 gene-LOC120905715    16327
## 17310     20675475       -1        1046 gene-LOC120905674    17178
## 16446      7853250        1         415 gene-LOC120905990    17474
## 17419     22569086        1        1091 gene-LOC120906317    19726
```

## 5. Synteny table

```
dongola_chr_2_end = 111988354
dongola_chr_3_end = 95710210
dongola_chr_X_end = 26913133
```

```
final_mapping$contig[final_mapping$contig == "X"] <- 1
final_mapping$seq_id[final_mapping$seq_id == "X"] <- 1

blue = "77dde7"
red = "ff5349"

start_zanu <- c()
end_zanu <- c()
fill <- c()
for (i in (1:nrow(final_mapping))){
    name <- final_mapping[i, "name"]
    fill <- if (final_mapping[i, "strand"] == final_mapping[i, "strand_d"]) append(fill, red)
    else append(fill, blue)
  start_zanu <- append(start_zanu, zanu[zanu$ID == name, "start"])
  end_zanu <- append(end_zanu, zanu[zanu$ID == name, "end"])
}
```

```
start_dong <- c()
end_dong <- c()
for (i in (1:nrow(final_mapping))){
    name <- final_mapping[i, "gene_name"]
    if (final_mapping[i, "contig"] == 1){
    start <- dongola_chr_X_end - dongola[dongola$ID == name, "start"]
    end <- dongola_chr_X_end - dongola[dongola$ID == name, "end"]
    } else if ((final_mapping[i, "contig"] == 2)){
      start <- dongola_chr_2_end - dongola[dongola$ID == name, "start"]
      end <- dongola_chr_2_end - dongola[dongola$ID == name, "end"]
    } else {
      start <- dongola_chr_3_end - dongola[dongola$ID == name, "start"]
      end <- dongola_chr_3_end - dongola[dongola$ID == name, "end"]
    }
  start_dong <- append(start_dong, start)
  end_dong <- append(end_dong, end)
}
```

```r
synteny_table <- data.frame(Species_1 = as.numeric(final_mapping$contig),
                            Start_1 = start_zanu,
                            End_1 = end_zanu,
                            Species_2 = as.numeric(final_mapping$seq_id),
                            Start_2 = start_dong, End_2 = end_dong, fill = fill)
head(synteny_table)
```

```
##   Species_1  Start_1    End_1 Species_2  Start_2    End_2   fill
## 1         1  7865247  7866349         1 19055658 19054278 77dde7
## 2         1 22553805 22555991         1  4351086  4349049 77dde7
## 3         1     5022    23194         1 26894161 26861576 ff5349
## 4         1 20657888 20658706         1  6238316  6237208 77dde7
## 5         1  7870052  7871396         1 19060967 19058761 77dde7
## 6         1 22548905 22549815         1  4344615  4343484 77dde7
```

### 6. Karyotype table

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

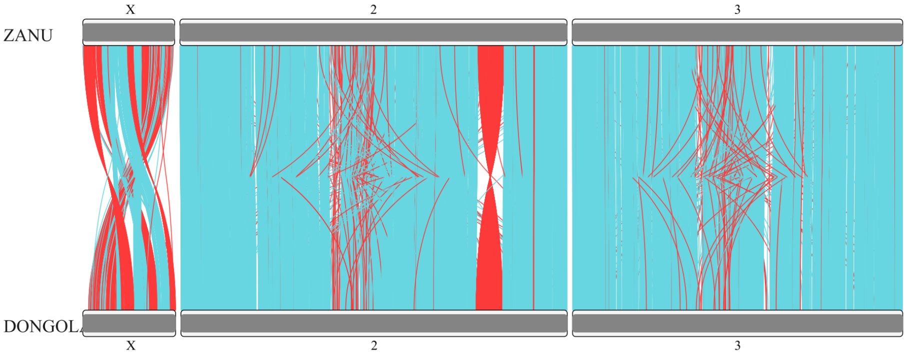
```r
#karyotype_table <- data.frame(Chr = c('X', '2', '3', 'X', '2', '3'),
#  start = rep(1),
#  end = c(27238055, 114783175, 97973315, 26913133, 111988354, 95710210),
#  fill =  rep(969696), species = c("ZANU", "ZANU", "ZANU", "DONGOLA", "DONGOLA", "DONGOLA"),
#  size = rep(12), color = rep(252525))
#head(karyotype_table)
```

```r
karyotype_table <- rbind(karyotype_table, data.frame(Chr=c('X','2','3'),
                                                     Start=c(1, 1, 1),
                                                     End=c(26913133, 111988354, 95710210),
                                                     fill='969696',
                                                     species='DONGOLA', size=12, color='252525'))
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  26913133 969696 DONGOLA   12 252525
## 5   2     1 111988354 969696 DONGOLA   12 252525
## 6   3     1  95710210 969696 DONGOLA   12 252525
```

### Plot

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


ZANU
X
2
3
DONGOLA
X
2
3