HW3

Svetlana Milrud

06 06 2022

Load the data

```
gene_mapping <- read.csv('~/HW3_R/gene_mapping.txt', sep='\t')
dongola <- read.csv('~/HW3_R/DONGOLA_genes.txt', sep='\t')
zanu <- read.csv('~/HW3_R/ZANU_genes.txt', sep='\t')
```

Data Exploration

Gene mapping

```
head(gene_mapping,n=4)
```

```
##   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
##                                                    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
```

Column description:

- contig: chromosome name in ZANU
- middle.position: position of gene center in ZANU chromosome coordinate
- strand: direction of gene in relation to chromosome scaffold direction
- ord: just an index of record
- name: gene name in ZANU
- ref.genes: how many genes are homologus to this one from ZANU
- DONG: complex string for DONGOLA gene(s) information separated by “,” for one gene and “;” between genes

For one gene this complex string has structure:

- sequence_id - id from NCBI where is this gene in DONGOLA genome (not only chromosomes here)
- middle coordinate of the gene
- strand
- length of the gene
- gene name from DONGOLA annotation.

Dongola genes

```
head(dongola,n=4)
```

```
##                  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
```

Zanu genes

```
head(zanu,n=4)
```

```
##           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
```

Editing Gene mapping dataframe

Editing DONG column

```
#create dataframe from DONG column in gene_mapping dataframe
dong <- gene_mapping$DONG
dong <- (strsplit(dong,",")) #separate by comma
dong <- as.data.frame(dong)
dong <- as.data.frame(t(dong)) # column to rows
rownames(dong) <- NULL
colnames(dong) <- c('sequence_id','middle_coordinate','strand_d','gene_length','gene_name')
```

```
# bind two dataframes and removing DONG column
gene_mapping <- cbind(gene_mapping[0:6],dong)
head(gene_mapping,n=4)
```

```
##   contig middle.position strand ord      name ref.genes sequence_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
##   middle_coordinate 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
```

Editing contig column

```
# contig includes not only 2, 3 and X chromosomes
unique(gene_mapping$contig)[0:8]
```

```
## [1] "2"                "3"                "HiC_scaffold_10"  "HiC_scaffold_104"
## [5] "HiC_scaffold_107" "HiC_scaffold_111" "HiC_scaffold_112" "HiC_scaffold_115"
```

```
# leave only only 2, 3 and X chromosomes in contig column
chromosomes <- c( "2", "3", "X")
gene_mapping$contig <- as.character(gene_mapping$contig)
gene_mapping <- gene_mapping[gene_mapping[,"contig"] %in% chromosomes,]
```

```
# check
unique(gene_mapping$contig)
```

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

Editing sequence_id column

https://www.ncbi.nlm.nih.gov/genome/?term=Anopheles%20Arabiensis%20DONGOLA

Chr	Seq id
2	NC_053517.1
3	NC_053518.1
X	NC_053519.1

```
# rename sequence_id to chromosome
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'] <- '1'

# convert X chromosome into numeric value for downstream analysis
```

```
# explore sequence_id column
unique(gene_mapping$sequence_id)[0:8]
```

```
## [1] "2"              "1"              "3"              "NW_024412154.1"
## [5] "NW_024412121.1" "NW_024412103.1" "NW_024412152.1" "NW_024412162.1"
```

```
# leave only only 2, 3 and X chromosomes in sequence_id column
chromosomes <- c( "2", "3", "1")
gene_mapping <- gene_mapping[gene_mapping[,"sequence_id"] %in% chromosomes,]
```

```
# check
unique(gene_mapping$sequence_id)
```

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

Editing gene_name column

```
# remove 'DONG_' in the beginnig of the gene_name
gene_mapping$gene_name <- lapply(gene_mapping$gene_name, sub, pattern = '^DONG_', replacement ="")
gene_mapping$gene_name <- as.character(gene_mapping$gene_name)
```

X as numeric value in ZANU

```
gene_mapping$contig <- sub("X", "1", gene_mapping$contig)
```

```
# final gene_mapping dataframe
head(gene_mapping, n=4)
```

```
##   contig middle.position strand ord      name ref.genes sequence_id
## 1      2           31135     -1   0 gene_3542         1           2
## 2      2           38868     -1   1 gene_3543         1           2
## 3      2           42746      1   2   gene_80         1           2
## 4      2           46243     -1   3 gene_3544         1           2
##   middle_coordinate strand_d gene_length         gene_name
## 1         111908344        1        6540 gene-LOC120894913
## 2         111899667        1        6539 gene-LOC120904110
## 3         111895084       -1        6538 gene-LOC120904105
## 4         111891588        1        6537 gene-LOC120904096
```

Creation of dataframe with closest Dongola and Zanu genes

```
# calculate distances between Dongola and Zanu genes
gene_mapping$distance <- abs(gene_mapping$middle.position - as.numeric(gene_mapping$middle_coordinate))
```

```
# remove rows where Dongola chromosomes not equal to Zanu chromosomes
gene_mapping<-subset(gene_mapping, contig==sequence_id)
```

```
# remove multiple Dongola genes according to closest distance
new_data<-data.frame()

unique_names<-unique(gene_mapping$gene_name)

for (i in unique_names){
  gene_collector<- gene_mapping[gene_mapping$gene_name == i, ]
  min_count<-min(gene_collector$distance)
  new_data<-rbind(new_data,gene_collector[gene_collector$distance == min_count, ])
}
new_data <- new_data[order(new_data$distance),]
```

```
# remove multiple Zanu genes according to closest distance
new_data1<-data.frame()

unique_names<-unique(new_data$name)

for (i in unique_names){
  gene_collector<- new_data[new_data$name == i, ]
  min_count<-min(gene_collector$distance)
  new_data1<-rbind(new_data1,gene_collector[gene_collector$distance == min_count, ])
}
final_mapping <- new_data1[order(new_data1$distance),]
head(final_mapping, n=4)
```

```
##       contig middle.position strand  ord       name ref.genes sequence_id
## 16445      1         7865798     -1  420 gene_13388         1           1
## 17420      1        22554898      1 1158 gene_13057         1           1
## 15952      1           14108     -1    0 gene_13164         1           1
## 17310      1        20658297      1 1063 gene_13015         1           1
##       middle_coordinate 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
```

Creating synteny__dual__comparison dataframe

```
# ZANU - Species_1
# create fill column according to strand of Zanu and Dongola: if direction
#is identical, than fill will be red (e41a1c), if not than fill will be
# gray (cccccc)
start_z <- c()
end_z <- 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, "e41a1c")
    else append(fill, "cccccc")
  start_z <- append(start_z, zanu[zanu$ID == name, "start"])
  end_z <- append(end_z, zanu[zanu$ID == name, "end"])
}
```

```
# length of X, 2, 3 chromosomes
# https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_016920715.1

don_end_2 = 111988354 # Chr2
don_end_3 = 95710210 # Chr3
don_end_1 = 26913133 # ChrX
```

```
# DONGOLA - Species_2
start_d <- c()
end_d <- c()
for (i in (1:nrow(final_mapping))){
    name <- final_mapping[i, "gene_name"]
    if (final_mapping[i, "contig"] ==1){
    start <- don_end_1 - dongola[dongola$ID == name, "start"]
    end <- don_end_1 - dongola[dongola$ID == name, "end"]
    } else if ((final_mapping[i, "contig"] ==2)){
      start <- don_end_2 - dongola[dongola$ID == name, "start"]
      end <- don_end_2 - dongola[dongola$ID == name, "end"]
    } else {
      start <- don_end_3 - dongola[dongola$ID == name, "start"]
      end <- don_end_3 - dongola[dongola$ID == name, "end"]
    }
  start_d <- append(start_d, start)
  end_d <- append(end_d, end)
}
```

```
# create synteny_dual_comparison dataframe
synteny_dual_comparison <- data.frame(Species_1 = as.numeric(final_mapping$contig),
 Start_1 = start_z, End_1 = end_z, Species_2 = as.numeric(final_mapping$sequence_id),
Start_2 = start_d, End_2 = end_d, fill =fill)

head(synteny_dual_comparison, n=4)
```

```
##   Species_1  Start_1    End_1 Species_2  Start_2    End_2   fill
## 1         1  7865247  7866349         1 19055658 19054278 cccccc
## 2         1 22553805 22555991         1  4351086  4349049 cccccc
## 3         1     5022    23194         1 26894161 26861576 e41a1c
## 4         1 20657888 20658706         1  6238316  6237208 cccccc
```

Creating karyotype_dual_comparison dataframe

```
# similar to https://cran.r-project.org/web/packages/RIdeogram/vignettes/RIdeogram.html

karyotype_dual_comparison <- data.frame(Chr = c('X', '2', '3', 'X', '2', '3'),
Start = rep(1,6),
End = c(27238055, 114783175, 97973315, 26913133, 111988354, 95710210),
fill =  rep(969696,6), species = c("ZANU", "ZANU", "ZANU", "DONGOLA", "DONGOLA", "DONGOLA"),
size = rep(12,6), color = rep(252525,6))
head(karyotype_dual_comparison,n=4)
```

```
##   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
```

Synteny between ZANU and DONGOLA

```
ideogram(karyotype = karyotype_dual_comparison, synteny = synteny_dual_comparison)
convertSVG("chromosome.svg", device = "png")
```

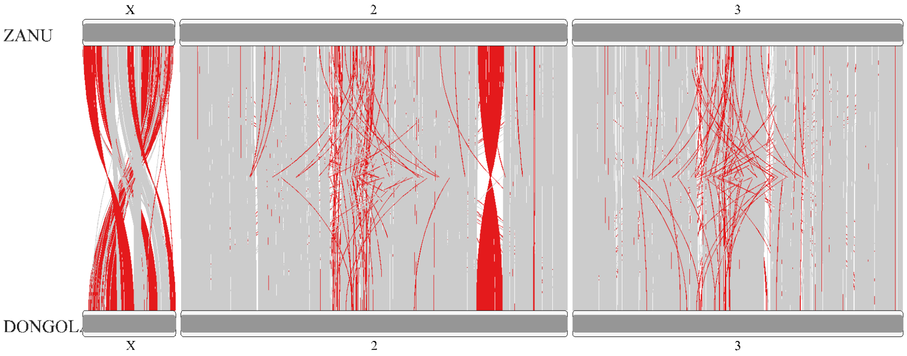

Figure 1: Synteny between ZANU and DONGOLA