Zhang-Lab 生信小课堂 第八期

和趣求真 Z 秉实生信

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

基因家族分析

2022.12.09 二综一楼C102 15:00 欢迎大家交流学习！

主讲人：李梦圆
2022/12/09

基因家族分析

主讲人：李梦圆
2022/12/09

Zhou L, Yarra R. Genome-Wide Identification and Characterization of AP2/ERF Transcription Factor Family Genes in Oil Palm under Abiotic Stress Conditions. Int J Mol Sci. 2021 Mar 10;22(6):2821. doi: 10.3390/ijms22062821. PMID: 33802225; PMCID: PMC8000548.

基因家族鉴定

hmm

隐马尔可夫模型HMM
PF00847
https://www.ebi.ac.uk/interpro/search/text/

#目标基因家族搜索

```
hmmsearch --cut_tc –domtblout AP2.out PF00847.hmm Oil_Palm.pep.all.fa.gz
```

#过滤筛选得到E-value小于1*10-20,先拿到序列号

```
grep -v  “#”  AP2.out|awk '($7 + 0) < 1E-20'|cut -f1 -d " "|sort -u > AP2_first_id.txt
```

#再根据序列号，从Oil_Palm.pep.all.fa.gz中提取序列

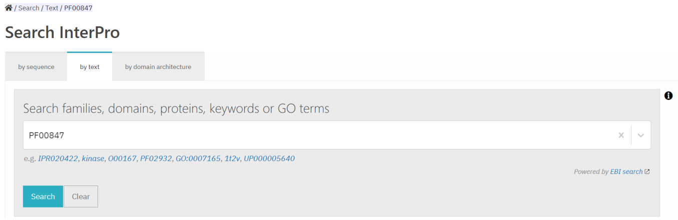
```
less Oil_Palm.pep.all.fa.gz | seqkit grep -f AP2_first_id.txt > AP2_first.fa
```

#对筛选出来的序列用clustalw进行多序列比对

```
clustalw2  -infile=AP2_first.fa -type=PROTEIN -output=fasta -outfile=AP2_align.out -outorder=input
```

#使用hmmbuild对这些置信的序列进行隐马尔可夫模型的构建，即构建更加准确的hmm模型来尽可能的预测目标物种中AP2基因家族中所有的成员。

```
hmmbuild AP2_first.hmm  AP2_align.out
hmmsearch --cut_tc --domtblout AP2.second.out AP2_first.hmm Oil_Palm.pep.all.fa
```



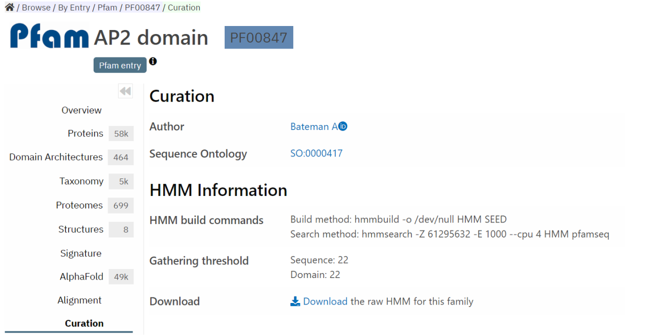
Search InterPro

Search families, domains, proteins, keywords or GO terms

PF00847

e.g. IPR020422, kinase, O00167, PF02932, GO:0007165, 1t2v, UP000005640

Powered by EBI search



Pfam AP2 domain PF00847

Curation

Author Bateman A

Sequence Ontology SO:0000417

HMM Information

HMM build commands
Build method: hmmbuild -o /dev/null HMM SEED
Search method: hmmsearch -Z 61295632 -E 1000 --cpu 4 HMM pfamseq

Gathering threshold
Sequence: 22
Domain: 22

Download
Download the raw HMM for this family

基因家族鉴定

Blast

```
#用makeblastdb建立blast数据库
makeblastdb -in ref.AP2.plant.fa -dbtype prot -out blastdb

#用blastp进行序列搜索，得到每个序列的相似序列
blastp -num_threads 20 -db blastdb -query Oil_Palm.pep.all.fa -outfmt 7 -seg yes > blastp.out

#筛选identity大于75%的序列
cat blastp.out |awk '$3>75' |cut -f1 |sort -u > blastp_result_id.list
```

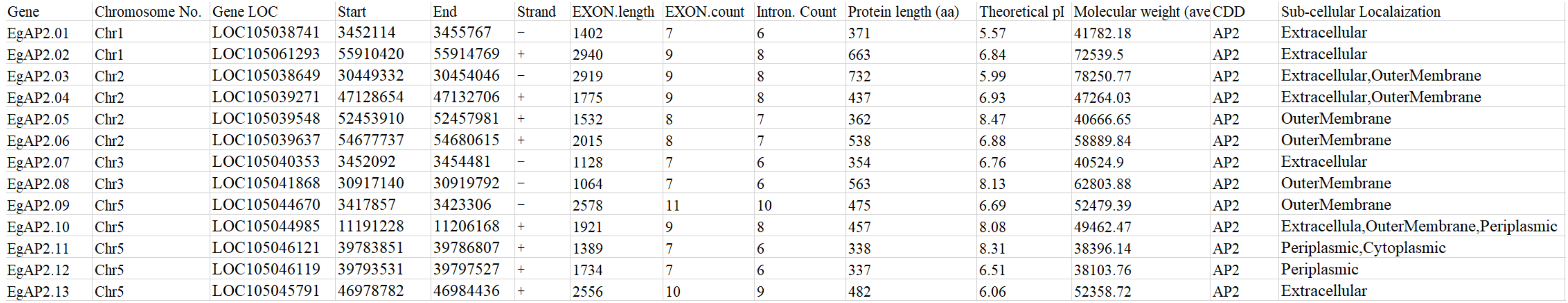
合并两种方法筛选结果

将上述两种方法得到gene id合并取交集，找出两种方法共有的基因家族成员，使结果更可信。

得到目标基因组中的AP2基因家族蛋白序列。

基因家族蛋白序列分析

蛋白基本信息



Gene	Chromosome No.	Gene LOC	Start	End	Strand	EXON.length	EXON.count	Intron. Count	Protein length (aa)	Theoretical pI	Molecular weight (ave	CDD	Sub-cellular Localaization
EgAP2.01	Chr1	LOC105038741	3452114	3455767	–	1402	7	6	371	5.57	41782.18	AP2	Extracellular
EgAP2.02	Chr1	LOC105061293	55910420	55914769	+	2940	9	8	663	6.84	72539.5	AP2	Extracellular
EgAP2.03	Chr2	LOC105038649	30449332	30454046	–	2919	9	8	732	5.99	78250.77	AP2	Extracellular,OuterMembrane
EgAP2.04	Chr2	LOC105039271	47128654	47132706	+	1775	9	8	437	6.93	47264.03	AP2	Extracellular,OuterMembrane
EgAP2.05	Chr2	LOC105039548	52453910	52457981	+	1532	8	7	362	8.47	40666.65	AP2	OuterMembrane
EgAP2.06	Chr2	LOC105039637	54677737	54680615	+	2015	8	7	538	6.88	58889.84	AP2	OuterMembrane
EgAP2.07	Chr3	LOC105040353	3452092	3454481	–	1128	7	6	354	6.76	40524.9	AP2	Extracellular
EgAP2.08	Chr3	LOC105041868	30917140	30919792	–	1064	7	6	563	8.13	62803.88	AP2	OuterMembrane
EgAP2.09	Chr5	LOC105044670	3417857	3423306	–	2578	11	10	475	6.69	52479.39	AP2	OuterMembrane
EgAP2.10	Chr5	LOC105044985	11191228	11206168	+	1921	9	8	457	8.08	49462.47	AP2	Extracellula,OuterMembrane,Periplasmic
EgAP2.11	Chr5	LOC105046121	39783851	39786807	+	1389	7	6	338	8.31	38396.14	AP2	Periplasmic,Cytoplasmic
EgAP2.12	Chr5	LOC105046119	39793531	39797527	+	1734	7	6	337	6.51	38103.76	AP2	Periplasmic
EgAP2.13	Chr5	LOC105045791	46978782	46984436	+	2556	10	9	482	6.06	52358.72	AP2	Extracellular

等电点，分子量

https://web.expasy.org/compute_pi/

亚细胞定位

WoLF PSORT: https://wolfpsort.hgc.jp

CELLO：http://cello.life.nctu.edu.tw/

Cell-PLoc 2.0 ：http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/

TargetP：http://www.cbs.dtu.dk/services/TargetP/

基因结构和保守序列

基因结构

从Gff文件筛选出对应基因的注释信息

保守序列motif

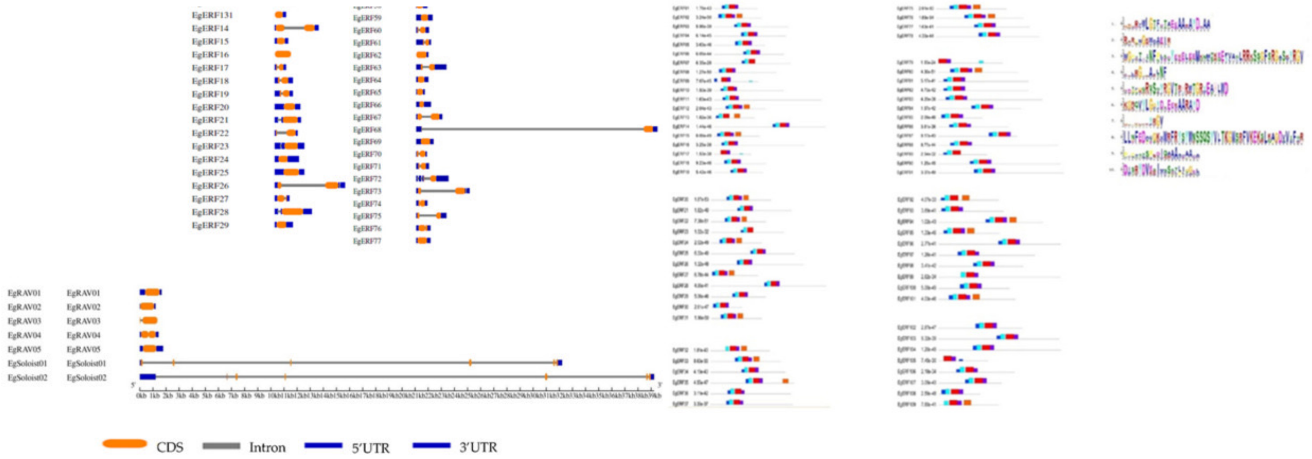
MEME

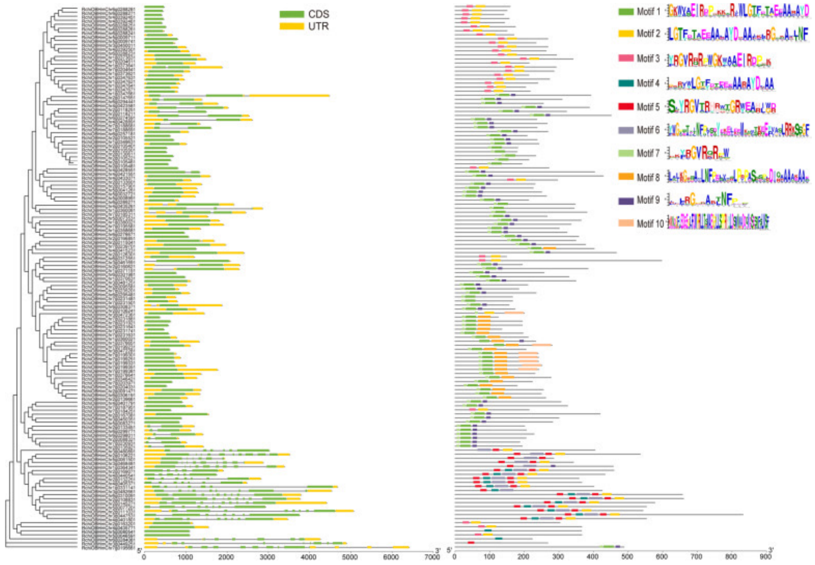
```
meme AT_AP2.fa -protein -oc ./memeresultzoops/ -nostatus -nmotifs 10 -mod zoops -minw 6 -maxw 13 -objfun classic -markov_order 0
```

https://meme-suite.org/meme/doc/meme.html

结果可视化

Tbtools
GSDS-2.0 http://gsds.gao-lab.org/
https://meme-suite.org/





系统发育分析

进化树构建

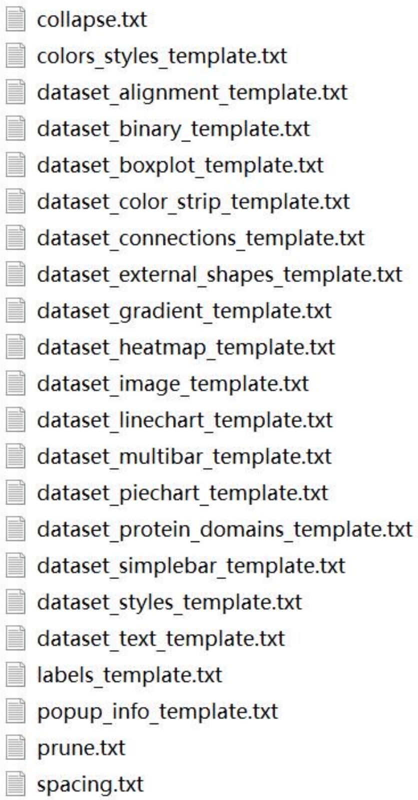
MEGA

多序列比对
进化树构建
导出后生成以.nwk结尾的树文件

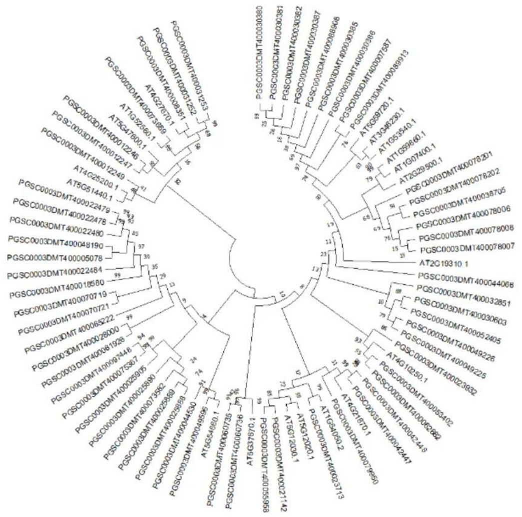
进化树美化

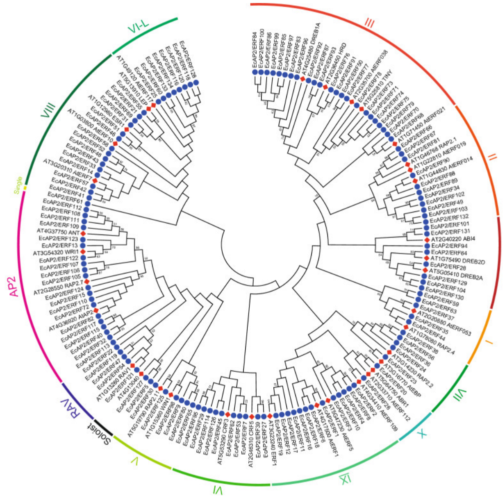
https://itol.embl.de/

https://www.evolgenius.info/evolview/#/



collapse.txt
colors_styles_template.txt
dataset_alignment_template.txt
dataset_binary_template.txt
dataset_boxplot_template.txt
dataset_color_strip_template.txt
dataset_connections_template.txt
dataset_external_shapes_template.txt
dataset_gradient_template.txt
dataset_heatmap_template.txt
dataset_image_template.txt
dataset_linechart_template.txt
dataset_multibar_template.txt
dataset_piechart_template.txt
dataset_protein_domains_template.txt
dataset_simplebar_template.txt
dataset_styles_template.txt
dataset_text_template.txt
labels_template.txt
popup_info_template.txt
prune.txt
spacing.txt





基因重复分析

获取同源基因对

Jcvi

```
python -m jcvi.formats.gff bed --type=mRNA --key=ID Oil_Palm.gff -o Oil_Palm.bed

python -m jcvi.formats.bed uniq Oil_Palm.bed

seqkit grep -f <(cut -f4 Oil_Palm.uniq.bed) Oil_Palm.pep.all.fa | seqkit seq -i > Oil_Palm.pep
seqkit grep -f <(cut -f4 Oil_Palm.uniq.bed) Oil_Palm.cds.all.fa | seqkit seq -i > Oil_Palm.cds

python -m jcvi.compara.catalog ortholog --no_strip_names Oil_Palm Oil_Palm

grep -v "#" Oil_Palm.lifted.anchors | awk '{print$1"\t"$2}' > Oil_Palm.homolog
```

计算ka/ks

利用ParaAT快速进行kaks批量运算

1.蛋白序列比对（可选 clustalw2 | t_coffee | mafft | muscle）

```
ParaAT.pl -g -t -h Oil_Palm.homolog –n Oil_Palm.cds.all.fa -a Oil_Palm.pepall.fa -m mafft -p proc -f axt -o Oil_Palm.paraat 2> paraat.log &
```

2.计算kaks值（KaKs_Calculator实现）

KaKs_Calculator3.0

```
cd Oil_Palm.paraat
for i in `ls |grep "axt"`;do KaKs -i $i -o ${i}.kaks -m YN;done
for i in `ls |grep "kaks"`;do awk 'NR>1{print $1"\t"$3"\t"$4"\t"$5}' $i >>../all.kaks;done
```

基因重复分析

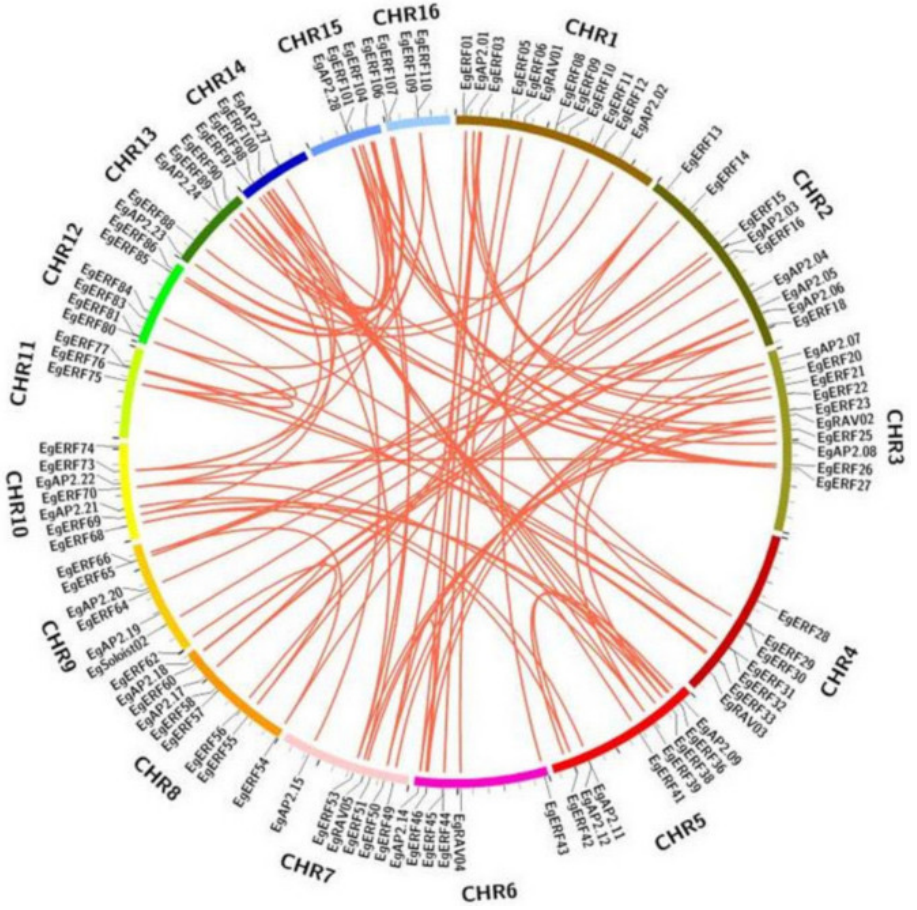
Circos可视化基因重复分析结果

```
circos -conf circos.conf

在circos.conf中添加
<plot>
        #show    = conf(show_text)
        type    = text
        file    = name.txt
        color   = black
        r1      = 1.20r
        r0      = 1.00r
        label_size    = 20p
</plot>

<link>
       show = yes
       thickness = 10
       color    = 102,205,170
       record_limit=5000
       file = highlight.txt
</link>
```

highlight.txt文件中为重复基因的对应关系及在其染色体上的位置，例
CHR1   38509565   38510844   CHR12   32333262   32334903。
name.txt文件中为重复基因在注释文件中的名称,在染色体上的位置及其在基因家族中的名称,例：CHR1   46716483   46719041   EgAP2.01。



染色体分布分析

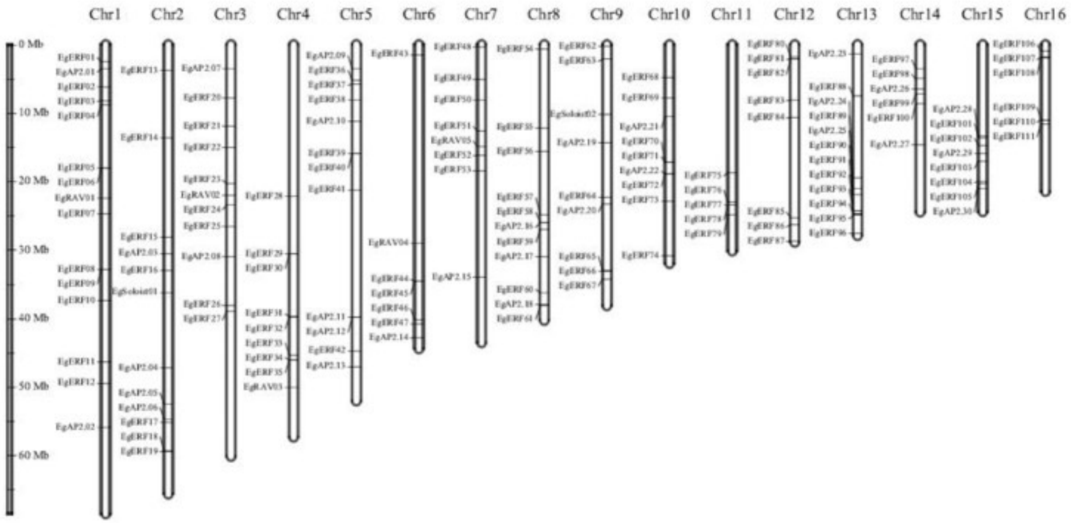
Tbtools可视化基因家族在染色体上的分布

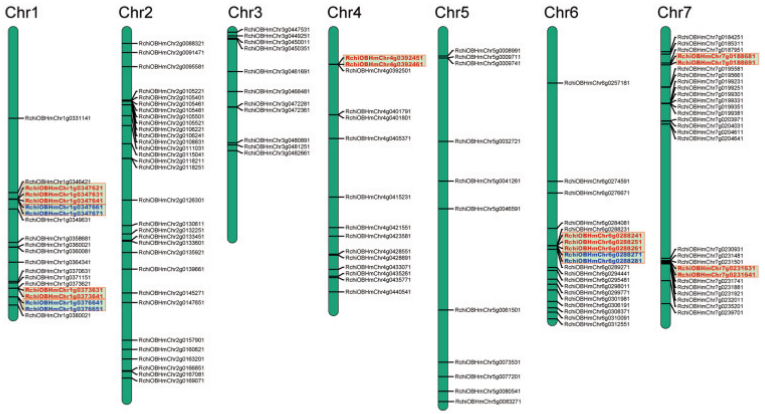
输入文件：

仅包含基因家族的Gff文件

Gene list文件

Chr id 文件





基因家族启动子区域中顺式作用元件的分析

提取启动子区域

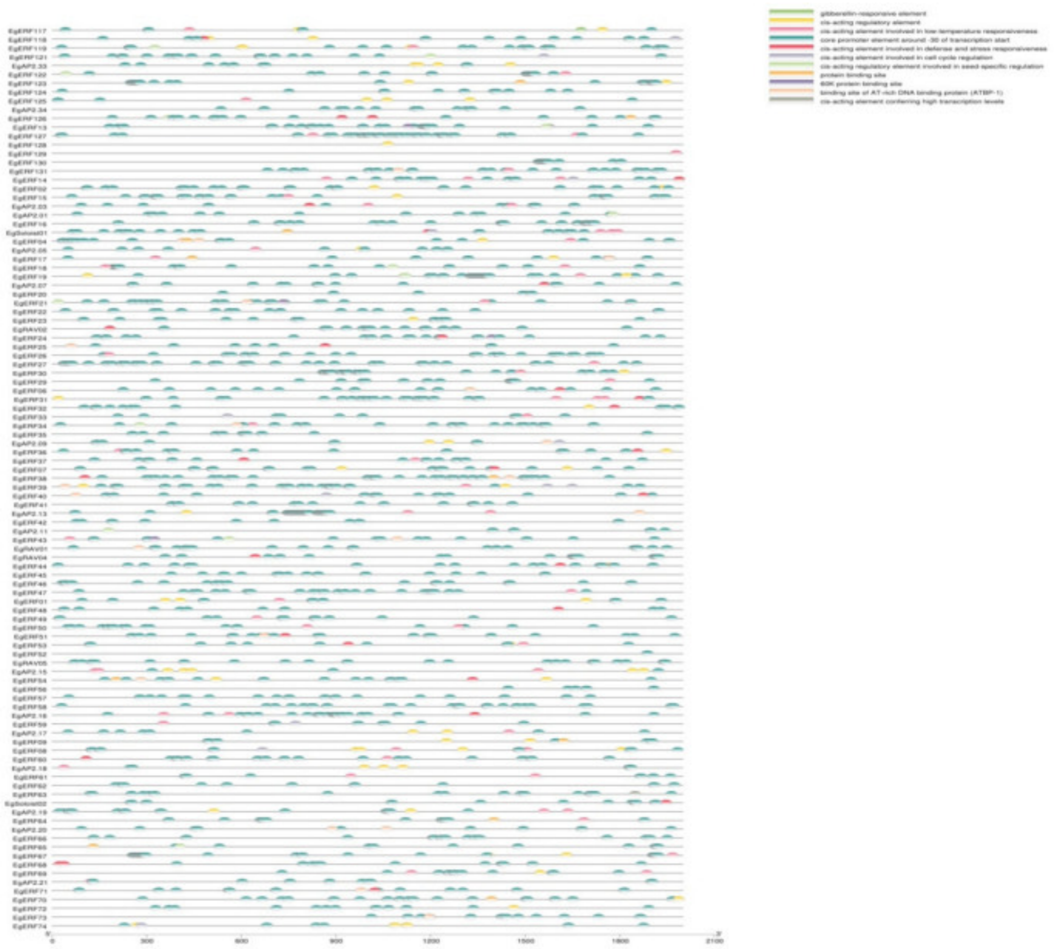
根据Gff文件，提取基因上游200bp序列

顺式作用元件分析

PlantCAR
http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

可视化顺式作用元件分析结果

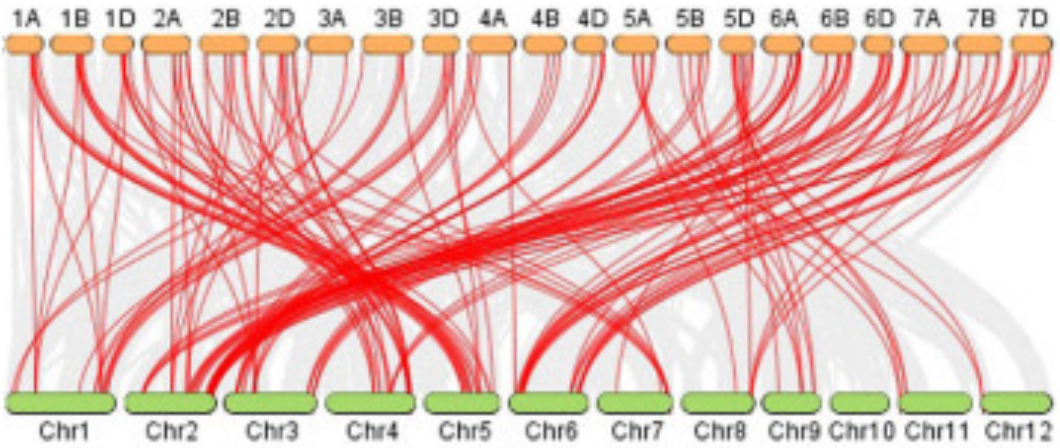
Tbtools
顺式作用元件类型及其在对应基因上的位置
基因的长度



基因家族物种间共线性分析

Jcvi分析物种间基因家族共线性

```
##生成bed文件
python -m jcvi.formats.gff bed --type=mRNA --key=ID species1.gff -o species1.bed
python -m jcvi.formats.gff bed --type=mRNA --key=ID species2.gff -o species2.bed
## 取最长转录本，对bed进行去重复
python -m jcvi.formats.bed uniq species1.bed
python -m jcvi.formats.bed uniq species2.bed
## 获取cds/pep序列
seqkit grep -f <(cut -f4 species1.uniq.bed) species1.pep.fa | seqkit seq -i >species1.pep
seqkit grep -f <(cut -f4 species1.uniq.bed) species1.cds.fa | seqkit seq -i >species1.cds
seqkit grep -f <(cut -f4 species2.uniq.bed) species2.pep.fa | seqkit seq -i >species2.pep
seqkit grep -f <(cut -f4 species2.uniq.bed) species2.cds.fa | seqkit seq -i >species2.cds
## 共线性分析
python -m jcvi.compara.catalog ortholog --no_strip_names species1 species2
## 配置species1.species2.anchors.simple
例#FF6A6A*gene1    gene1       gene2    gene2     1       +
## 共线性图可视化
python -m jcvi.compara.synteny screen --minspan=30 --simple species1. species2.anchors species1. species2.anchors.new
### 配置seqids文件
awk '{print $1}' species1.bed  |grep NC | sort -u|sort -k 1.14n | xargs echo | sed 's/ /,/g' > seqids
awk '{print $1}' species2.bed  |grep NC | sort -u|sort -k 1.14n | xargs echo | sed 's/ /,/g' >> seqids
#配置layout文件
echo ".6, .1, .8, 0, #ff8484, species1, bottom, species1.bed" > layout
echo ".4, .1, .8, 0, #ff8484, species2, top, species2.bed" >> layout
echo "e, 0, 1, species1. species2.anchors.simple" >> layout
python -m jcvi.graphics.karyotype seqids layout
```



谢谢