

# 基因家族分析

吴田振  
南京师范大学  
2022-01-13

## □ 单个基因家族的系统分析

- ✓ 基因家族成员鉴定;
- ✓ 进化树构建;
- ✓ 保守domain和motif鉴定;
- ✓ 基因结构可视化;
- ✓ 染色体定位;

## □ 基因组水平基因家族扩张与收缩分析

- ✓ 直系同源基因的鉴定;
- ✓ 构建系统发育树;
- ✓ 构建超度量树;
- ✓ 基因家族扩张与收缩分析。

# 基因家族

**概念：** 由具有类似功能的很多基因组成的一组基因集合，这些基因往往具有类似的一个或几个相同的保守结构域。

**划分：** 按功能划分：把一些功能类似的基因聚类，形成一个家族；

按照序列相似程度划分：一般将同源的基因放在一起认为是一个家族。

- 热激蛋白70家族(HSP70)是一类高度保守的分子伴侣蛋白,在细胞中协助蛋白质正确折叠；
- NBS-LRR(nucleotide-binding site and leucine-rich-repeat)是植物中最大类抗病基因家族之一；

# motif 和 domain的区别是什么？

## 1、层级不同

Motif是位于二级和三级结构之间的层次，Motif的层次接近二级结构。

Domain是位于二级和三级结构之间的层次，Domain的层次接近三级结构。

## 2、独立性不同：

domain是独立稳定的。

motif不是独立稳定的。

## 3、蛋白质结构不同：

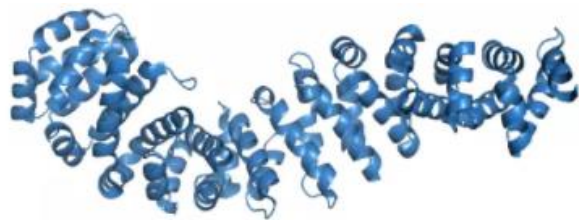
Motif在生物学中是基于数据的数学统计模型，是特定的group的序列预测。对蛋白质来说，motifs可以被定义为蛋白质（蛋白质序列）属于一个给定的蛋白质家族。

domains是一种结构实体，通常代表蛋白质结构中独立折叠和行使功能的一部分。因此蛋白质经常是这些结构域的不同组合构建起来的。

## 4、组合形式不同：

motif通常是螺旋-环-螺旋，贝塔折叠的组合、阿而法螺旋组合等。

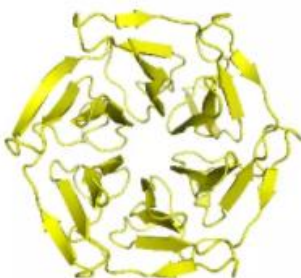
domain通常是球状压缩区或纤维状压缩区。



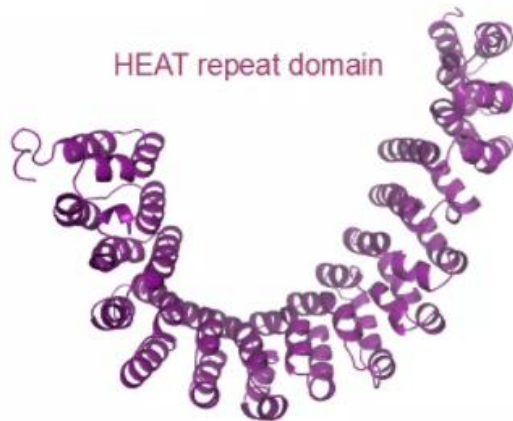
Ankyrin repeat domain



Kelch repeat domain



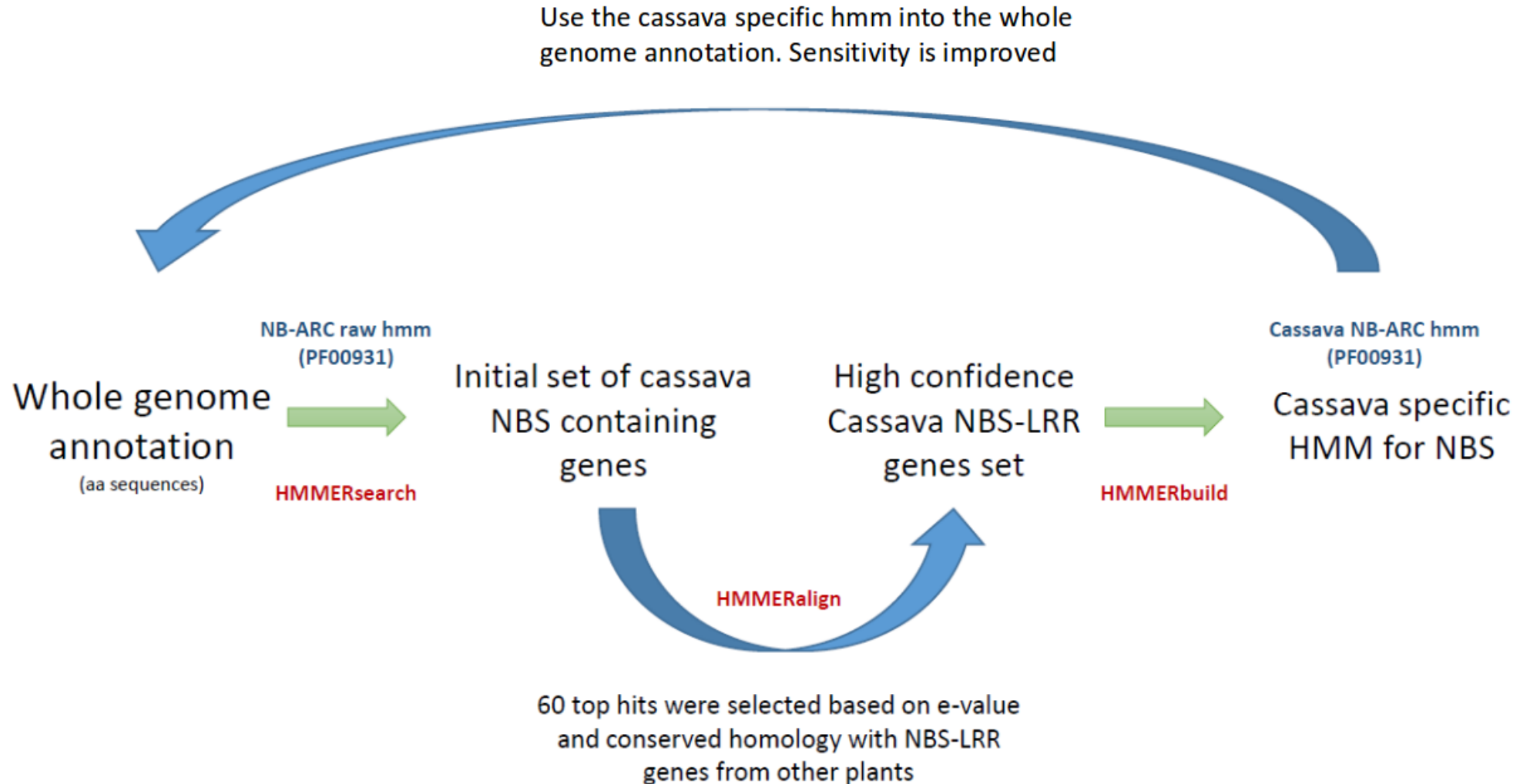
HEAT repeat domain



# ✓特定物种基因组中单个基因家族的所有成员鉴定

准备文件：物种的pep文件、基因家族隐马可夫模型hmm文件；

软件：hmmer。



# ✓特定物种基因组中单个基因家族的所有成员鉴定——文件下载

`wget -c http://ftp.ensemblgenomes.org/pub/plants/release-52/fasta/arabidopsis_thaliana/pep/Arabidopsis_thaliana.TAIR10.pep.all.fa.gz` #从NCBI 或Ensembl下载pep文件

`wget -c http://pfam.xfam.org/family/PF00931/hmm` #在pfam网站下载hmm文件

#下载模式物种或近缘物种的NBS序列，用于blast方法搜索同源序列

## Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

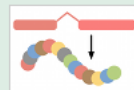
[More about this genebuild](#)

[Download genes, cDNAs, ncRNA, proteins - FASTA](#) - [GFF3](#)

[Update your old Ensembl IDs](#)

Carboxy\* CAB  
RuBisCO  
ADH F-box  
PSII

Example gene



Example transcript

[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)



Pfam 35.0 (November 2021, 19632 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

### QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW A PFAM ENTRY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

### QUERY PFAM BY KEYWORD

Search for keywords in text data in the Pfam database.

[Go](#) [Example](#)

You can also use the [keyword search box](#) at the top of every page.

Summary  
Domain organisation  
Clan  
Alignments  
HMM logo  
Trees  
Curation & model  
Species  
Structures  
AlphaFold Structures  
trRosetta Structure

Jump to...

[Go](#)

### Curation and family details

This section shows the detailed information about the Pfam family. You can see the definitions of ma

#### Curation

Seed source:	[1]
Previous IDs:	none
Type:	Domain
Sequence Ontology:	SO:0000417
Author:	Bateman A. <a href="#">ORCID</a>
Number in seed:	9
Number in full:	71983
Average length of the domain:	210.20 aa
Average identity of full alignment:	20 %
Average coverage of the sequence by the domain:	24.99 %

#### HMM information

HMM build commands: <code>build method: hmmbuild -o /dev/null HMM SEED</code> <code>search method: hmmsearch -Z 61295632 -E 1000 --cpu 4 HMM pfamseq</code>		
Model details:	Parameter	Sequence Domain
	Gathering cut-off	23.5 23.5
	Trusted cut-off	23.5 23.5
	Noise cut-off	23.4 23.4
Model length:	252	
Family (HMM) version:	25	
Download:	<a href="#">download</a> the raw HMM for this family	

## Index of /pub/plants/release-52/fasta/arabidopsis\_thaliana/pep/

<a href="#">..</a>	30-Oct-2021 18:39	6907144
<a href="#">Arabidopsis_thaliana.TAIR10.pep.abinitio.fa.gz</a>	30-Oct-2021 18:24	9691161
<a href="#">Arabidopsis_thaliana.TAIR10.pep.all.fa.gz</a>	17-Nov-2021 12:42	132
<a href="#">CHECKSUMS</a>	30-Oct-2021 18:39	2432
<a href="#">README</a>		

## ✓ 特定物种基因组中单个基因家族的所有成员鉴定——**hmmer**分析流程

```
hmmsearch --cut_tc --domtblout atha_NBS.txt hmmfile Arabidopsis_thaliana.TAIR10.pep.all.fa #搜索比对数据库，找到拟南芥的NB-ARC结构域
```

```
grep -v “#” atha_NBS.txt|awk ‘($7 + 0) < 1E-20’|cut -f1 -d “ ”|sort -u > atha_NBS_qua_id.txt #过滤掉e_value高于1e-20的id，将蛋白id存入atha_NBS_qua_id.txt
```

```
seqtk subseq Arabidopsis_thaliana.TAIR10.pep.all.fa atha_NBS_qua_id.txt > atha_NBS_qua.fa #将初步筛选出来的拟南芥中含有NB-ARC结构域的蛋白提取出来，存为NBS-ARC_qua.fa
```

```
muscle -in atha_NBS_qua.fa -out aln_atha_NBS_qua.fa #比对蛋白序列
```

```
hmmbuild second_hmm aln_atha_NBS_qua.fa #构建拟南芥NBS基因家族隐马尔科夫模型文件
```

```
hmmsearch --domtblout second_atha_NBS.txt second_hmm Arabidopsis_thaliana.TAIR10.pep.all.fa #再次比对
```

```
grep -v “#” second_atha_NBS.txt|awk ‘($7 + 0) < 1E-20’|cut -f1 -d “ ”|sort -u > second_atha_NBS_qua_id.txt  
#再次过滤掉e_value高于1e-20的id，将蛋白id存入second_atha_NBS_qua_id.txt
```

```
grep -F -f atha_NBS_qua_id.txt second_atha_NBS_qua_id.txt | sort | uniq >first_overlap_id.txt #取两次id交集
```

## ✓ 特定物种基因组中单个基因家族的所有成员鉴定——blast分析流程

NCBI下载所有植物的存在于Ref-seq（一般认为还是比较置信的植物基因序列）中的NBS序列

```
makeblastdb -in ref.nbs.plant.fa -dbtype prot #建库
```

```
Blast #序列比对
```

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

```
comm -12 blastp_result_id.list first_overlap_id.txt > common.list #取两方法id的交集
```

```
seqtk subseq Arabidopsis_thaliana.TAIR10.pep.all.fa common.list > atha_NB-  
ARC_final.fas #提取最终序列，用于后续分析
```



## ✓ 基因家族成员进化树构建——建树的简化步骤

意义：可揭示家族各成员的进化轨迹，功能差异等。

##比对蛋白序列

```
muscle -in atha_NB-ARC_final.fas -out aln_atha_NB-ARC_final.fas
```

##简化基因名

```
sed -i 's/pep.*//g' aln_atha_NB-ARC_final.fas
```

##剪切

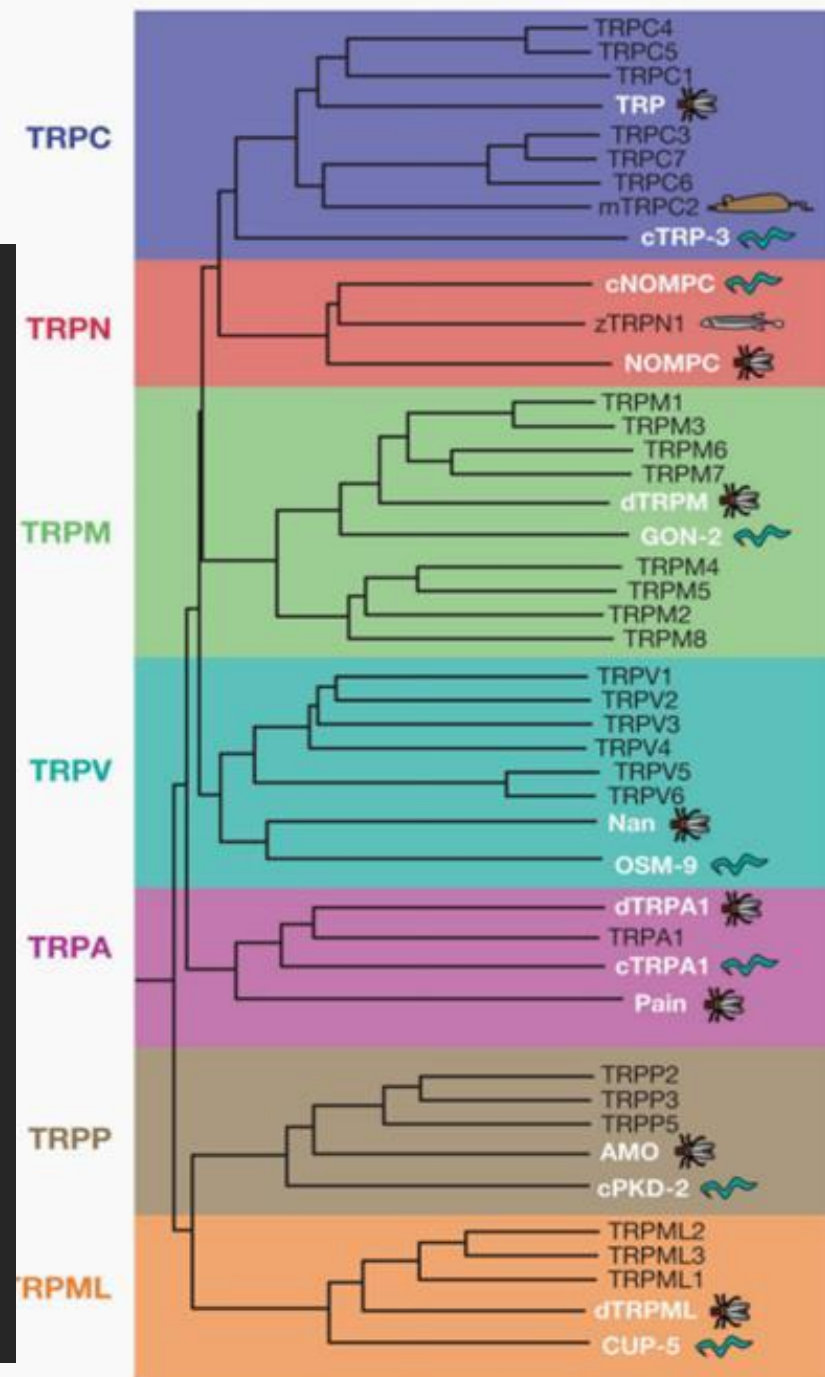
```
trimal -in aln_atha_NB-ARC_final.fas -out trimal-  
aln_atha_NB-ARC_final.fas -automated1
```

```
Gblocks trimal-aln_atha_NB-ARC_final.fas -t=p -b5=h
```

##iqtree构树,结果查看一致树文件

```
iqtree -s trimal-aln_atha_NB-ARC_final.fas-gb -pre  
outtree -bb 1000 -m MFP -nt AUTO
```

##利用Figtree等进行树的美化



## ✓ 保守motif鉴定与可视化——利用MEME鉴定

**意义：** 利用MEME搜索基因家族中成员的motif可以揭示基因家族在物种内的多样化及其功能，如果它们都含有相同的motif'表明其功能具有相似性，如果部分家族成员含有其他不同的motif，很可能这些成员有其他特异功能，或者可以归分为一个亚族。

```
meme head_cds.fas -dna -revcomp -nmotifs 10 -mod zoops -minw 5 -maxw 50 > meme_format.html
```

或

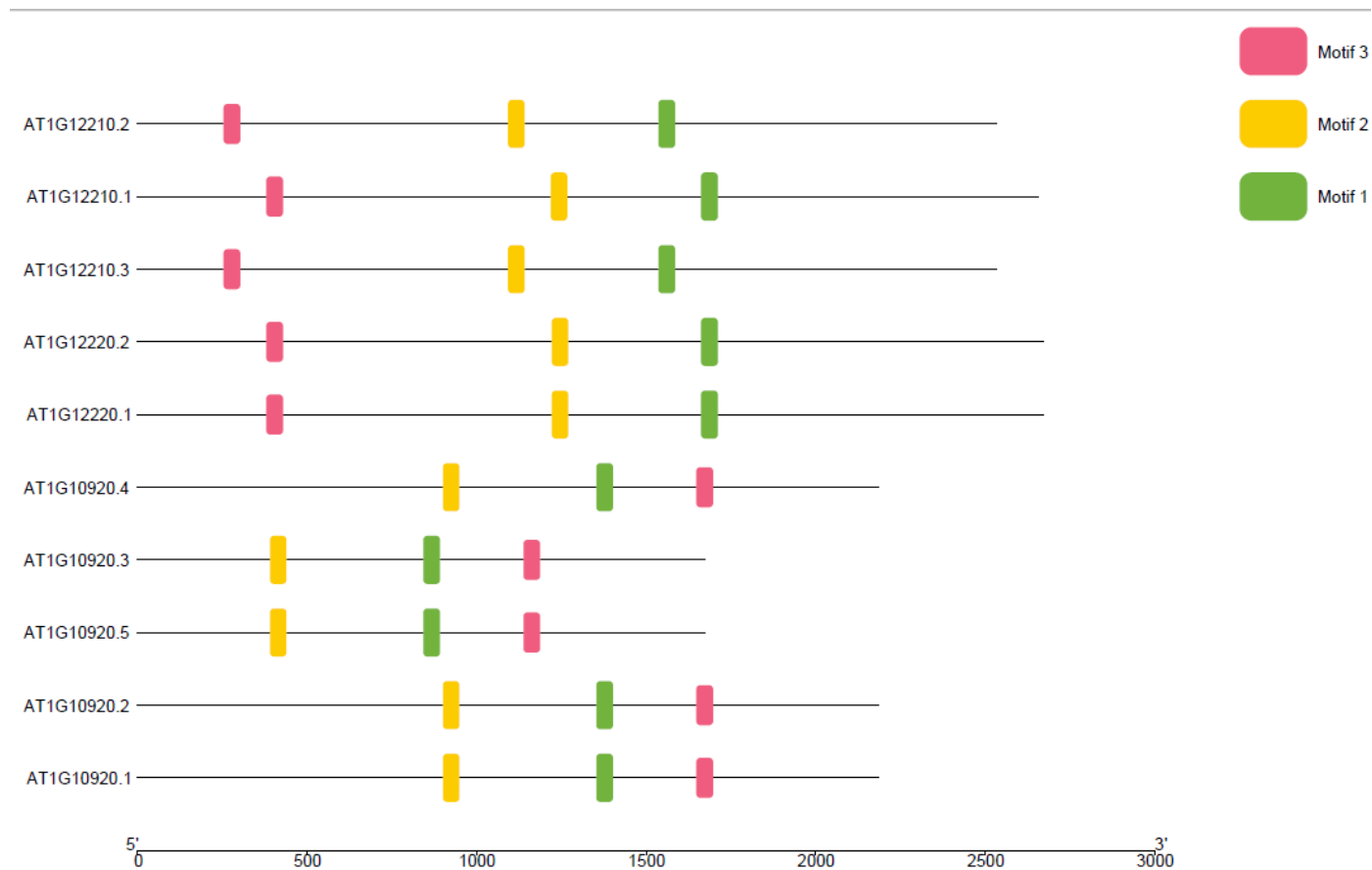
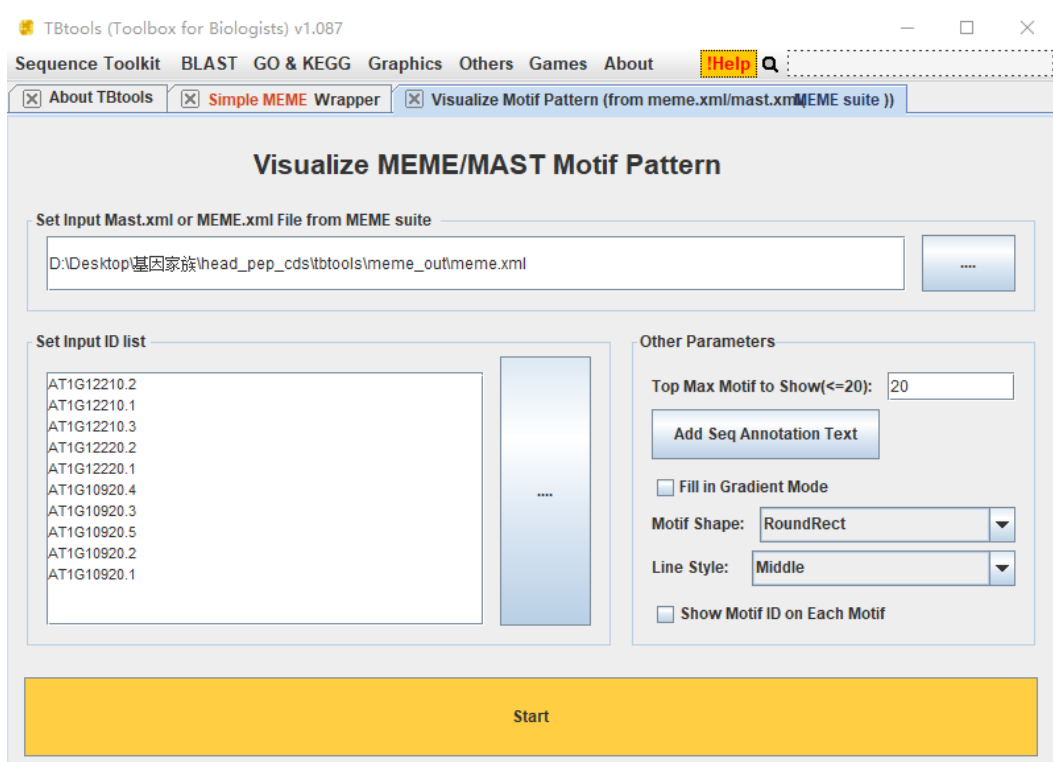


**Simple MEME Wrapper**

**Set Nucltide/Protein Sequences for Mining**

**Set Output/Working Directory**

## ✓ 保守motif鉴定与可视化——结果展示



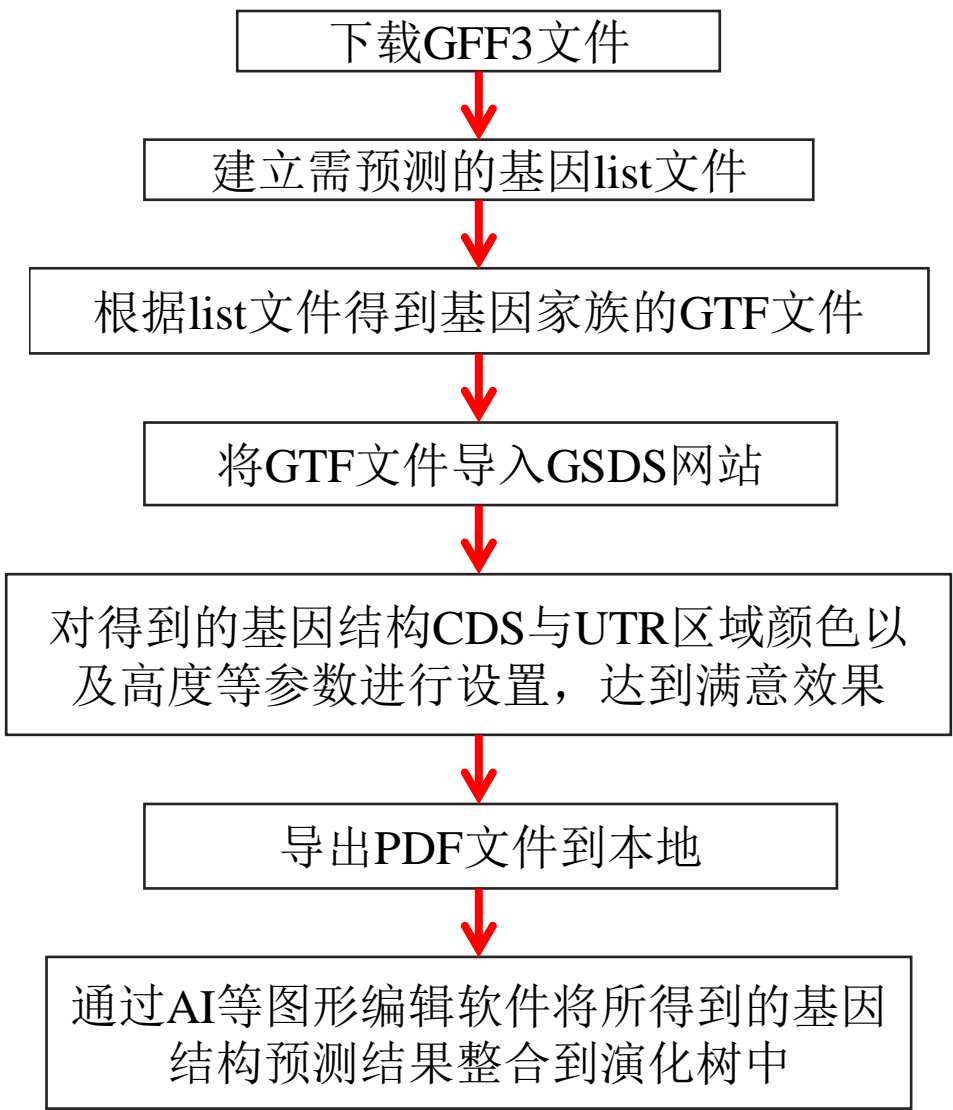
可在左侧添加分子系统发育树

Mast分析（重在确定motif的存在，较全面，是由motif查找domain的过程）

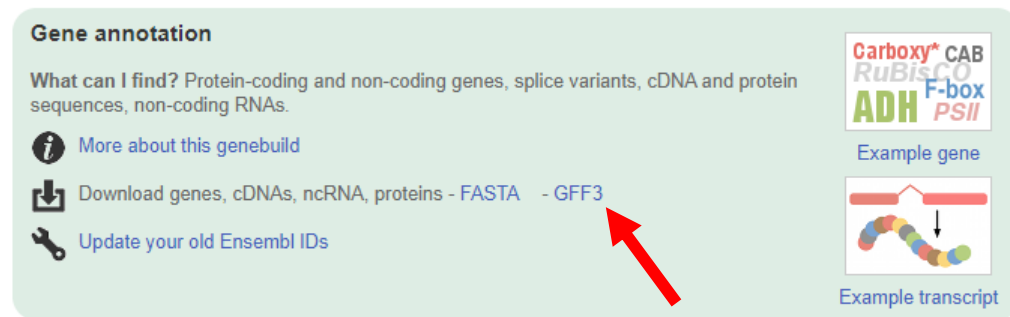
✓ 基因结构可视化——GTF文件、GSDS网站

- GFF (general feature format): 可以用于任何基因组注释的存储
- GTF (gene transfer format): 严格的用于基因注释信息的存储

列	GTF2	GFF3
reference sequence name	same	same
annotation source	same	same
feature type	feature requirements depend on software	can be anything
start coordinate	same	same
5. end coordinate	same	same
score	not used	optional
strand	same	same
frame	same	same
attributes	空格分隔	分隔



## ✓ 基因结构可视化——获取GTF文件



```
conda activate gene_family #进入分析环境
```

```
wget -c http://ftp.ensemblgenomes.org/pub/plants/release-52/gff3/arabidopsis\_thaliana/Arabidopsis\_thaliana.TAIR10.52.gff3.gz #下载gff3文件
```

```
conda install gffread #安装gffread软件
```

```
gffread Arabidopsis_thaliana.TAIR10.52.gff3 -T -o  
Arabidopsis_thaliana.TAIR10.52.gtf #将gff文件转化为gtf文件
```

利用GSDS网站  
进行可视化

```
for i in `cat first_overlap_id.txt`;do grep $i Arabidopsis_thaliana.TAIR10.52.gtf  
>> out.gtf;done #得到基因家族的转录本信息
```

```
sed -i -e 's/transcript://g' -e 's/gene_id "gene:.*//g' out.gtf #处理第九行
```

## ✓ 基因结构可视化——GSDS网站

out.gtf文件

**GSDS 2.0 Gene Structure Display Server**

Home | Help | About | FAQ | Links: PlantRegMap

● Gene Features

Format: BED

Input features in BED format:

Input data:

or upload file:

● Other Features to Display

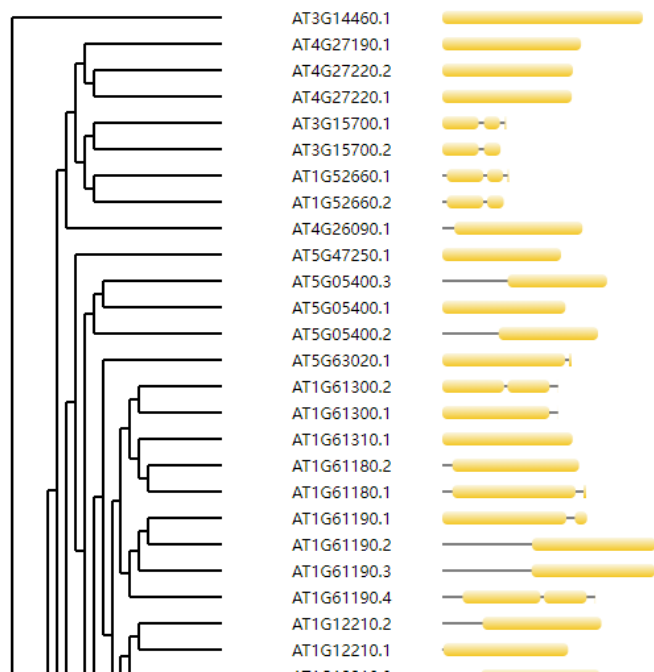
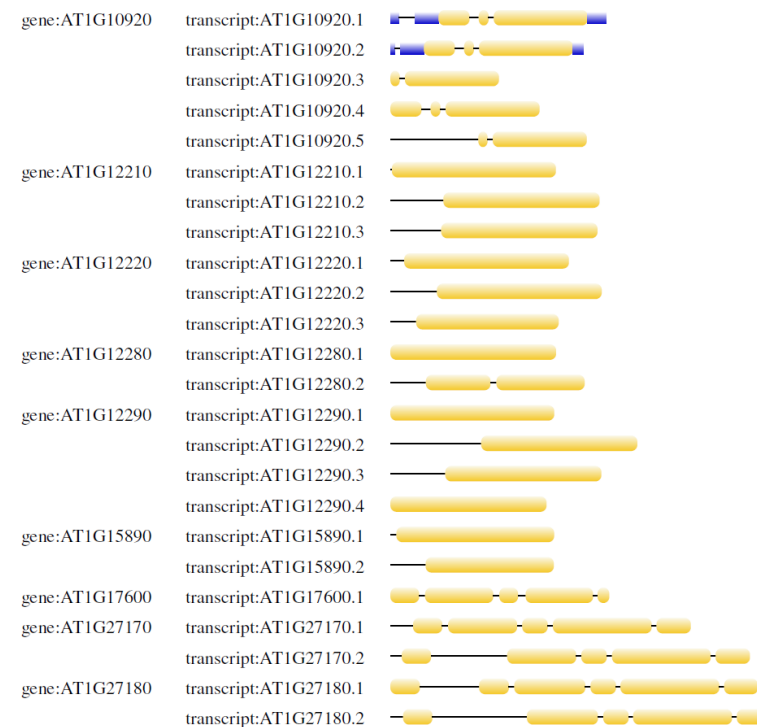
● Output (Phylogenetic Tree/Order)

● Image Format: SVG

**How to Cite:**  
Bo Hu, Jinpu Jin, An-Yuan Guo, He Zhang, Jingchu Luo and Ge Gao. (2015). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics*, 31(8):1296-1297.

© Center for Bioinformatics(CBI), Peking University.  
Any comments or suggestions, please contact:  
[gsds2@mail.cbi.pku.edu.cn](mailto:gsds2@mail.cbi.pku.edu.cn)  
[hubo.bnu@gmail.com](mailto:hubo.bnu@gmail.com)

Supported By



## ✓ 染色体定位——准备文件

1 拟南芥NBS基因id

2 拟南芥基因组的注释文件 (gff3文件)

3 拟南芥基因组长度

4 在线绘图工具: MapGene2Chrom web v2

[http://mg2c.iask.in/mg2c\\_v2.0/](http://mg2c.iask.in/mg2c_v2.0/)

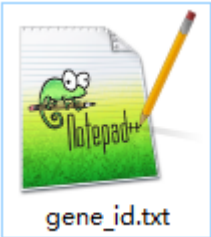
✓ 染色体定位——准备基因家族ID

	A	
1	AT1G10920.1	
2	AT1G10920.2	
3	AT1G10920.3	
4	AT1G10920.4	
5	AT1G10920.5	
6	AT1G12210.1	
7	AT1G12210.2	
8	AT1G12210.3	
9	AT1G12220.1	
10	AT1G12220.2	
11	AT1G12220.3	
12	AT1G12280.1	
13	AT1G12280.2	
14	AT1G12290.1	
15	AT1G12290.2	

数据处理

	F	
	AT1G10920	
	AT1G12210	
	AT1G12220	
	AT1G12280	
	AT1G12290	
	AT1G15890	
	AT1G17600	
	AT1G27170	
	AT1G27180	
	AT1G33560	
	AT1G50180	

存入文件





## ✓ 染色体定位——准备基因家族成员的gff3信息文件

```
grep ID=gene Arabidopsis_thaliana.TAIR10.52.gff3 > only_id_gene.txt #将标签为基因的提出来
```

```
awk '{print $9,$4,$5,$1}' only_id_gene.txt > col9451.txt #取第1、4、5、9行
```

```
awk -F '[;:]+ ' '{print $2,$(NF-2),$(NF-1),$NF}' col9451.txt > col9451.yes.txt #简化处理第1、4、5、9行
```

```
sed -i 's/[[[:space:]]//g' gene_id.txt #对基因id文件去首尾空格
```

```
for i in `cat gene_id.txt`;do grep $i col9451.yes.txt >>  
gene_fam_info_gff.txt;done #根据基因id提取家族成员位置信息，存入文件
```

## ✓ 染色体定位——准备每个染色体长度

可直接在gff3文件中查看染色体长度

或

```
conda install samtools #安装samtools软件
```

准备基因组长度文件

```
wget -c http://ftp.ensemblgenomes.org/pub/plants/release-52/fastq/arabidopsis\_thaliana/dna/Arabidopsis\_thaliana.TAIR10.dna.toplevel.fa.gz #下载拟南芥dna文件
```

```
samtools faidx Arabidopsis_thaliana.TAIR10.dna.toplevel.fa  
less Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.fai #创建索引并查看染色体长度
```

# ✓ 染色体定位——将准备文件导入网站

基因家族成员位置信息

染色体长度

MG2C

v1.0 | v1.1 | v2.0 | v2.1

parameters setting

chromosome id  
font Times N | size 12 | color black

SVG container  
width 1000 | height 900 | color white

single chromosome container  
width 270 | height 400 | fill none  
border-width 1 | border-color none

chromosome  
width 10 | height 300 | fill none  
RX 14 | RY 14 | border-width 1 | border-color black

gene lines  
color black | width 0.5 | type 1

gene id  
gene\_display\_type 1 | font Times N | size 11 | margin 15

connection between gene id and gene line  
width 0.5 | color black

download [template](#) of input1 and input2 | [FAQ\\_English](#) [常见问题\\_中文版](#)

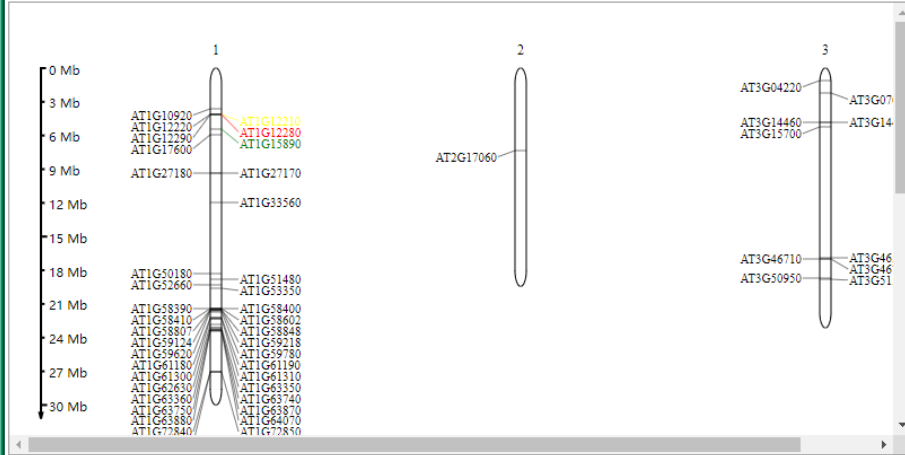
input1: Gene location which includes five fields: 1.gene\_id; 2.gene\_start; 3.gene\_end; 4.chr\_id; 5.gene\_color.

AT1G10920	364780	3647784	1	
AT1G12210	414216	4143916	1	yellow
AT1G12220	414257	4147939	1	
AT1G12280	414375	4178021	1	red
AT1G12290	4177917	4182593	1	
AT1G15890	5461309	5464241	1	green

DRAW RESET

input2: chromosome length which includes two fields: 1.chr\_id; 2.chr\_length.

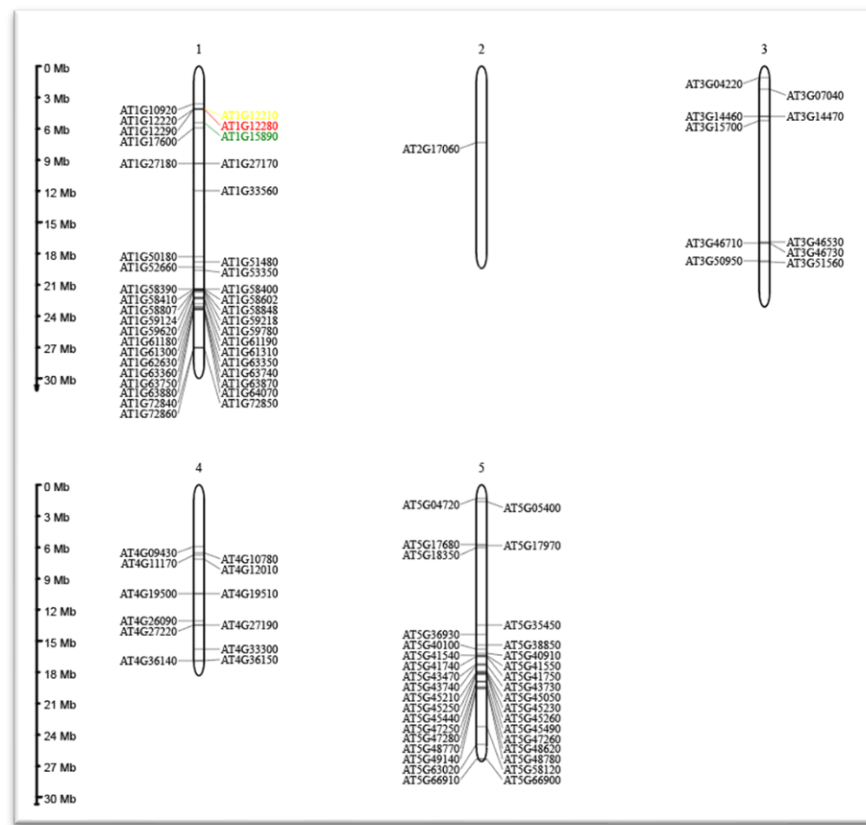
1	3042771
2	1969789
3	2347930
4	1856056
5	26975502



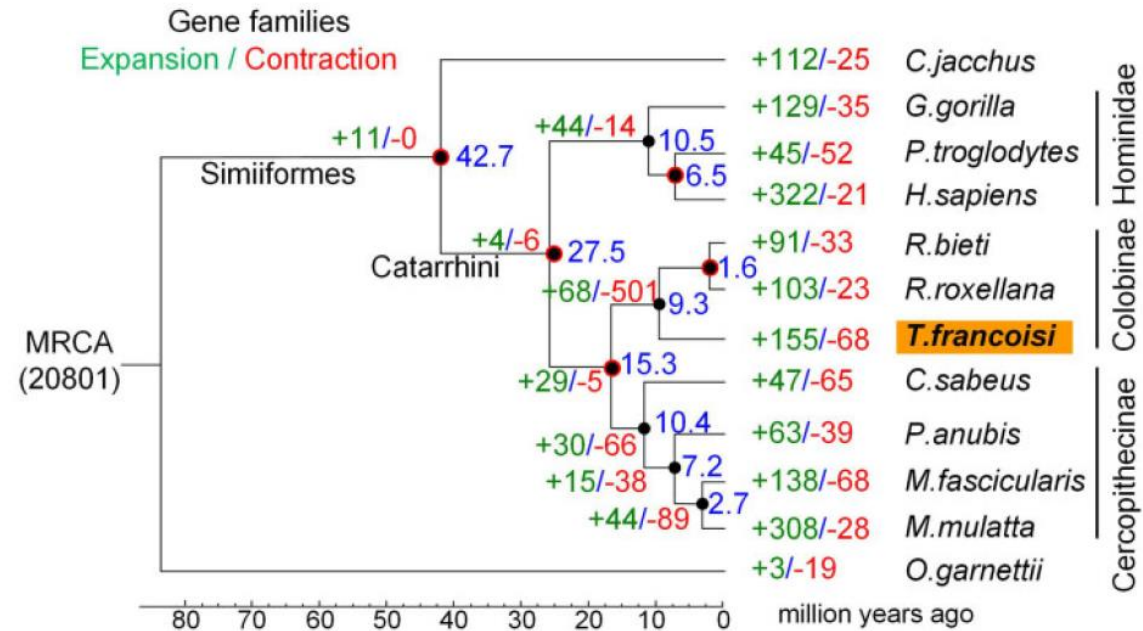
The interface displays two chromosome maps. The left map shows gene locations on chromosome 1 with a scale from 0 to 30 Mb. The right map shows chromosome lengths for chromosomes 1 to 5. The left map also includes a legend for gene colors: yellow, red, and green.



保存为SVG格式，利用AI进一步编辑



## ✓全基因组水平的基因家族收缩与扩张分析



Liu, Z., Zhang, L., Yan, Z., Ren, Z., Han, F., Tan, X., ... & Li, M. (2020). Genomic mechanisms of physiological and morphological adaptations of limestone langurs to karst habitats. *Molecular biology and evolution*, 37(4), 952-968.

流程可参考：黄鑫. 2021. 基因家族扩张收缩分析流程.

待续……