MCScanX's manual

Overview

The MCScanX package has two main components: 1) a modified version of MCScan algorithm allowing users to conveniently conduct synteny and collinearity detection and to clearly view multiple alignments of collinear blocks, and 2) a variety of tools to visualize and analyze the synteny and collinearity data generated by the modified MCScan algorithm.

All programs are executed using command line options on either MAC OS or Linux systems. Usage information is built into the programs. To show usage on the screen, users just need to run the program without giving any options:

```
$ ./program_name (for executable binary files)

$ perl program_name.pl (for perl scripts)

$ java program_name (for java classes)
```

All code is copiable, distributable, modifiable, and usable without any restrictions.

Contact: Yupeng Wang, wyp1125@uga.edu; Xu Tan, tanxu87@gmail.com

Installation

On Mac OS, Xcode (http://developer.apple.com/xcode/) should be installed prior to the installation of MCScanX package. On Linux systems, the Java SE Development Kit (JDK) and "libpng" should be installed before the installation of MCScanX package.

Then simply put MCscanX.zip into a directory and run:

```
$ unzip MCscanX.zip
$ cd MCScanX
$ make
```

The following is the list of executable programs

Core programs (in the main folder)

MCScanX

MCScanX_h

duplicate_gene_classifier

Downstream analysis programs (in the downstream_analyses folder)

Tool 1. detect_collinear_tandem_arrays

Tool 2. dissect_multiple_alignment

Tool 3. dot_plotter.java

Tool 4. dual_synteny_plotter.java

Tool 5. circle_plotter.java

Tool 6. bar_plotter.java

Tool 7. add_ka_and_ks_to_collinearity.pl

Tool 8. group_collinear_genes.pl

Tool 9. detect_collinearity_within_gene_families.pl

Tool 10. origin_enrichment_analysis.pl

Tool 11. family_circle_plotter.java

Tool 12. family_tree_plotter.java

Core programs

1. MCscanX

This program, implementing a modified MCScan algorithm, detects collinear blocks and progressively aligns multiple collinear blocks against reference chromosomes.

---Usage

MCScanX reads in two data files: xyz.blast and xyz.gff.

The xyz.blast file is simply the direct BLASTP output of m8 format as following:

```
AT1G50920   AT1G50920   100.00  671    0     0     1     671    1     671    0.0    1316
```

Here is a typical parameter setting for generating the xyz.blast file:

```
$ blastall  -i  query_file  -d database –p blastp –e 1e-10 –b 5 –v 5 –m 8 –o xyz.blast
```

The xyz.gff file holds gene positions, following a tab-delimited format:

```
 sp#      gene     starting_position          ending_position
```

Note: for sp#, sp is the two-letter short name for the species; # is the chromosome number. (For example, the second chromosome of Arabidopsis thaliana should be denoted as at2.)

The xyz.gff file can be generated by parsing the .gff3 file released by the sequencing initiatives.

Repeat of the same gene is not allowed in the .gff file.

When comparing multiple genomes, simply concatenate all inter-/intra-species m8 blast output into xyz .blast file and concatenate all gene positions of different species into xyz.gff file.

It is advised that to make MCscanX generate more reasonable results, the number of BLASTP hits for a gene should be restricted to around top 5.

When you have xyz.blast and xyz.gff ready, put them in the same folder. Then you can simply use:

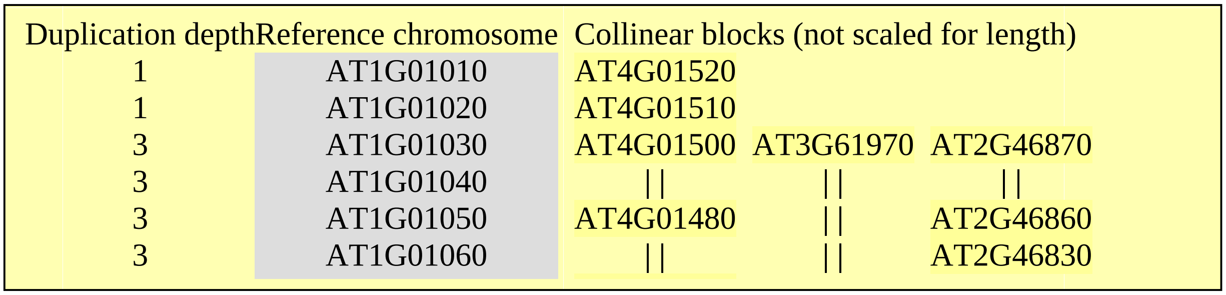
```
$ ./MCScanX  dir/xyz
```

---Output

The execution of MCScanX outputs one text file xyz.collinearity, containing pairwise collinear blocks as follows:

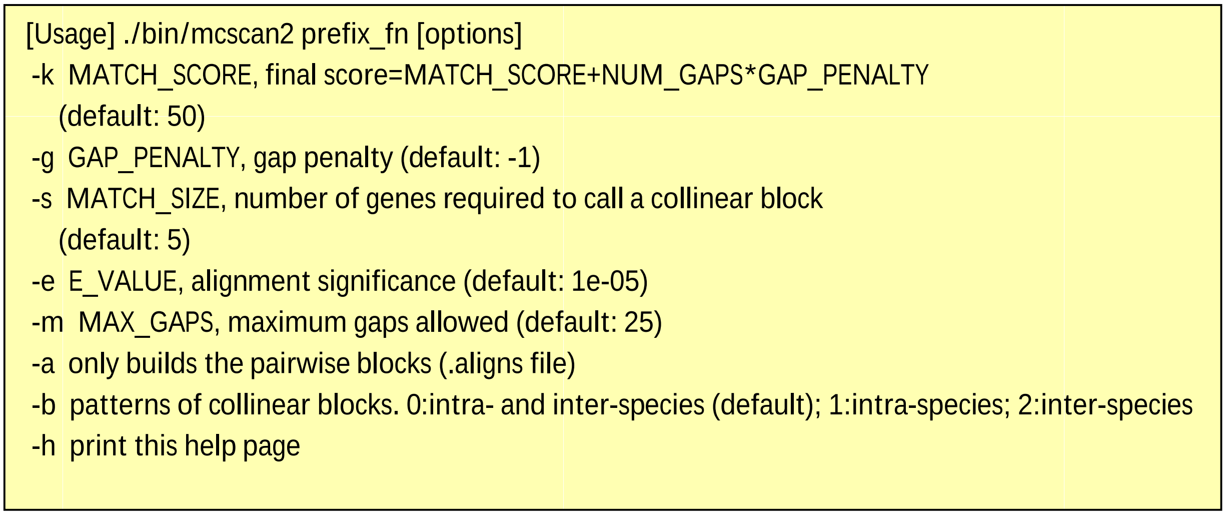
```
## Alignment 0: score=9171.0 e_value=0 N=187 at1&at1 plus
  0-  0:            AT1G17240     AT1G72300          0
  0-  1:            AT1G17290     AT1G72330          0
  ...                ...           ...               ...
  0-185:            AT1G22330     AT1G78260       1e-63
  0-186:            AT1G22340     AT1G78270      3e-174
##Alignment 1: score=5084.0 e_value=5.6e-251 N=106 at1&at1 plus
  ...
```

and one directory xyz.html , containing html files that display multiple alignment of collinear blocks against each reference chromosome. The HTML files should be viewed through a web browser. In a HTML file, the first column shows duplication depth at each gene locus, the second column shows the genes in reference chromosomes where tandem genes are marked in red, and the following is aligned collinear blocks where only match genes are displayed. For example:



Duplication depth	Reference chromosome	Collinear blocks (not scaled for length)		
1	AT1G01010	AT4G01520		
1	AT1G01020	AT4G01510		
3	AT1G01030	AT4G01500	AT3G61970	AT2G46870
3	AT1G01040	||	||	||
3	AT1G01050	AT4G01480	||	AT2G46860
3	AT1G01060	||	||	AT2G46830

---MCScanX parameters (for advanced users)



```
[Usage] ./bin/mcscan2 prefix_fn [options]
 -k  MATCH_SCORE, final score=MATCH_SCORE+NUM_GAPS*GAP_PENALTY
     (default: 50)
 -g  GAP_PENALTY, gap penalty (default: -1)
 -s  MATCH_SIZE, number of genes required to call a collinear block
     (default: 5)
 -e  E_VALUE, alignment significance (default: 1e-05)
 -m  MAX_GAPS, maximum gaps allowed (default: 25)
 -a  only builds the pairwise blocks (.aligns file)
 -b  patterns of collinear blocks. 0:intra- and inter-species (default); 1:intra-species; 2:inter-species
 -h  print this help page
```

2. MCScanX_h

The BLASTP input of MCScanX can be replaced by a tab-delimited file containing pair-wise homologous relationships detected by third party software. In this case, users should use MCScanX_h instead. The executation of MCScanX_h is very similar to that of MCScanX, except that the "xyz.blast" file should be replaced by "xyz.homology" file. At the bottom of screen output, statistics on numbers / percentages of collinear homolog pairs are shown.

The "xyz.homology" file may contain 2 or 3 tab-delimited columns. The first two columns show pair-wise homologous relationships. The optional third column shows the scores of pair-wise homologous relationships. When the third column is used, users need to specify whether higher or lower values are preferred.

As an example, users can use the combination of “orthologs.txt” and “coortholog.txt” file generated by OrthoMCL as the input (“xyz.homology”) of MCScanX_h.

```
AT1G01020   Glyma03g28190      0.343
```

3. duplicate_gene_classifier

Users may use this program, which incorporates the MCScanX algorithm, to classify origins of the duplicate genes of ONE genome into whole genome /segmental (i.e. collinear genes in collinear blocks), tandem (consecutive repeat), proximal (in nearby chromosomal region but not adjacent) or dispersed (other modes than segmental, tandem and proximal) duplications.

---Usage

```
$ ./duplicate_gene_classifier  dir/xyz
```

The input of duplicate_gene_classifier is the same with MCscanX, except an additional option for defining the maximum distance (# of genes) between 2 proximal duplicates.

---Output

The output is a text file in the same directory with input files named xyz.gene_type. It contains origin information for all the genes in xyz.gff file with a tab-delimited format:

```
Gene     gene_type(0/1/2/3/4)
```

Note: 0, 1, 2, 3, 4 stand for singleton, dispersed, proximal, tandem, WGD/segmental respectively.

It is not reasonable to apply this program to data of multiple genomes.

Downstream analyses

1. detect_collinear_tandem_arrays

Tandem duplications often complicate collinearity detection. To enhance the power of collinearity detection, MCScan algorithms use the gene with best BLASTP hit to represent a tandem array. This program transforms match genes in collinear blocks into tandem arrays if tandem duplications exist there.

---Usage

```
$ ./detect_collinear_tandem_arrays -g gff_file -b blast_file -c collinearity_file -o output_file
```

---Output

The path of output_file should be specified by the user. If any gene of a collinear pair is located in a tandem array, the collinear pair will be written into the output_file.

2. dissect_multiple_alignment

This program dissects the number of collinear blocks at each gene locus of the reference chromosomes into the number of intra-species collinear blocks and the number of inter-species collinear blocks.

---Usage

```
$ ./dissect_multiple_alignment -g gff_file -c collinearity_file -o output_file
```

---Output

The path of output_file should be specified by the user. The first and second columns of output_file show the chromosomes and genes in reference chromosomes. The 3rd, 4th and 5th columns show the numbers of intra-species collinear blocks, inter-species collinear blocks and outgroup species respectively.

3. dot_plotter.java

This java script generates a dot plot for all the collinear blocks on two sets of chromosomes given by the user. Note that JDK is needed for executing Java programs.

---Usage

```
$ java dot_plotter -g gff_file -s collinearity_file -c control_file -o output_PNG_file
```

The input files include a *gff* file containing all gene positions, a collinearity file generated by MCScanX, and a control file (.ctl) containing plot size and chromosome IDs.

The control file can be easily made by modifying the dot.ctl file:

```
800     //dimension (in pixels) of x axis
800     //dimension (in pixels) of y axis
sb1,sb2,sb3,sb4,sb5,sb6,sb7,sb8,sb9,sb10        //chromosomes in x axis
os1,os2,os3,os4,os5,os6,os7,os8,os9,os10,os11,os12       //chromosomes in y axis
```

Note that no space is allowed between adjacent chromosome IDs.

---Output

Output is an image file (PNG format) which can be viewed with an image viewer.

Each dot is a collinear gene pair between the two sets of chromosomes. Different colors of dots, generated randomly, represent different collinear blocks.

4. dual_synteny_plotter.java

This java script generates a dual synteny plot which links all the collinear blocks between two sets of chromosomes using straight lines.

---Usage

```
$ java dual_synteny_plotter -g gff_file -s collinearity_file -c control_file -o output_PNG_file
```

The input files include a *gff* file containing all gene positions, a collinearity file generated by MCScanX, and a control file (.ctl) containing plot size and chromosome IDs.

The control file can be easily made by modifying the column.ctl file:

```
200     //plot width (in pixels)
800     //plot height (in pixels)
sb1,sb2 //chromosomes in the left column
os1,os2,os3     //chromosomes in the right column
```

Note that no space is allowed between adjacent chromosome IDs.

---Output

Output is an image file (PNG format) which can be viewed with an image viewer.

Each line links a pair of collinear genes between the two sets of chromosomes. Different colors of lines, generated randomly, represent different collinear blocks.

5. circle_plotter.java

This Java scripts generates a circular plot which links all the collinear blocks with curved lines between and within the chromosomes given by users.

---Usage

```
$ java circle_plotter -g gff_file -s collinearity_file -c control_file -o output_PNG_file
```

The input files include a *gff* file containing all gene positions, a collinearity file generated by MCScanX, and a control file (.ctl) containing plot size and chromosome IDs.

The control file can be easily made by modifying the *circle.ctl* file:

```
800     //plot width and height (in pixels)
sb1,sb2,os1,os2,os3     //chromosomes in the circle
```

Note that no space is allowed between adjacent chromosome IDs.

---Output

Output is an image file (PNG format) which can be viewed with an image viewer.

Each curved line links a pair of collinear genes between or within the given chromosomes. Different colors of lines, generated randomly, represent different collinear blocks.

6. bar_plotter.java

This Java scripts generates a bar plot displaying chromosome rearrangement between reference and target chromosome sets given by users.

--Usage:

```
java bar_plotter -g gff_file -s collinearity_file -c control_file -o output_PNG_file
```

The input files include a gff file containing all gene positions, a collinearity file generated by MCScanX, and a control file (.ctl) containing plot size and chromosome IDs. The control file can be easily made by modifying the bar.ctl file:

```
800     //dimension (in pixels) of x axis
800     //dimension (in pixels) of y axis
sb1,sb2,sb3,sb4,sb5,sb6,sb7,sb8,sb9,sb10       //reference chromosomes
os1,os2,os3,os4,os5,os6,os7,os8,os9,os10,os11,os12      //target chromosomes
```

Note that no space is allowed between adjacent chromosome IDs.

7. add_ka_and_ks_to_collinearity.pl

This program calculates the Ka & Ks value of each collinear gene pair shown in the MCScanX output (.collinearity file). Clustalw and Bio-perl are needed for executing this program.

---Usage

The input is a xyz.syteny file generated by MCScanX and a coding sequence file of corresponding gene set in fasta format.

```
$ perl add_ka_and_ks_to_synteny.pl -i ../data/xyz.collinearity -d ../data/xyz.cds -o
xyz.collinearity.kaks
```

---Output

Users should specify the path of output file. The output file is a modified version of xyz.collinearity file with each line containing a collinear gene pair and its ka & ks values.

8. group_collinear_genes.pl

This program groups genes through connecting collinear genes until any gene in each group has no collinear gene outside the group. This analysis can be used to construct gene families based on collinear relationships.

---Usage

```
$ perl group_collinear_genes.pl -i collinearity_file -o output_file
```

Input includes a *xyz.collinearity* file generated by MCScanX.

---Output

The output file displays each group in one line in a tab-delimited format.

Note, the first group (the largest size) usually contains much more genes than other groups, should be regarded as non-informative.

9. detect_collinearity_within_gene_families.pl

This program detects collinear gene pairs within gene families.

---Usage

```
$ perl detect_collinearity_within_gene_families.pl -i gene_family_file -d xyz.collinearity  -o
output_file
```

Input includes a *xyz.collinearity* file generated by MCScanX and a gene family file in tab-delimited format with gene family name in the first column:

```
Gene_family_1  gene1          gene2          gene3...          genex

Gene_family_2  gene1          gene2          gene3...          genex
```

---Output

The output file gives the collinear pairs of the given gene families in tab-delimited format:

```
Gene_family_1  gene_pair1      gene_pair2    ...  gene_pairx

Gene_family_2  gene_pair1      gene_pair2
```

10. origin_enrichment_analysis.pl

This program identifies potential enrichment of duplicate gene origins for input gene families according to the result of *duplicate_gene_classifier.*

---Usage

```
$ perl origin_enrichment_analysis.pl -i gene_family_file -d gene_origin_file  -o output_file
```

This perl program takes in a gene family file with the same format as the above ones and the gene origin file generated by *duplicate_gene_classifier.*

---Output

The output is the p-values of different origins for the given gene families

11. family_circle_plotter.java

This java script generates a circular plot which links all collinear genes within a gene family with red curved lines, and places the gene family collinearity into a genomic collinearity background.

---Usage

```
$ java family_circle_plotter -g gff_file -s collinearity_file -c control_file -f gene_family_file -o
output_PNG_file
```

The input files include a *.gff* file containing all gene positions, a *.collinearity* file generated by MCScanX, a control file (*.ctl*) containing the plot size and chromosome IDs and a gene family file containing only one gene family with the aforementioned format.

The control file can be easily made by modifying the *family.ctl* file:

```
800     //plot width and height (in pixels)
at1, at2, at3, at4, at5     //chromosomes in the circle
```

Note: users can input just the chromosomes of interest into the *family.ctl* file. This will generate a circular plot within the given chromosomes set.

---Output

Output is an image file which can be viewed with any image. Each red curved line links a pair of collinar genes within the given gene family. The grey lines stand for genomic collinearity background.

12. family_tree_plotter.java

This java script generates a gene family tree on which collinear gene pairs and tandem gene groups are linked with red and blue curves respectively.

--Usage:

```
$ java family_tree_plotter -t tree_file -s collinearity_file -o output_PNG_file (show collinear gene pairs only)
$ java family_tree_plotter -t tree_file -s collinearity_file –d tandem_pair_file -o output_PNG_file (show both collinear and tandem gene pairs)
```

The input files include a .collinearity file generated by MCScanX and a tree file for the gene family in newick format (bracket tree).

Users can set up the plot width, plot height, and font_size with the following options: -x plot_width -y plot height -f font_size

--Output

The output is an image file (PNG format) which can be viewed with an image viewer;

Note: this script aims to show the collinear and tandem overview for a gene family. The branch lengths are disregarded, thus do not reflect the true value.