NeoChr

NeoChr is used to construct new chromosome denovo. It would assist users to grab related genes in different pathways of various organism manually, to rewire genes' relationship logically*, and to replace genes with orthologs that score higher*. Then it would allow users to define gene order and orientation in DRAG&DROP way, and decouple the genes with overlap. In the end, it would add or delete features, such as encrypted watermarks*, telomere, loxp sites to build a brand new genome.

Note:
*These function are unavailable now, and are explained in Next Version.

Plugin Scripts

This module contains three plugins: Decouple.pl, Add.pl and Delete.pl.

2.1 Decouple.pl

This plugin is to decouple the genes which have overlap regions. These overlapping genes can be decoupled if meet the following conditions: (1)One gene's 5'UTR does not cover another gene's initial codon (ATG); (2)Overlapping region initial coordinate is in the coding DNA sequences(CDS) of gene which is need to be decoupled; (3)The decouple site of CDS have synonymous substitute codon to replace; After decoupling, we use these non-redundancy genes to generate a GFF file and a FASTA file.

2.1.1 Internal operation

First, this plugin extracts base sequence from the genome file according to the gene order list, and records the gene order in the list. And then plugin records the annotation information according to the specie GFF file, moreover, plugin extends gene CDS upstream 600bp as 5'-UTR and downstream 100bp as 3'-UTR if the GFF file does not contain annotated these two features.

Second, this plugin detects the overlapping genes in the same chromosome. In case the overlapping genes are detected, it will judge whether the overlapping initial site is located in the CDS region, and identify the site is belong to phase0/1/2.

Third, the plugin attempts to synonymous substitute codon to break the initial codon intra the CDS. Printing information whether or not be decoupled successfully, such as:

```
YDR512C and YDR513W can not be decoupled in the 646
YIL177C and YIL177W-A can not be decoupled in the 963
YIL172C and YIL171W-A are decoulped successfully in the 893
```

And non-redundancy genes are generated.

Finally, the plugin links non-redundancy genes to construct a new chromosome according to the gene order.

2.1.2 Example

We have two input forms to execute the plugin:

Using string format as gene order list input form:

perl GeneDecouple.pl --species saccharomyces_cerevisiae_chr --list_format string --gene_order="YAL054C -,YAL038W +,YBR019C -,YBR145W +,YCL040W +,YCR012W +,YCR105W +,YDL168W +,YPL017C -,YIL177C -,YIL177W-A +,YIL172C -,YIL171W-A +," --geneset_dir ../gene_set --upstream_extend 600 --downstream_extend 100 --neo_chr_gff neochr.gff --neo_chr_fa neochr.fa

Using file format as gene order list input form:

perl GeneDecouple.pl --species saccharomyces_cerevisiae_chr --list_format file --gene_order gene_ordre.list --geneset_dir ../gene_set --upstream_extend 600 --downstream_extend 100 --neo_chr_gff neochr.gff --neo_chr_fa neochr.fa

2.1.3 Parameters

Parameter	Description	Default	Selectable range
species	set the species name (general use latin name)		
list_format	set the input form of gene order list	string	string/file
gene_order	set the input gene order list file(include pathway genes and addition genes)		
geneset_dir	set the species annotation directory	600	
upstream_extend	set the length of gene downstram(bp)	100	
neo_chr_gff	set the name of output neochr gff file		
neo_chr_fa	set the name of output neochr fasta file		
help	Show help information		

2.1.4 The format of output file

The output files are standard GFF and FASTA format files which are decoupled.

1.decoupled GFF file

```
NeoChr  Genovo  gene    31513   33982   .       -       .       ID=YIL172C;display=Alpha-glucosidase;
NeoChr  Genovo  3UTR    31513   31612   .       -       .       Parent=YIL172C;
NeoChr  Genovo  mRNA    31613   33382   .       -       .       Parent=YIL172C;
NeoChr  Genovo  CDS     31613   33382   .       -       .       Parent=YIL172C;
NeoChr  Genovo  decouple        33089   33089   .       .       .       Parent=YIL172C;
NeoChr  Genovo  5UTR    33383   33982   .       -       .       Parent=YIL172C;
```

2.decoupled FASTA file

```
>NeoChr
AGAGAAGGTGAAATAATAATAAGTAAGCAGCTCGGTTATAAGAGAACAAAAACACACGAAAAAAAAAAAAGTCGTCAATATAAAAAG
TTACAACTTGACCGAATCAATTAGATGTCTAACAATGCCAGGGTTTGACAATGTAGAAACGTCGCCTAGTTGGTCACTTTCTCCTG
AAAATACGTCTCATAATTTTGCCGGATCTTGTCTTGGGCAAGTCATCCACTAAAATGATCAATTTTGGTGCGGCAAATGGCCCGAT
TAAAGACCAAATGCTTCTTGATATCTTGTAATTCATCATCTGTTGCGGTGGACCAACTAGATTTGTTTTTCAACACCACAAATGCA
AGTCAAGTCATCGTTGAATCCGACAACAGCACACTCGGCCACAATTGGATCTTCGATAATAGCAGCCTCAATTTCAGCGGTAGACA
```

2.2 Add.pl

This plugin will add the LoxPsym sequence and the customized left and right telomeres, centromere and autonomously replicating sequence (ARS) into the FASTA file and GFF file which are generated by Decouple.pl.

2.2.1 Internal operation

The plugin adds LoxPsym behind the first 3bp of 3'-UTR in each gene and adds telomere, centromere and ARS according this mode:

left_telomere + gene1 + centromere + gene2 + ARS + gene3 + right_telomere

The distance between centromere and ARS is less than 30Kb.

Finally, user can see the new added features chromosome according to the JBrowse.

2.2.2 Example

perl 04.Add.pl --loxp loxPsym.feat --left_telomere UTC_left.feat --right_telomere UTC_right.feat --ars chromosome_I_ARS108.feature --centromere chromosome_I_centromere.feat --chr_gff neochr.gff --chr_seq neochr.fa --neochr_seq neochr.final.fa --neochr_gff neochr.final.gff

All the feature file format is 4 lines format, for example:

name = site_specific_recombination_target_region

type = loxPsym

source = BIO

sequence = ATAACTTCGTATAATGTACATTATACGAAGTTAT

Note: the first line is the detail name of feature, the second line is the type of feature, the third line is the source of feature and the last line is the sequence of feature.

2.2.3 Parameters

Parameter	Description	Default	Selectable range
loxp	set the sequence of loxp	ATAACTTCGTATAATGTATGCTATACGAAGTTAT	
left_telomere	set the sequence of left telomere		
right_telomere	set the sequence of right telomere		
chr_gff	set the input neorchr_gff file		
chr_seq	set the input neorchr_gff file		
neochr_seq	set the name of output added loxps and telomeres neochr_fa file		
neochr_gff	set the name of output added loxps and telomeres neochr_gff file		

2.2.4 The format of output

The output files are standard GFF and FASTA format of adding features chromosome.

added features GFF file

```
NeoChr  Genovo  left_telomere   1       689     .       +       .       ID=universal_telomere_cap_left;
NeoChr  Genovo  gene    690     3565    .       -       .       ID=YAL054C;display=Acetyl-coA_synthetase_isoform;
NeoChr  Genovo  3UTR    690     823     .       -       .       Parent=YAL054C;
NeoChr  Genovo  loxp    693     727     .       -       .       ID=site_specific_recombination_target_region;Parent=YAL054C;
NeoChr  Genovo  mRNA    824     2965    .       -       .       Parent=YAL054C;
NeoChr  Genovo  CDS     824     2965    .       -       .       Parent=YAL054C;
NeoChr  Genovo  5UTR    2966    3565    .       -       .       Parent=YAL054C;
```

2.3 Delete.pl

This plugin can modify the GFF and FASTA file which are generated by Add.pl according to the user drags a window in the JBrowse and delete any gene in the window.

2.3.1 Internal operation

Firstly, user uses mouse to drag a window in the added features FASTA file which is showed in the JBrowse and JBrowse displays all the genes in this window.

(需要 JBrowse 截图)

Secondly, user decides which genes is need to be delected from the new chromosome and plugin deletes genes from GFF file and modify FASTA in the same time.

2.3.2 Example

perl 05.delete.pl --delete="YAL054C,YAL038W" --neochr_gff neochr.refine.final.gff --neochr_fa neochr.refine.final.fa --slim_gff neochr.refine.delete.gff --slim_fa neochr.refine.delete.fa

2.3.3 Parameters

Parameter	Description	Default	Selectable range
delete	Set the to be deleted gene list		
neochr_gff	Set the input GFF file which is generated by Add.pl		
neochr_fa	Set the input FASTA file which is generated by Add.pl		
slim_gff	Set the output GFF file		
slim_fa	Set the output FASTA file		

2.3.4 The format of ouput

The output files are standard GFF and FASTA format of deleted genes chromosome.