Protocol

Required programing languages and tools

- Perl 5.18.2
- Python 2.7.13
- EMBOSS:6.6.0.0 (use seqret program in the package)
- Genbank_splicer.py (error occurs when we run with python3) https://github.com/jrjhealey/bioinfo-tools/blob/master/Genbank_slicer.py
- sed
- Easyfig 2.2.2 (http://mjsull.github.io/Easyfig/)
- (optional) inkscape 0.92.2

1. Preparation of input files (Genbank format) for Easyfig

First, we extract annotation information for each single chromosome from the annotation file (annotation.gff3).

```
$  perl -ne 'print if/^1/' data/annotation.gff3 >  1.gff3
$  perl -ne 'print if/^2/' data/annotation.gff3 >  2.gff3
$  perl -ne 'print if/^3/' data/annotation.gff3 >  3.gff3
$  perl -ne 'print if/^4/' data/annotation.gff3 >  4.gff3
$  perl -ne 'print if/^5/' data/annotation.gff3 > 5.gff3
```

We merge DNA sequence information (fasta format) and annotation information (GFF3 format) of each single chromosome (annotation.gff3) into a single genbank file using the 'seqret' program in the EMBOSS package.

```
$  seqret -sequence data/1.fa -feature -fformat gff -fopenfile 1.gff3 -osformat genbank
-outseq 1.gbk -osdirectory ./
$  seqret -sequence data/2.fa -feature -fformat gff -fopenfile 2.gff3 -osformat genbank
-outseq 2.gbk -osdirectory ./
$  seqret -sequence data/3.fa -feature -fformat gff -fopenfile 3.gff3 -osformat genbank
-outseq 3.gbk -osdirectory ./
$  seqret -sequence data/4.fa -feature -fformat gff -fopenfile 4.gff3 -osformat genbank
-outseq 4.gbk -osdirectory ./
$  seqret -sequence data/5.fa -feature -fformat gff -fopenfile 5.gff3 -osformat genbank
-outseq 5.gbk -osdirectory ./
```

Then, extract DNA sequence and annotation information within a specific chromosomal region using the Genbank_splicer.py python script.

```
$  python2 scripts/Genbank_slice.py -g 1.gbk -o 1_pnr700.gbk -s 1842912 -e 2542912 -v
$  python2 scripts/Genbank_slice.py -g 2.gbk -o 2_pnr700.gbk -s 1817796 -e 2517796 -v
$  python2 scripts/Genbank_slice.py -g 3.gbk -o 3_pnr700.gbk -s 585879 -e 1285879 -v
$  python2 scripts/Genbank_slice.py -g 4.gbk -o 4_pnr700.gbk -s 1576680 -e 2276680 -v
$  python2 scripts/Genbank_slice.py -g 5.gbk -o 5_pnr700.gbk -s 1255217 -e 1955217 -v
```

Finally, we replace the "misc_RNA" feature in our annotation file for the "gene" feature so that the annotation information will be recognized in the later drawing process by Easyfig software.

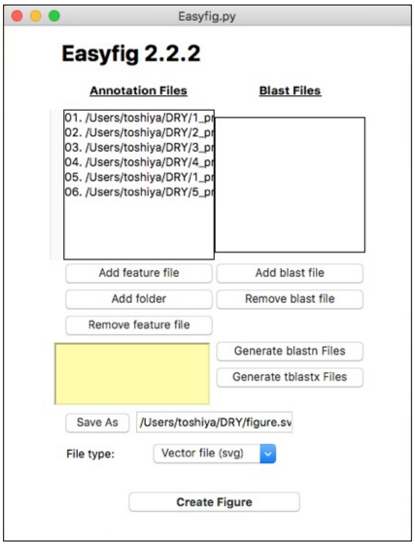
```
$  sed -i '' 's/misc_RNA/gene/' 1_pnr700.gbk
$  sed -i '' 's/misc_RNA/gene/' 2_pnr700.gbk
$  sed -i '' 's/misc_RNA/gene/' 3_pnr700.gbk
$  sed -i '' 's/misc_RNA/gene/' 4_pnr700.gbk
$  sed -i '' 's/misc_RNA/gene/' 5_pnr700.gbk
```

Drawing homologous regions between chromosomes using Easyfig

First, we launch Easyfig.

```
$  /Applications/Easyfig_2.2.2_OSX/Easyfig
```

Then, we register input files used in drawing a figure by clicking the "Add feature file" button.



Easyfig 2.2.2

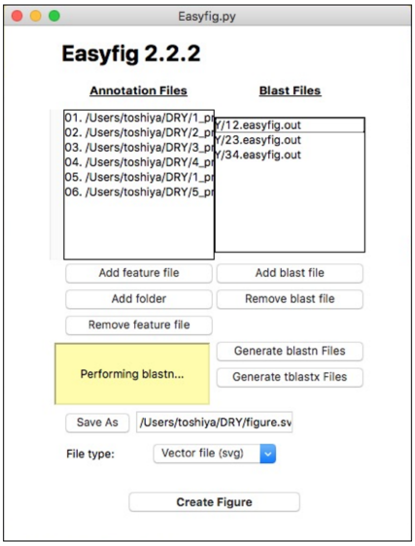
Annotation Files

01. /Users/toshiya/DRY/1_pr
02. /Users/toshiya/DRY/2_pr
03. /Users/toshiya/DRY/3_pr
04. /Users/toshiya/DRY/4_pr
05. /Users/toshiya/DRY/1_pr
06. /Users/toshiya/DRY/5_pr

Save As: /Users/toshiya/DRY/figure.sv

File type: Vector file (svg)

We can identify homologous regions between the chromosomal regions of interest using built-in blastn program by clicking the "Generate blastn files" button.



01. /Users/toshiya/DRY/1_pr
02. /Users/toshiya/DRY/2_pr
03. /Users/toshiya/DRY/3_pr
04. /Users/toshiya/DRY/4_pr
05. /Users/toshiya/DRY/1_pr
06. /Users/toshiya/DRY/5_pr

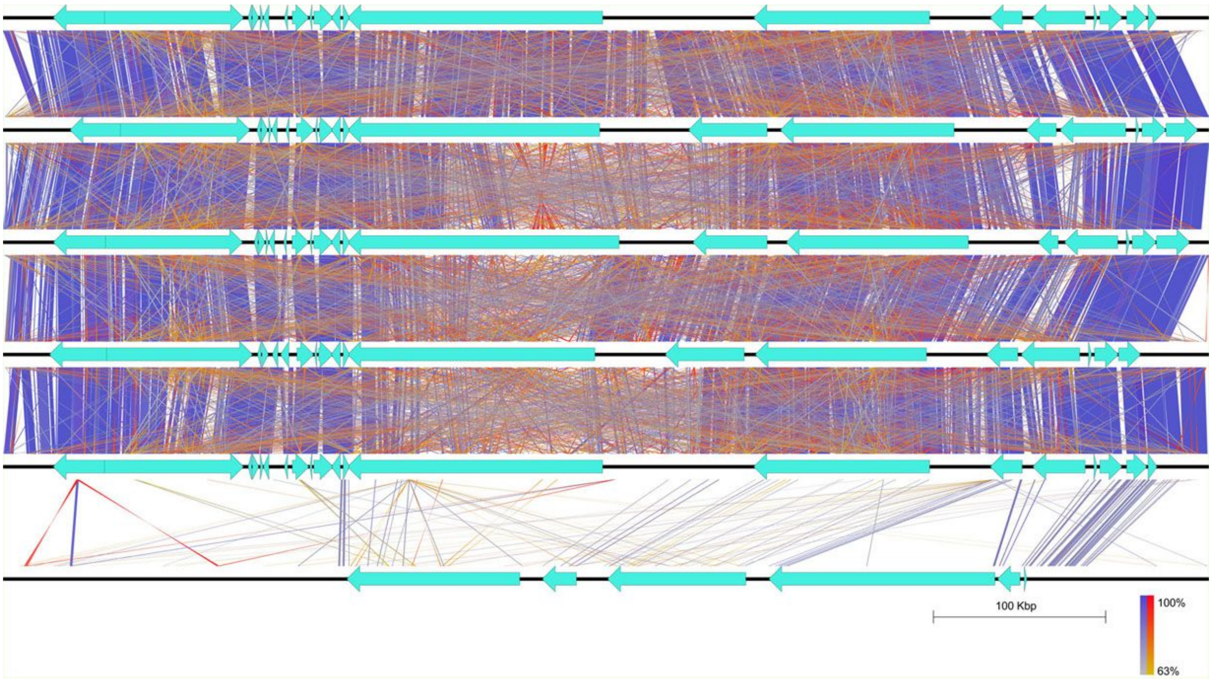
Y/12.easyfig.out
Y/23.easyfig.out
Y/34.easyfig.out

Performing blastn...

Save As: /Users/toshiya/DRY/figure.sv

File type: Vector file (svg)

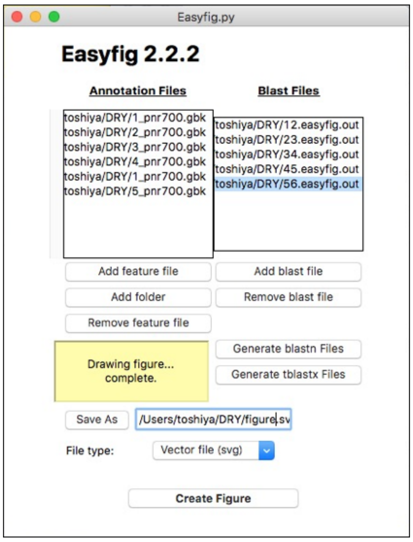
Raw blastn result files include repeat sequences that will generate noisy background in the final figure as in the figure below.



Therefore, we need to filter blast hits at repetitive sequences using the filter_repeat.pl Perl script.

```
$  perl scripts/filter_repeat.pl -i ./
$  perl scripts/filter_repeat.pl -s -i ./56.easyfig.out -m 250
```

After filtering blast hits at repeat sequences, we remove the unfiltered Blast result files by clicking the "Remove blast file" button. Then, we register the filtered Blast result files by clicking the "Add blast file" button.



Easyfig 2.2.2

Annotation Files

toshiya/DRY/1_pnr700.gbk
toshiya/DRY/2_pnr700.gbk
toshiya/DRY/3_pnr700.gbk
toshiya/DRY/4_pnr700.gbk
toshiya/DRY/1_pnr700.gbk
toshiya/DRY/5_pnr700.gbk

Blast Files

toshiya/DRY/12.easyfig.out
toshiya/DRY/23.easyfig.out
toshiya/DRY/34.easyfig.out
toshiya/DRY/45.easyfig.out
toshiya/DRY/56.easyfig.out

Drawing figure...
complete.

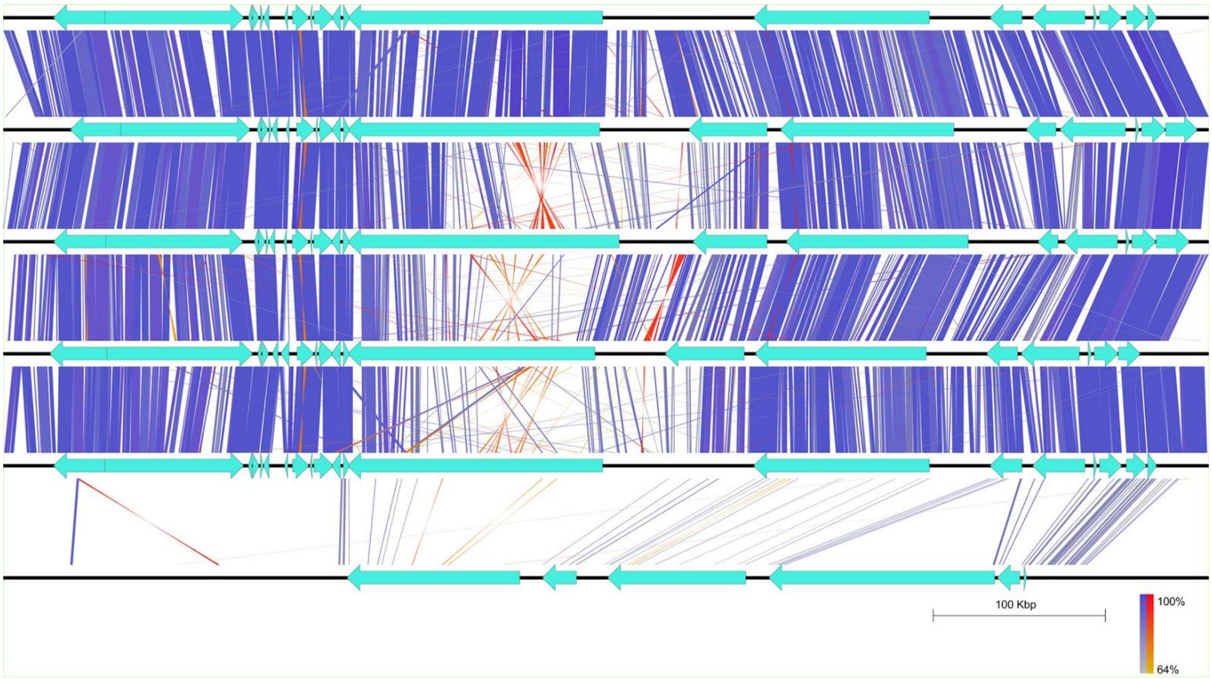
Save As: /Users/toshiya/DRY/figure.sv

File type: Vector file (svg)

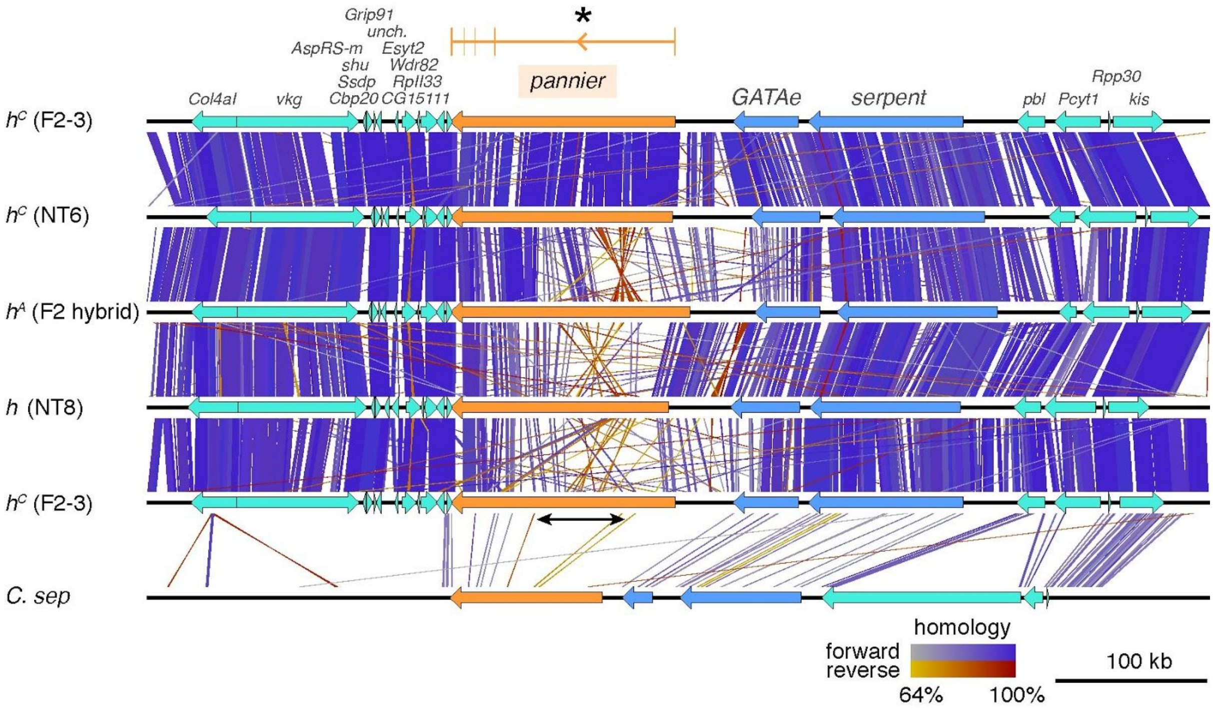
Finally, we change drawing settings for the present dataset as below,

```
------------------------------------------------------------------------------------------------------------------------
Menu bar: Image>Figure
        Width of Figure: 5000 > 7000
        Length of scale legend (in base pairs): 0 > 100000
Menu ba: Image>Blast
        Choose minimum blast colour: inverted, Gray > Yellow
        Choose minimum blast colour: normal, Dark_gray > Blue, inverted, Dark_grey > Red
        Outline blast hits in black: checked > unchecked
Save As :  File name (or File path)
File type : svg
------------------------------------------------------------------------------------------------------------------------
```

and export the result figure by clicking the "Create Figure" button.



The result image file (svg format) can be imported to the free illustration software inkscape. We can manually modify colors of the different classes of genes, and insert annotations as below.


Grip91
unch.
AspRS-m
Esyt2
shu
Wdr82
Ssdp
RpII33
Cbp20
CG15111
Col4al
vkg
*
pannier
GATAe
serpent
pbl
Pcyt1
Rpp30
kis
h^C (F2-3)
h^C (NT6)
h^A (F2 hybrid)
h (NT8)
h^C (F2-3)
C. sep
homology
forward
reverse
100 kb