Homework 3

Alexey Serdyukov

2022-06-16

```
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
```

```
gene_mapping <- read.csv('gene_mapping.tsv', sep='\t')
dong_genes <- read.csv('dongola_genes.tsv', sep='\t')
zanu_genes <- read.csv('ZANU_genes.tsv', sep='\t')
```

Clean and merge data

```
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
```

```
gene_mapping_clean <- gene_mapping %>%
  tidyr::separate(
    col = DONG,
    into = c(
      "dong_ncbi_id",
      "dong_middle",
```

```
      "dong_strand",
      "dong_length",
      "dong_name"
    ),
    sep=",",
    convert = TRUE
  ) %>%
  rename(
    zanu_name = name,
    zanu_strand = strand,
    zanu_middle = middle.position,
    ref_genes = ref.genes
  ) %>%
  # Map NCBI ids to chromosome names
  mutate(
    dong_contig = recode(
      dong_ncbi_id,
      `NC_053519.1` = "X",
      `NC_053517.1` = "2",
      `NC_053518.1` = "3",
    ),
    dong_name = stringr::str_remove(dong_name, "DONG_")
  ) %>%
  # Keep only required contigs
  filter(
    contig %in% c("X", "2", "3") &
      dong_contig == contig
  ) %>%
  select(
    contig,
    zanu_name,
    zanu_strand,
    dong_name,
    dong_strand,
    zanu_middle,
    dong_middle,
    ref_genes
  )

head(gene_mapping_clean)
```

```
##   contig zanu_name zanu_strand         dong_name dong_strand zanu_middle
## 1      2 gene_3542          -1 gene-LOC120894913           1       31135
## 2      2 gene_3543          -1 gene-LOC120904110           1       38868
## 3      2   gene_80           1 gene-LOC120904105          -1       42746
## 4      2 gene_3544          -1 gene-LOC120904096           1       46243
## 5      2 gene_3545          -1 gene-LOC120895288           1       53442
## 6      2   gene_81           1 gene-LOC120895290          -1       60574
##   dong_middle ref_genes
## 1   111908344         1
## 2   111899667         1
## 3   111895084         1
## 4   111891588         1
## 5   111884408         1
```

```
## 6   111877309         1
```

Keep closest of multimapped genes

```
print(sum(duplicated(gene_mapping_clean$zanu_name)))
```

```
## [1] 2668
```

```
print(sum(duplicated(gene_mapping_clean$dong_name)))
```

```
## [1] 3311
```

```
gene_mapping_dedup <- gene_mapping_clean %>%
  mutate(distance = abs(zanu_middle - dong_middle)) %>%
  group_by(zanu_name) %>%
  slice_min(order_by = distance) %>%
  group_by(dong_name) %>%
  slice_min(order_by = distance) %>%
  ungroup() %>%
  select(-distance)
head(gene_mapping_dedup)
```

```
## # A tibble: 6 x 8
##   contig zanu_name zanu_strand dong_name     dong_strand zanu_middle dong_middle
##   <chr>  <chr>           <int> <chr>               <int>       <int>       <int>
## 1 2      gene_5019          -1 gene-LOC1208~           1    48531603    65514822
## 2 2      gene_6182          -1 gene-LOC1208~           1    86040949    28681053
## 3 2      gene_2643           1 gene-LOC1208~          -1    86040395    28681607
## 4 2      gene_5313          -1 gene-LOC1208~           1    58398932    55921684
## 5 2      gene_2537           1 gene-LOC1208~          -1    82790246    31941591
## 6 2      gene_5008          -1 gene-LOC1208~          -1    48220819    60987618
## # ... with 1 more variable: ref_genes <int>
```

No duplicates left:

```
print(sum(duplicated(gene_mapping_dedup$zanu_name)))
```

```
## [1] 0
```

```
print(sum(duplicated(gene_mapping_dedup$dong_name)))
```

```
## [1] 0
```

Ideogram

Karyotype

```
dong_len_2 = 111990000L
dong_len_3 = 95710000L
dong_len_X = 26910000L
karyotype <- data.frame(
  Chr = rep(c("X", "2", "3"), 2),
  Start = rep(1L, 6),
  End = c(
    27238055L,
    114783175L,
```

```
    97973315L,
    dong_len_X,
    dong_len_2,
    dong_len_3
  ),
  fill = "777777",
  species = c(rep("ZANU", 3), rep("DONG", 3)),
  size = 12L,
  color = "000000"
)
karyotype
```

```
##   Chr Start       End   fill species size  color
## 1   X     1  27238055 777777    ZANU   12 000000
## 2   2     1 114783175 777777    ZANU   12 000000
## 3   3     1  97973315 777777    ZANU   12 000000
## 4   X     1  26910000 777777    DONG   12 000000
## 5   2     1 111990000 777777    DONG   12 000000
## 6   3     1  95710000 777777    DONG   12 000000
```

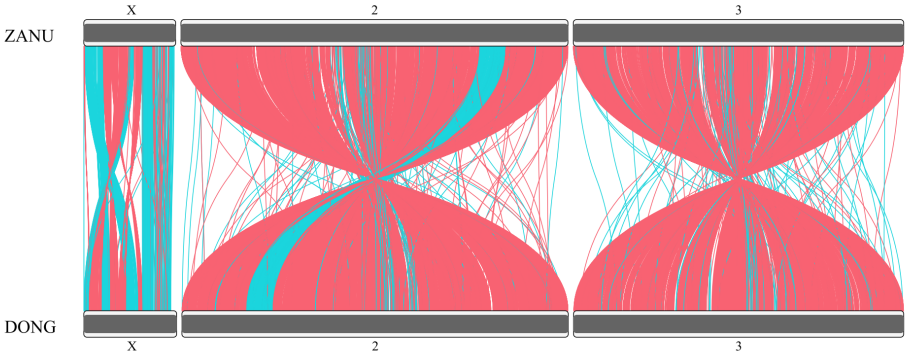
Synteny

```
synteny <- gene_mapping_dedup %>%
  mutate(
    Species_1 = recode(contig, `X` = 1L, `2` = 2L, `3` = 3L),
    Species_2 = Species_1,
    fill = ifelse(zanu_strand == dong_strand, "00DBE4", "FC7A85")
  ) %>%
  merge(zanu_genes, by.x = "zanu_name", by.y = "ID") %>%
  merge(dong_genes, by.x = "dong_name", by.y = "ID", suffixes = c("_zanu", "_dong")) %>%
  rename(
    zanu_start = start_zanu,
    zanu_end = end_zanu,
    dong_start = start_dong,
    dong_end = end_dong
  ) %>%
  select(
    Species_1,
    Start_1 = zanu_start,
    End_1 = zanu_end,
    Species_2,
    Start_2 = dong_start,
    End_2 = dong_end,
    fill
  )
head(synteny)
```

```
##   Species_1  Start_1    End_1 Species_2  Start_2    End_2   fill
## 1         2 48528403 48534803         2 65511152 65519724 FC7A85
## 2         2 86040710 86041188         2 28680597 28681368 FC7A85
## 3         2 86040192 86040598         2 28681316 28681908 FC7A85
## 4         2 58381587 58416277         2 55853085 55941166 FC7A85
## 5         2 82789431 82791062         2 31940683 31942410 FC7A85
## 6         2 48219362 48222277         2 60986210 60989026 00DBE4
```

Plot

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


X
2
3
ZANU
DONG
X
2
3