Zhang-Lab 生信小课堂 第十期

Applied Bioinformatics Club（ABC）

和趣求真 秉实生信

（张建伟生物信息学课题组 https://zhang.hzau.edu.cn）

同源基因鉴定

拥有物种蛋白质序列，如何鉴定物种间的同源基因？

2023.3.3 二综一楼C102 15:00 欢迎大家交流学习！

主讲人：李姗莹

2023/03/03

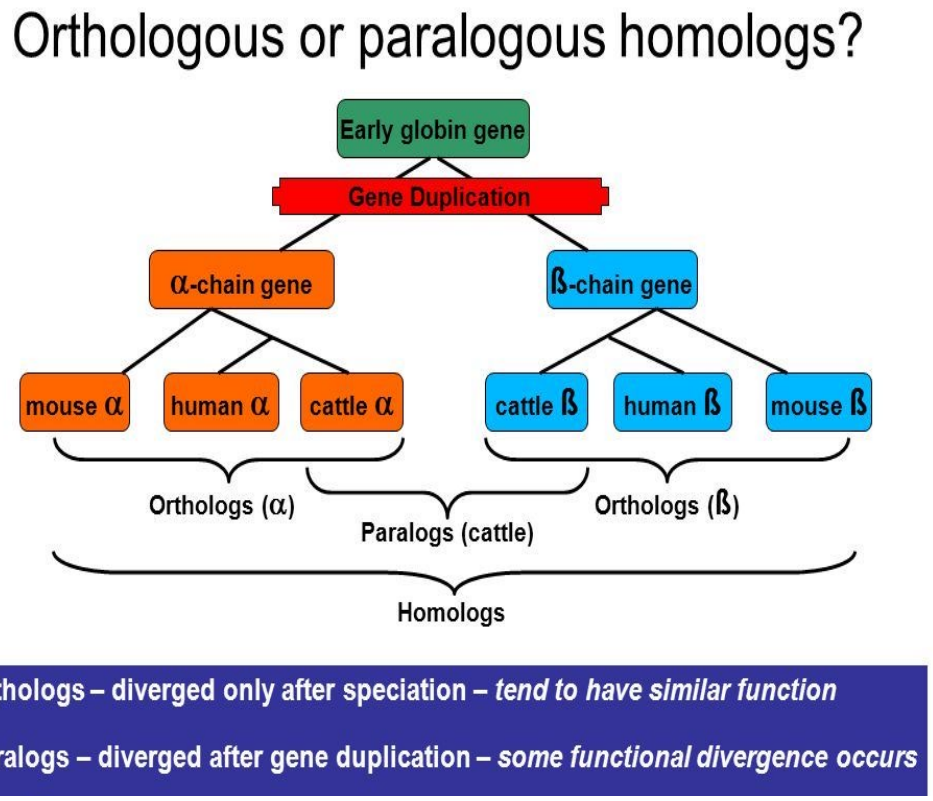
同源基因

Homologs(同源基因): 由一个共同祖先在不同物种中遗传的基因。同源基因在序列上是相似的，但相似的序列不一定是同源的。

- Orthologs(直系同源基因): 来自于不同物种的，从同一祖先垂直进化而来的基因，保留了与原始基因有相同的功能。直系同源基因通常是编码生命必需的酶、辅酶或关键性的调控蛋白的基因，功能保守，进化缓慢。大多数直系同源基因功能相同或相近，调控途径也相似，**常用来构建系统发育树**。

- Paralogs(旁系同源基因): 由于基因复制而产生的同源基因，可能会进化出与原来基因相似的功能但是也可以进化成不同的特征，旁系同源基因并不局限于同一物种内，不同物种中由于始祖基因的复制而分化的基因也称旁系同源基因。

- xenologs(异同源基因): 通过水平基因转移，来源于共生或病毒侵染所产生的相似基因。异同源的产生不是垂直进化而来的，也不是平行复制产生的，而是由于原核生物与真核生物的接触，比如病毒感染，在跨度巨大的物种间跳跃转移产生的。



同源基因鉴定软件

- Genetribe
- Orthofinder
- Inparanoid
- MCScanX

Genetribe

Genetribe安装

需要依赖python3和三个工具使用
#BLAST (v2.9.0)
conda install blast -c bioconda
#MCscan (v1.0.6)
pip install jcvi
#BEDTools (v2.29.2)
conda install bedtools -c bioconda

#Genetribe安装
git clone https://github.com/chenym1/genetribe.git
cd genetribe
./install.sh
vi ~/.bash_profile
add the following lines to the end of ~/.bash_profile
export PATH=/path/to/genetribe/:$PATH
source ~/.bash_profile
genetribe -h

GeneTribe

A tool for performing collinearity-incorporating homology inference

官网：https://chenym1.github.io/genetribe/

Yongming Chen, Wanjun Song, Xiaoming Xie, Zihao Wang, Panfeng Guan, Huiru Peng, Yuannian Jiao, Zhongfu Ni, Qixin Sun, and Weilong Guo. (2020) A Collinearity-incorporating Homology Inference Strategy for Connecting Emerging Assemblies in Triticeae Tribe as a Pilot Practice in the Plant Pangenomic Era. Molecular Plant, 13, 1694–1708.

Genetribe

Genetribe输入文件

File1 Protein Sequences in Fasta Format (name.fa)

```
>AT5G16970.1 pep chromosome:TAIR10:5:5575973:5578086:-1 gene:AT5G16970 transcript:AT5G16970.1 gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:AER descri
ption:alkenal reductase [Source:NCBI gene (formerly Entrezgene);Acc:831560]
MTATNKQVILKDYVSGFPTESDFDFTTTTVELRVPEGTNSVLVKNLYLSCDPYMRIRMGK
PDPSTAALAQAYTPGQPIQGYGVSRIIESGHPDYKKGDLLWGIVAWEEYSVITPMTHAHF
KIQHTDVPLSYYTGLLGMPGMTAYAGFYEVCSPKEGETVYVSAASGAVGQLVGQLAKMMG
CYVVGSAGSKEKVDLLKTKFGFDDAFNYKEESDLTAALKRCFPNGIDIYFENVGGKMLDA
VLVNMNMHGRIAVCGMISQYNLENQEGVHNLSNIIYKRIRIQGFVVSDFYDKYSKFLEFV
LPHIREGKITYVEDVADGLEKAPEALVGLFHGKNVGKQVVVVARE
>AT4G32100.1 pep chromosome:TAIR10:4:15511757:15512218:-1 gene:AT4G32100 transcript:AT4G32100.1 gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:AT4G3210
0 description:Beta-1,3-N-Acetylglucosaminyltransferase family protein [Source:NCBI gene (formerly Entrezgene);Acc:829341]
MATNACKFLCLVLLFAFVTQGYGDDSYSLESLSVIQSKTGNMVENKPEWEVKVLNSSPCY
FTHTTLSCVRFKSVTPIDSKVLSKSGDTCLLGNGDSIHDISFKYVWDTSFDLKVVDGYIA
CS
```

File2 Annotation File in Bed Format (name.bed)

```
1       3630    5899    AT1G01010       0       +
1       6787    9130    AT1G01020       0       -
1       11648   13714   AT1G01030       0       -
1       23120   31227   AT1G01040       0       +
1       31169   33171   AT1G01050       0       -
1       33364   37871   AT1G01060       0       -
1       38443   41017   AT1G01070       0       -
1       44969   47059   AT1G01080       0       -
1       47233   49304   AT1G01090       0       -
1       49908   51210   AT1G01100       0       -
1       51952   54737   AT1G01110       0       +
1       57163   59215   AT1G01120       0       -
1       61904   63811   AT1G01130       0       -
1       64165   67774   AT1G01140       0       -
1       69910   72138   AT1G01150       0       -
1       72338   74096   AT1G01160       0       +
```

File3 Chromosome Group Information (name.chrlist)

```
N
Arabidopsis_thaliana.chrlist (END)
```

Genetribe

Genetribe使用

```
#BSUB -J genetribe
#BSUB -n 10
#BSUB -o genetribe.%J.out
#BSUB -e genetribe.%J.err
#BSUB -R span[hosts=1]
#BSUB -q smp
```

core

Usage: genetribe core -l <FirstName> -f <SecondName> [options]

genetribe core -l Physcomitrium_patens -f Arabidopsis_thaliana

longestcds

Usage: genetribe longestfasta -i pep.fa -s strsplit

genetribe longestfasta -i Physcomitrium_patens.fa -s strsplit

Genetribe

Genetribe结果

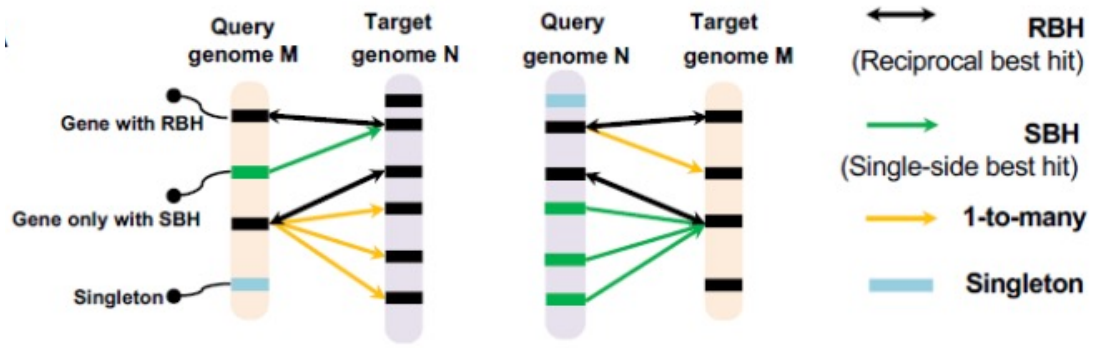
```
Arabidopsis_thaliana.bed                              Arabidopsis_thaliana_Physcomitrium_patens.one2one    Physcomitrium_patens_Arabidopsis_thaliana.SBH
Arabidopsis_thaliana.cds                              Arabidopsis_thaliana_Physcomitrium_patens.SBH        Physcomitrium_patens_Arabidopsis_thaliana.singleton
Arabidopsis_thaliana.chrlist                          Arabidopsis_thaliana_Physcomitrium_patens.singleton  Physcomitrium_patens.bed
Arabidopsis_thaliana.fa                               Arabidopsis_thaliana.TAIR10.54.gff3                  Physcomitrium_patens.cds
Arabidopsis_thaliana.longestcds.fa                    Physcomitrium_patens_Arabidopsis_thaliana.block_pos  Physcomitrium_patens.chrlist
Arabidopsis_thaliana_Physcomitrium_patens.block_pos   Physcomitrium_patens_Arabidopsis_thaliana.one2many   Physcomitrium_patens.fa
Arabidopsis_thaliana_Physcomitrium_patens.csv         Physcomitrium_patens_Arabidopsis_thaliana.one2one    Physcomitrium_patens.longestcds.fa
Arabidopsis_thaliana_Physcomitrium_patens.one2many    Physcomitrium_patens_Arabidopsis_thaliana.RBH        Physcomitrium_patens.Phypa_V3.54.gff3
```

AT1G10630	Pp3c12_14910	0.97
AT1G69550	Pp3c5_6380	0.80
AT3G22930	Pp3c14_8590	0.80
AT5G03240	Pp3c18_2470	0.98
AT1G07920	Pp3c1_23850	0.94
AT3G55590	Pp3c8_17380	0.78
AT3G11940	Pp3c23_6150	0.88
AT5G60670	Pp3c27_660	0.89
AT2G34420	Pp3c10_3020	0.76
AT1G55060	Pp3c5_1920	0.98
AT3G09790	Pp3c19_8830	0.98
AT1G49300	Pp3c2_24890	0.80
AT2G47170	Pp3c4_20580	0.95
AT2G29570	Pp3c12_18160	0.84
AT1G61580	Pp3c10_25460	0.82

one2many

AT1G01050	Pp3c13_16700	RBH	N
AT1G01230	Pp3c11_20050	RBH	N
AT1G01620	Pp3c13_18810	RBH	N
AT1G01940	Pp3c14_5290	RBH	N
AT1G02140	Pp3c1_16440	RBH	N
AT1G02500	Pp3c19_3060	RBH	N
AT1G03150	Pp3c10_6780	RBH	N
AT1G03190	Pp3c10_16570	RBH	N
AT1G03330	Pp3c2_19710	RBH	N
AT1G03950	Pp3c13_6090	RBH	N
AT1G04170	Pp3c7_5020	RBH	N
AT1G04270	Pp3c12_9180	RBH	N
AT1G04300	Pp3c12_920	RBH	N

one2one (RBH+SBH)



Orthofinder

Orthofinder安装

```
#1.canda安装
conda install -c bioconda orthofinder
orthofinder -h

#2.直接调用集群
module load OrthoFinder/2.3.8
```

Emms, D. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16: 157
Emms, D. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20: 238

Orthofinder

Orthofinder使用

```
#BSUB -J orthofinder
#BSUB -n 10
#BSUB -o orthofinder.%J.out
#BSUB -e orthofinder.%J.err
#BSUB -R span[hosts=1]
#BSUB -q smp

orthofinder -f orthofinder/ -t 2
```

```
$ orthofinder -f data \ # 数据目录
    -S diamond \ # 比对blast, mmseqs, blast_gz, diamond（推荐）
    -M msa \ # 基因树推断方法, dendroblast, msa（推荐）
    -T fastatree \ # 建树软件, iqtree, raxml-ng, fasttree（推荐）, raxml
    -t 6 \ # 线程数, 根据服务器配置选择
```

Orthofinder

Orthofinder结果

```
Citation.txt                        Log.txt          Phylogenetically_Misplaced_Genes       Single_Copy_Orthologue_Sequences
Comparative_Genomics_Statistics     Orthogroups      Phylogenetic_Hierarchical_Orthogroups  Species_Tree
Gene_Duplication_Events             Orthogroup_Sequences  Putative_Xenologs                 WorkingDirectory
Gene_Trees                          Orthologues      Resolved_Gene_Trees
```

```
Orthogroups.GeneCount.tsv
Orthogroups_SingleCopyOrthologues.txt
Orthogroups.tsv
Orthogroups.txt
Orthogroups_UnassignedGenes.tsv
```

```
SpeciesTree_rooted_node_labels.txt
SpeciesTree_rooted.txt
```

Inparanoid

Inparanoid安装

```
#blastall
conda install -c bioconda blast-legacy=2.2.26

#Inparanoid
wget https://bitbucket.org/sonnhammergroup/inparanoid4/get/359f8ea484ba.zip
unzip 359f8ea484ba.zip
perl inparanoid
```

```
blast_parser.pl  BLOSUM62  EC            LICENSE  PAM70       SC
BLOSUM45         BLOSUM80  inparanoid.pl  PAM30    README.txt  seqstat.jar
```

InParanoid 7: new algorithms and tools for eukaryotic orthology analysis" Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O and Sonnhammer ELL Nucleic Acids Res. 38:D196-D203 (2009)

Inparanoid

Inparanoid使用

第一步：建库
formatdb -i Arabidopsis_thaliana.fa

第二步：运行Inparanoid
Usage:perl inparanoid.pl <FASTAFILE with sequences of species A> <FASTAFILE with sequences of species B> [FASTAFILE with sequences of species C]
perl inparanoid.pl Arabidopsis_thaliana.pep.fa Arabidopsis_thaliana.pep.fa

Inparanoid

Inparanoid结果文件

1	5042	Arabidopsis_thaliana.longest.fa	1.000	AT2G17930
1	5042	Arabidopsis_thaliana.longest.fa	0.677	AT4G36080
1	5042	Physcomitrium_patens.longest.fa	1.000	Pp3c17_20000
2	4552	Arabidopsis_thaliana.longest.fa	1.000	AT1G80070
2	4552	Physcomitrium_patens.longest.fa	1.000	Pp3c8_25090
2	4552	Physcomitrium_patens.longest.fa	0.859	Pp3c24_20910
3	4420	Arabidopsis_thaliana.longest.fa	1.000	AT3G02260
3	4420	Physcomitrium_patens.longest.fa	1.000	Pp3c11_22340
3	4420	Physcomitrium_patens.longest.fa	0.646	Pp3c7_7030
4	3973	Arabidopsis_thaliana.longest.fa	1.000	AT5G23110
4	3973	Physcomitrium_patens.longest.fa	1.000	Pp3c9_15860
5	3381	Arabidopsis_thaliana.longest.fa	1.000	AT1G48090
5	3381	Physcomitrium_patens.longest.fa	1.000	Pp3c9_20320
6	3289	Arabidopsis_thaliana.longest.fa	1.000	AT1G03060
6	3289	Arabidopsis_thaliana.longest.fa	0.608	AT4G02660
6	3289	Physcomitrium_patens.longest.fa	1.000	Pp3c6_26100
6	3289	Physcomitrium_patens.longest.fa	0.700	Pp3c5_2897
6	3289	Physcomitrium_patens.longest.fa	0.700	Pp3c5_2890
6	3289	Physcomitrium_patens.longest.fa	0.096	Pp3c16_11700
6	3289	Physcomitrium_patens.longest.fa	0.096	Pp3c16_11705
7	3270	Arabidopsis_thaliana.longest.fa	1.000	AT5G53460
7	3270	Physcomitrium_patens.longest.fa	1.000	Pp3c5_20080
7	3270	Physcomitrium_patens.longest.fa	0.364	Pp3c16_19110
8	3263	Arabidopsis_thaliana.longest.fa	1.000	AT1G20960
8	3263	Arabidopsis_thaliana.longest.fa	0.314	AT2G42270
8	3263	Physcomitrium_patens.longest.fa	1.000	Pp3c1_35880
9	3252	Arabidopsis_thaliana.longest.fa	1.000	AT1G55860
9	3252	Arabidopsis_thaliana.longest.fa	0.851	AT1G70320
9	3252	Physcomitrium_patens.longest.fa	1.000	Pp3c11_25020
9	3252	Physcomitrium_patens.longest.fa	0.705	Pp3c7_8070
9	3252	Physcomitrium_patens.longest.fa	0.700	Pp3c7_8076
10	3078	Arabidopsis_thaliana.longest.fa	1.000	AT2G26890
10	3078	Physcomitrium_patens.longest.fa	1.000	Pp3c13_3460
10	3078	Physcomitrium_patens.longest.fa	0.601	Pp3c4_24650

SQLtable

```
27628 sequences in file Arabidopsis_thaliana.longest.fa
32234 sequences in file Physcomitrium_patens.longest.fa
18952 sequences Arabidopsis_thaliana.longest.fa have homologs in dataset Physcomitrium_patens.longest.fa
17039 sequences Physcomitrium_patens.longest.fa have homologs in dataset Arabidopsis_thaliana.longest.fa
83594 Arabidopsis_thaliana.longest.fa-Arabidopsis_thaliana.longest.fa matches
64463 Physcomitrium_patens.longest.fa-Physcomitrium_patens.longest.fa matches
################################
7092 groups of orthologs
11860 in-paralogs from Arabidopsis_thaliana.longest.fa
13766 in-paralogs from Physcomitrium_patens.longest.fa
Grey zone 0 bits
Score cutoff 40 bits
In-paralogs with confidence less than 0.05 not shown
Sequence overlap cutoff 0.5
Group merging cutoff 0.5
Scoring matrix BLOSUM62
################################
--------------------------------------------------------------------------------
Group of orthologs #1. Best score 5042 bits
Score difference with first non-orthologous sequence - Arabidopsis_thaliana.longest.fa:5042   Physcomitrium_patens.longest.fa:5042
AT2G17930               100.00%         Pp3c17_20000            100.00%
AT4G36080               67.69%
--------------------------------------------------------------------------------
Group of orthologs #2. Best score 4552 bits
Score difference with first non-orthologous sequence - Arabidopsis_thaliana.longest.fa:228   Physcomitrium_patens.longest.fa:4552
AT1G80070               100.00%         Pp3c8_25090             100.00%
                                        Pp3c24_20910            85.87%
--------------------------------------------------------------------------------
Group of orthologs #3. Best score 4420 bits
Score difference with first non-orthologous sequence - Arabidopsis_thaliana.longest.fa:4420   Physcomitrium_patens.longest.fa:4420
AT3G02260               100.00%         Pp3c11_22340            100.00%
                                        Pp3c7_7030              64.65%
--------------------------------------------------------------------------------
Group of orthologs #4. Best score 3973 bits
Score difference with first non-orthologous sequence - Arabidopsis_thaliana.longest.fa:3973   Physcomitrium_patens.longest.fa:3973
AT5G23110               100.00%         Pp3c9_15860             100.00%
--------------------------------------------------------------------------------
Group of orthologs #5. Best score 3381 bits
Score difference with first non-orthologous sequence - Arabidopsis_thaliana.longest.fa:3381   Physcomitrium_patens.longest.fa:3381
AT1G48090               100.00%         Pp3c9_20320             100.00%
```

Stats

MCScanX

MCScanX安装

官网：http://chibba.pgml.uga.edu/mcscan2/#tm
github：https://github.com/wyp1125/MCScanx

```
#blast
conda install blast2.2.31

#MCScanX
wget https://codeload.github.com/wyp1125/MCScanX/zip/refs/heads/master
unzip master
cd MCScanX-master # master解压出来文件名是MCScanX-mastercd MCScanX
Make
```

Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res, 40(7): e49.

MCScanX

MCScanX使用

第一步：blastp（protein-protein BLAST）比对

#合并

cat at.fa vv.fa >> all.fasta

#建库

makeblastdb -in all.fa -dbtype prot -parse_seqids -out all

#蛋白比对

blastp -query all.fa -db all -out at_vv.blast -evalue 1e-10 -num_threads 16 -outfmt 6 -num_alignments 5

第二步：运行MCScanX

input files：at_vv.gff; at_vv.blast

MCScanX at_vv/

AT1G50920	AT1G50920	100.00	671	0	0	1	671	1	671	0.0	1316
AT1G50920	AT1G10300	75.33	673	149	3	1	671	30	687	0.0	985
AT1G50920	AT1G80770	27.24	290	204	3	62	349	147	431	2e-27	120
AT1G50920	AT3G23860	35.68	185	102	6	6	179	6	184	9e-18	88.6
AT1G50920	AT1G78010	25.11	219	136	8	171	382	320	517	1e-06	52.0
AT1G36960	AT1G36960	100.00	181	0	0	1	181	1	181	9e-72	265
AT1G36960	AT5G36228	53.52	71	33	0	8	78	108	178	2e-17	84.7
AT1G36960	AT5G25600	60.61	66	25	1	11	76	111	175	8e-17	82.8
AT1G36960	AT1G48095	77.53	89	20	0	89	177	117	205	2e-14	75.5
AT1G36960	AT1G48095	80.65	31	6	0	46	76	18	48	1e-08	55.8
AT1G36960	AT2G02103	53.97	63	29	0	16	78	116	178	2e-13	71.6
AT1G36960	AT5G18636	53.97	63	29	0	16	78	116	178	3e-13	71.2
AT1G44020	AT1G44020	100.00	577	0	0	1	577	1	577	0.0	1134

at_vv.blast

at3	AT3G19630	6818676	6820674
at5	AT5G11220	3577057	3577854
at2	AT2G29110	12506880	12510552
at3	AT3G19790	6874937	6875440
at3	AT3G60770	22460525	22461656
at3	AT3G21570	7600297	7600707
at5	AT5G07420	2349547	2351366
at4	AT4G12030	7210981	7212713
at2	AT2G15010	6484422	6485001

at_vv.gff

MCScanX

MCScanX结果

```
[Usage] MCScanX 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)
 -w  OVERLAP_WINDOW, maximum distance (# of genes) to collapse BLAST matches (default: 5)
 -a  only builds the pairwise blocks (.collinearity file)
 -b  patterns of collinear blocks. 0:intra- and inter-species (default); 1:intra-species; 2:inter-species
 -h  print this help page
```

结果文件：at_vv.html, at_vv.collinearity, at_vv.tandem

```
############### Parameters ###############
# MATCH_SCORE: 50
# MATCH_SIZE: 5
# GAP_PENALTY: -1
# OVERLAP_WINDOW: 5
# E_VALUE: 1e-05
# MAX GAPS: 25
############### Statistics ###############
# Number of collinear genes: 24437, Percentage: 47.49
# Number of all genes: 51452
##########################################
## Alignment 0: score=8972.0 e_value=0 N=190 at1&at1 plus
  0-  0:        AT1G17240       AT1G72300             0
  0-  1:        AT1G17290       AT1G72330             0
  0-  2:        AT1G17310       AT1G72350         5e-41
  0-  3:        AT1G17350       AT1G72420        2e-113
  0-  4:        AT1G17380       AT1G72450         7e-63
  0-  5:        AT1G17400       AT1G72490         2e-82
  0-  6:        AT1G17420       AT1G72520             0
  0-  7:        AT1G17430       AT1G72620        1e-143
  0-  8:        AT1G17455       AT1G72630         1e-53
```

at_vv.collinearity里记录了共线性信息

MCScanX

MCScanX结果可视化

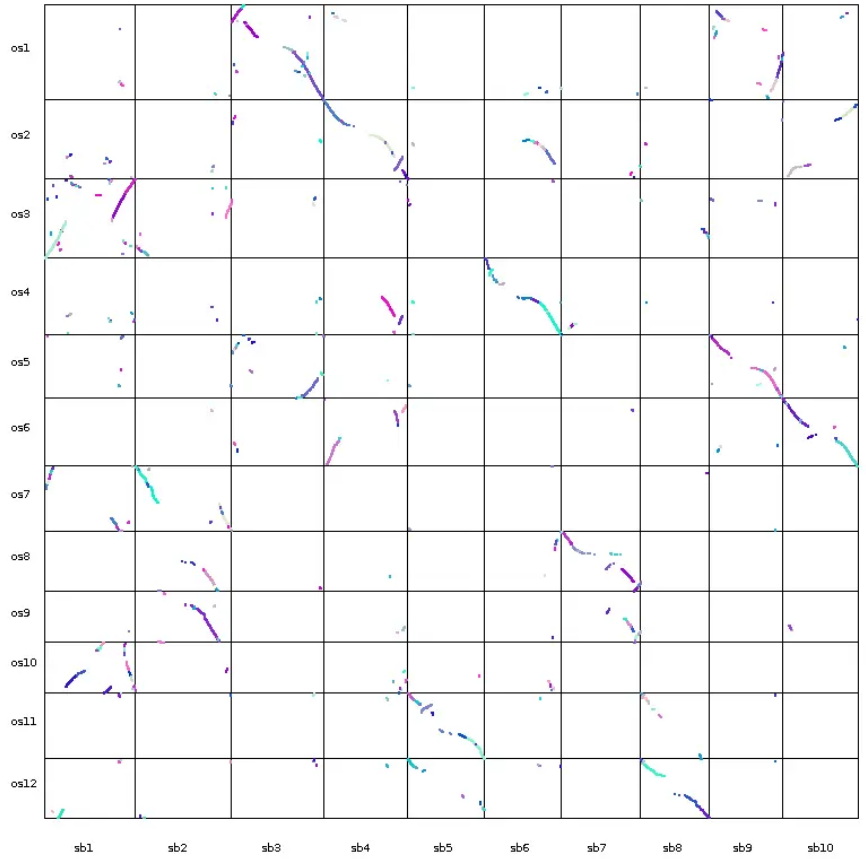
Input files： at_vv.gff; at_vv.collinearity; dot.ctl

dot.ctl

```
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
```

```
#BSUB -J dot_plotter
#BSUB -n 10
#BSUB -o dot_plotter.%J.out
#BSUB -e dot_plotter.%J.err
#BSUB -R span[hosts=1]
#BSUB -q normal
```

java dot_plotter -g at_vv.gff -s at_vv.collinearity -c dot.ctl -o dot.PNG



MCScanX

MCScanX结果可视化

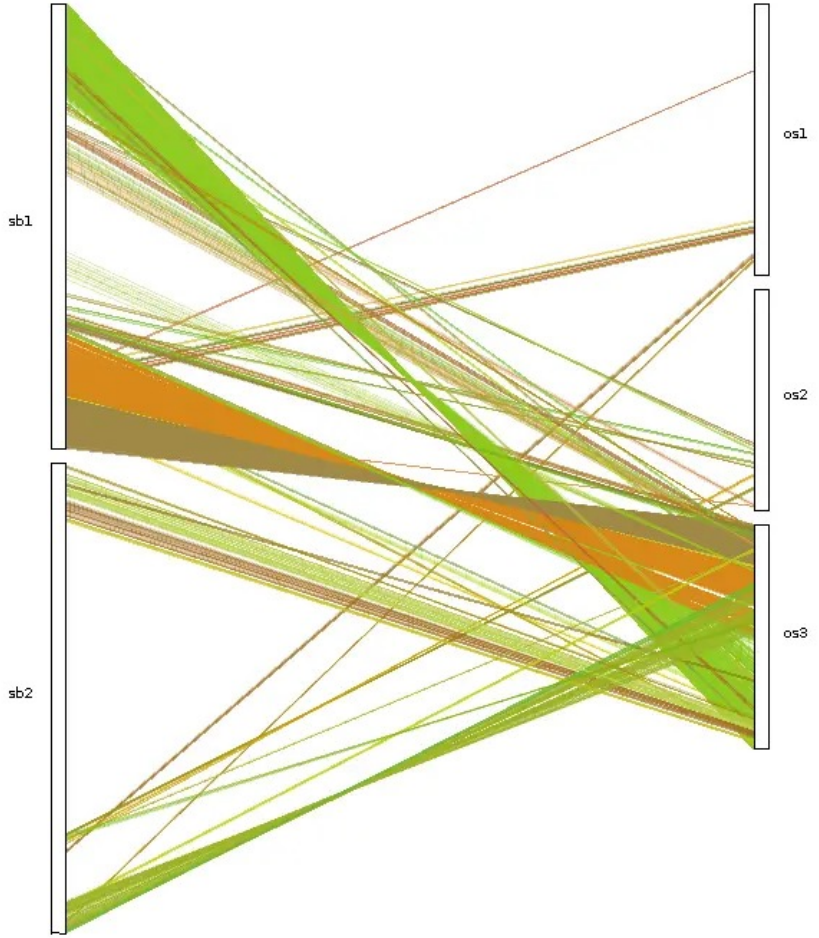
input files：at_vv.gff; at_vv.collinearity; dual_synteny.ctl

dual_synteny.ctl

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

```
#BSUB -J dual_synteny_plotter
#BSUB -n 10
#BSUB -o dual_synteny_plotter.%J.out
#BSUB -e dual_synteny_plotter.%J.err
#BSUB -R span[hosts=1]
#BSUB -q normal
```

java dual_synteny_plotter -g at_pp.gff -s at_pp.collinearity -c dual_synteny.ctl -o dual_synteny.PNG



MCScanX

MCScanX结果可视化

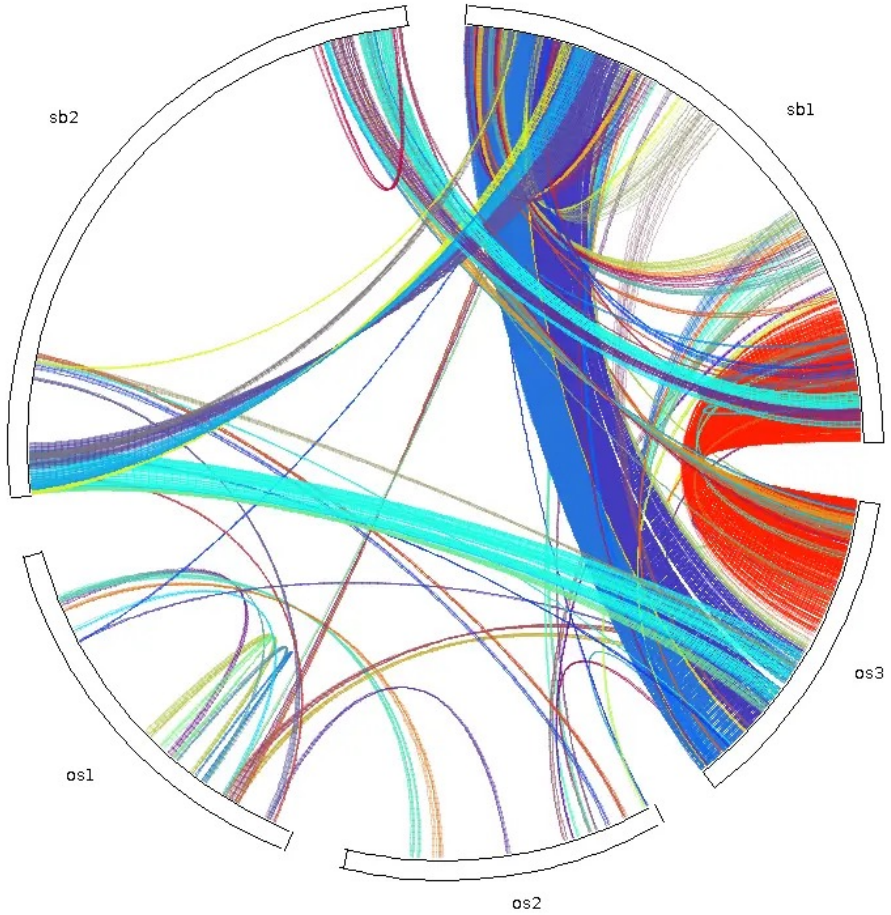
input files：at_vv.gff; at_vv.collinearity; circle.ctl

circle.ctl

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

```
#BSUB -J circle_plotter
#BSUB -n 10
#BSUB -o circle_plotter.%J.out
#BSUB -e circle_plotter.%J.err
#BSUB -R span[hosts=1]
#BSUB -q normal
```

java circle_plotter -g at_vv.gff -s at_vv.collinearity -c circle.ctl -o circle.PNG



感谢大家学习交流