R Notebook

Dear students, this is your third homework.

Sinteny blocks by gene homology.

Studying comparative genomics we are often interesting in visualisation of synteny blocks to look at structural variations between different species.

If you are not familiar with this concept you can start with en.wikipedia.org/wiki/Synteny

There are different ways to construct synteny blocks. We can use pairwise alignments of genomes, complicated structures of genome fragmentation and so on.

In this assignment we want to construct synteny using genes homology.

In this homework I want to teach you how to: * work with real data, how to clean it and reshape for you purposes. * work with unfamiliar R libraries, how to use its documentation to achieve result. * work with aditional data sources

So the goal of this work is to create a .png picture of gene synteny between chromosomal scaffolds of two mosquito species Anopheles Gambiae ZANU and Anopheles Arabiensis DONGOLA.

Here you can see the example of such figure for another pair of mosquitoes ZANU and PEST strains of An.Gambiae:

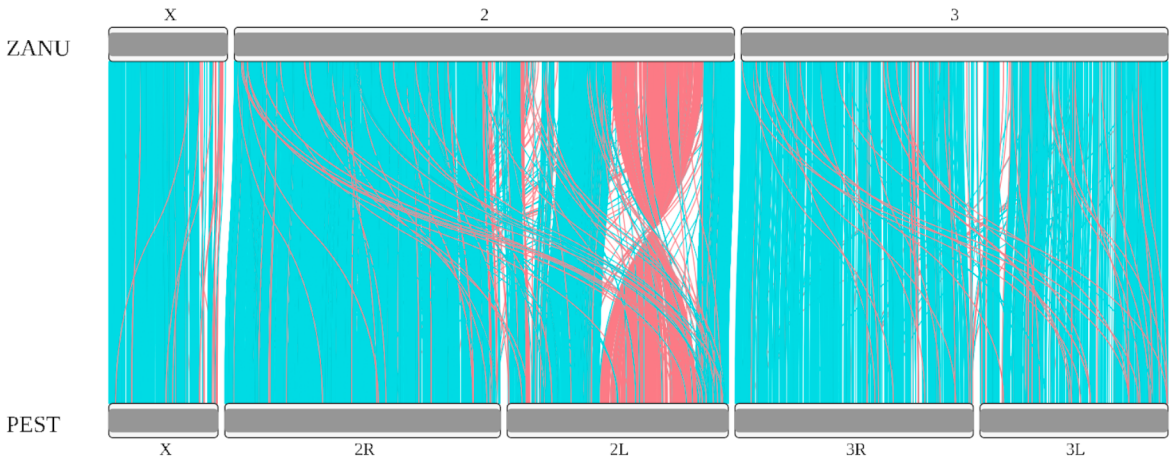

Figure 1: synteny

So here we have two stripes of mosquitoes chromosomes X, 2 and 3 (they have only 3) and between them we have blue and red lines. Each line is for pair of genes that are homologus.

It is blue when in both mosquitoes genes are in the same direction and red in opposit.

Your task is to construct the same picture from input data using RIdeogram package.

Please find how to install it and documentation by yourself.

Input description

First genome (ZANU) is a new assembled genome and in purpose to construct synteny it was annotated using annotation from the second An.Arabiensis DONGOLA genome (with GeMoMa pipeline)

1. Anopheles Arabiensis DONGOLA genome description is in NCBI database.

There you can find mapping between chromosome names and sequence IDs, and information about chromosome lengths

2. Chromosome lengths for ZANU are:

X: 27238055

2: 114783175

3: 97973315

3. You have data table with results of gene mapping between two species:

```
str(read.csv('gene_mapping.tsv', sep='\t'))
```

```
## 'data.frame':    17643 obs. of  7 variables:
##  $ contig         : chr  "2" "2" "2" "2" ...
##  $ middle.position: int  31135 38868 42746 46243 53442 60574 97823 107787 112233 115544 ...
##  $ strand         : int  -1 -1 1 -1 -1 1 1 -1 -1 1 ...
##  $ ord            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ name           : chr  "gene_3542" "gene_3543" "gene_80" "gene_3544" ...
##  $ ref.genes      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ DONG           : chr  "NC_053517.1,111908344,1,6540,DONG_gene-LOC120894913" "NC_053517.1,11189966
```

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

4. You have two tables that describe genes for ZANU and DONGOLA:

```
str(read.csv("DONGOLA_genes.tsv", sep='\t'))
```

```
## 'data.frame':    15289 obs. of  4 variables:
##  $ ID    : chr  "gene-LOC120906950" "gene-LOC120906947" "gene-LOC120906949" "gene-LOC120906948" ...
##  $ start : int  59885 61728 88010 90190 657 23986 26497 51121 53297 69497 ...
```

```
##  $ end   : int  60345 64249 88555 90789 1316 24588 26764 51663 53859 70001 ...
##  $ strand: int  -1 1 -1 -1 -1 1 1 -1 -1 1 ...
```

two tables ZANU_genes.tsv and DONGOLA_genes.tsv have the same structure:

- ID: gene_name in species annotation
- start: gene start coordinate in chromosome coordinate system
- end: gene end coordinate
- strand: gene direction in relation with chromosomal scaffold direction

All of these information will be enough to construct synteny using RIdeogram

Result needed:

.png file with synteny between ZANU and DONGOLA X, 2, and 3 chromosomes.

You should visualize only genes that maps only between the same chromosomes: X-X, 2-2, 3-3

Genes with the same direction and genes that change direction between genomes must be colored differently. That allows us to see borders of inversions.

Task with * (additional complexity)

If there are multiple DONGOLA genes for one ZANU gene you can take only one which coordinates are closest.