hw3

Amosov Artem

01 06 2022

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

#make karyotype table
karyotype <- data.frame('Chr' = c('X', 2, 3, 'X', 2, 3),
                  'Start' = c(rep(1, 6)),
                  'End' = c(27238055, 114783175, 97973315, 26910000, 111990000, 95710000),
                  'fill' = c(rep('chartreuse', 3), rep('cornflowerblue', 3)),
                  'species' = c(rep('Zanu', 3), rep('Dongola', 3)),
                  'size' = rep(12, 6),
                  'color' = rep(252525, 6))
```

```
# Make columns from DONG string:
mapping$d_id <- str_extract(mapping$DONG, '(\\w*\\.\\d{1})')

mapping$mid <- as.numeric(str_extract(mapping$DONG, '(\\d{7,11})'))

mapping$d_strand <- str_extract(mapping$DONG, '(?<=\\,{1})\\-?\\d{1}(?=\\,+)')

mapping$len <- str_extract(mapping$DONG, '(?<=\\,{1})\\d{3,5}(?=\\,+)')

mapping$dong_name <- str_extract(mapping$DONG, '(?<=DONG_)gene-[A-Z]{3}\\d*')

# delete raw Dongola string
mapping <- mapping[-7]
```

```
#Take only chr 2, 3, X
ids <- unique(mapping$d_id)
unplaced_d <- ids[-1:-3]

unplaced_z <- unique(mapping$contig)[-1:-2]
unplaced_z <- unplaced_z[-82]

mapping <- subset(mapping, mapping$d_id %in% unplaced_d == FALSE)
mapping <- subset(mapping, mapping$contig %in% unplaced_z == FALSE)

#Change ids to chr names
mapping[mapping$d_id == 'NC_053517.1', ]['d_id'] = 2
mapping[mapping$d_id == 'NC_053518.1', ]['d_id'] = 3
mapping[mapping$d_id == 'NC_053519.1', ]['d_id'] = 'X'
```

```r
# take only genes from same chromosomes
mapping <- mapping[mapping$contig == mapping$d_id,]


for(i in 1:nrow(mapping)) {
    row <- mapping[i, ]

    geneZ <- row$name
    geneD <- row$dong_name
    rowD <- dongola[dongola$ID == geneD,]
    rowZ <- zanu[zanu$ID == geneZ,]

    # Add starts and ends for zanu:
    mapping[i, 'z_start'] <- rowZ$start
    mapping[i, 'z_end'] <- rowZ$end

    chr2 <- karyotype[(karyotype['species'] == 'Dongola')&(karyotype['Chr'] == '2'),]
    chr3 <- karyotype[(karyotype['species'] == 'Dongola')&(karyotype['Chr'] == '3'),]


    if (row$contig == 2 || row$contig == 3)
      {
        mapping[i, 'fill'] <- ifelse(row$strand == row$d_strand, 'db4527', '5891bf')
# picture for chr 2 and 3 looked like reversed, that's why we will reverse coordinates in these chromos
        mapping[i, 'd_start'] <- ifelse(row$contig == 2, chr2$End - rowD$start, chr3$End - rowD$start)
        mapping[i, 'd_end'] <- ifelse(row$contig == 2, chr2$End - rowD$end, chr3$End - rowD$end)
      }
    else
      {
        mapping[i, 'fill'] <- ifelse(row$strand == row$d_strand, '5891bf', 'db4527')
        mapping[i, 'd_start'] <- rowD$start
        mapping[i, 'd_end'] <- rowD$end
      }

}
head(mapping)
```

```
##   contig middle.position strand ord      name ref.genes d_id       mid d_strand
## 1      2           31135     -1   0 gene_3542         1    2 111908344        1
## 2      2           38868     -1   1 gene_3543         1    2 111899667        1
## 3      2           42746      1   2   gene_80         1    2 111895084       -1
## 4      2           46243     -1   3 gene_3544         1    2 111891588        1
## 5      2           53442     -1   4 gene_3545         1    2 111884408        1
## 6      2           60574      1   5   gene_81         1    2 111877309       -1
##    len         dong_name z_start z_end   fill d_start  d_end
## 1 6540 gene-LOC120894913   29035 33235 5891bf   85776  79809
## 2 6539 gene-LOC120904110   37467 40269 5891bf   91997  86783
## 3 6538 gene-LOC120904105   41638 43855 5891bf   96116  93507
## 4 6537 gene-LOC120904096   44541 47945 5891bf  100784  96331
## 5 6536 gene-LOC120895288   50702 56183 5891bf  110050 101480
## 6 6535 gene-LOC120895290   58892 62256 5891bf  114945 110622
```

```
final <- mapping[c('contig', 'z_start', 'z_end','d_id', 'd_start','d_end','fill')]
colnames(final) = c("Species_1", "Start_1", "End_1", "Species_2", "Start_2", "End_2", "fill")

final[final$Species_1 == 'X', ]['Species_1'] = '1'
final$Species_1 <- as.numeric(final$Species_1)
final[final$Species_2 == 'X', ]['Species_2'] = '1'
final$Species_2 <- as.numeric(final$Species_2)
```

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