

基因家族分析基础理论与文献讲解

生信帮

云南农业大学

云南省生物大数据重点实验室

2023年7月22日

目录



雲南農業大學
Yunnan Agricultural University

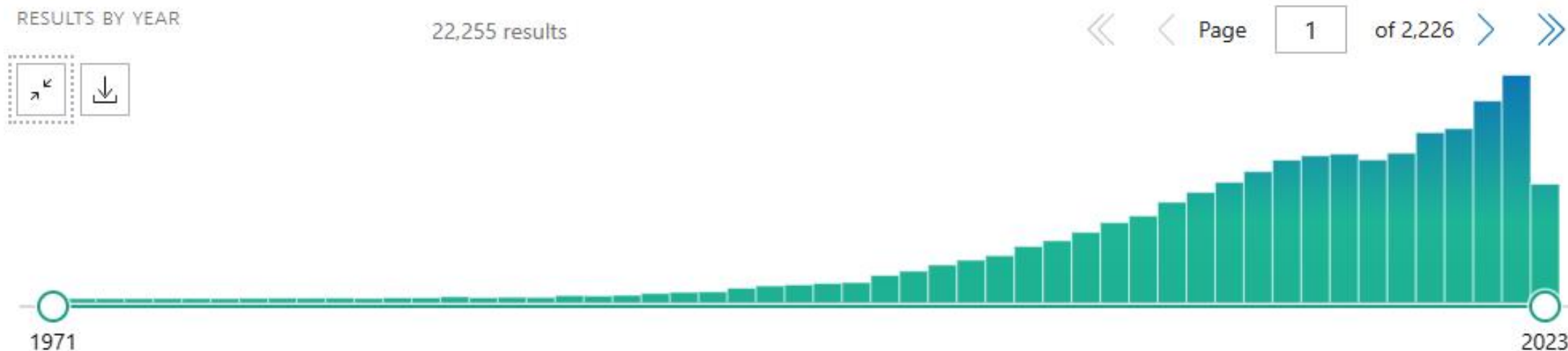
云南省生物大数据重点实验室



生信帮

- 基因家族数据下载
- 家族成员鉴定
- 进化树构建
- 基因结构展示
- 结构域展示
- **motif**展示
- 启动子鉴定
- 共线性分析
- 复制基因鉴定和Ka/Ks分析

测序价格便宜->基因组完成组装的越来越多->基因家族分析变得越来越容易



发表期刊 (2023)



云南农业大学
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮

统计了截止2022-2023年度发表的基因家族文章数量，及对应的期刊，并把近两年影响因子也列出来了！

期刊	发文数	分区	IF_2023	IF_2022	涨跌分值
Int J Mol Sci	191	BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q1); CHEMISTRY, MULTIDISCIPLINARY - SCIE(Q2)	5.6	6.208	-0.608
Front Plant Sci	156	PLANT SCIENCES - SCIE(Q1)	5.6	6.627	-1.027
Genes (Basel)	70	GENETICS & HEREDITY - SCIE(Q2)	3.5	4.141	-0.641
BMC Genomics	48	GENETICS & HEREDITY - SCIE(Q2);	4.4	4.547	-0.147
Front Genet	42	GENETICS & HEREDITY - SCIE(Q1)	3.7	4.772	-1.072
BMC Plant Biol	40	PLANT SCIENCES - SCIE(Q1)	5.3	5.26	0.04
Int J Biol Macromol	31	CHEMISTRY, APPLIED - SCIE(Q1); BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q1); POLYMER SCIENCE - SCIE(Q1)	8.2	8.025	0.175
Gene	30	GENETICS & HEREDITY - SCIE(Q2)	3.5	3.913	-0.413
PeerJ	29	MULTIDISCIPLINARY SCIENCES - SCIE(Q2)	2.7	3.061	-0.361
Plant Physiol Biochem	14	PLANT SCIENCES - SCIE(Q1)	6.5	5.437	1.063
Sci Rep	13	MULTIDISCIPLINARY SCIENCES - SCIE(Q2)	4.6	4.996	-0.396
Funct Integr Genomics	9	GENETICS & HEREDITY - SCIE(Q2)	2.9	3.674	-0.774
Plant Sci	8	PLANT SCIENCES - SCIE(Q1); BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q2)	5.2	5.363	-0.163
PLoS One	8	MULTIDISCIPLINARY SCIENCES - SCIE(Q2)	3.7	3.752	-0.052
Physiol Mol Biol Plants	8	PLANT SCIENCES - SCIE(Q2)	3.5	3.023	0.477
Plant Signal Behav	8	PLANT SCIENCES - SCIE(Q2); BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q4)	2.9	2.532	0.358

Genet Mol Biol	1	GENETICS & HEREDITY - SCIE(Q4); BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q4)	2.1	2.087	0.013
J Insect Sci	1	ENTOMOLOGY - SCIE(Q2)	2.2	2.066	0.134
Plant Physiol	1	PLANT SCIENCES - SCIE(Q1)	7.4	8.005	-0.605
Plants (Basel)	57	PLANT SCIENCES - SCIE(Q2)	4.5	4.22	0.28
Biology (Basel)	6	BIOLOGY - SCIE(Q3)	4.2	5.17	-0.97
Planta	6	PLANT SCIENCES - SCIE(Q1)	4.3	4.54	-0.24
Plant Genome	6	PLANT SCIENCES - SCIE(Q1); GENETICS & HEREDITY - SCIE(Q2)	4.2	4.219	-0.019
Protoplasma	6	PLANT SCIENCES - SCIE(Q2); CELL BIOLOGY - SCIE(Q4)	2.9	3.186	-0.286
Curr Issues Mol Biol	6	BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q3)	3.1	2.976	0.124
Mol Biol Rep	6	BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q4)	2.8	2.742	0.058
Biochem Genet	6	BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q4); GENETICS & HEREDITY - SCIE(Q4)	2.4	2.22	0.18
Fish Shellfish Immunol	5	VETERINARY SCIENCES - SCIE(Q1); MARINE & FRESHWATER BIOLOGY - SCIE(Q1); FISHERIES - SCIE(Q1); IMMUNOLOGY - SCIE(Q3)	4.7	4.622	0.078
3 Biotech	5	BIOTECHNOLOGY & APPLIED MICROBIOLOGY - SCIE(Q3)	2.8	2.893	-0.093
Plant Cell Rep	4	PLANT SCIENCES - SCIE(Q1)	6.2	4.964	1.236
Genomics	4	BIOTECHNOLOGY & APPLIED MICROBIOLOGY - SCIE(Q2); GENETICS & HEREDITY - SCIE(Q2)	4.4	4.31	0.09
Insect Sci	4	ENTOMOLOGY - SCIE(Q1)	4	3.605	0.395
Insects	4	ENTOMOLOGY - SCIE(Q1)	3	3.139	-0.139
Cells	3	CELL BIOLOGY - SCIE(Q2)	6	7.666	-1.666
Biomolecules	3	BIOCHEMISTRY & MOLECULAR BIOLOGY - SCIE(Q2)	5.5	6.064	-0.564
Physiol Plant	3	PLANT SCIENCES - SCIE(Q1)	6.4	5.081	1.319

基因家族研究的目的



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

- 鉴定基因家族的各个成员信息
- 阐明基因家族整体结构特征
- 对基因家族进行亚家族的分类
- 研究基因家族扩张与收缩情况
- 阐明各成员基因表达的规律

如何研究？



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

如何猜想？怎么入手？

- 转录因子、基因家族
 - 一个物种、多个物种联合分析
- 单个物种单个基因家族、单个基因家族
多个物种

怎样分析讨论？

- 同源基因功能研究汇总分析：已知推未知
- 进化角度分析进化来源：扩张、收缩规律，功能演化，复制事件等

查找序列

如何查找？

- 在线数据库
- BLAST
- HMM Search

可视化分析

如何可视化分析？

- 染色体定位
- 基因结构、分类
- 进化树、共线性、基因复制等

猜想



Keyword

分析讨论

具体分析流程图

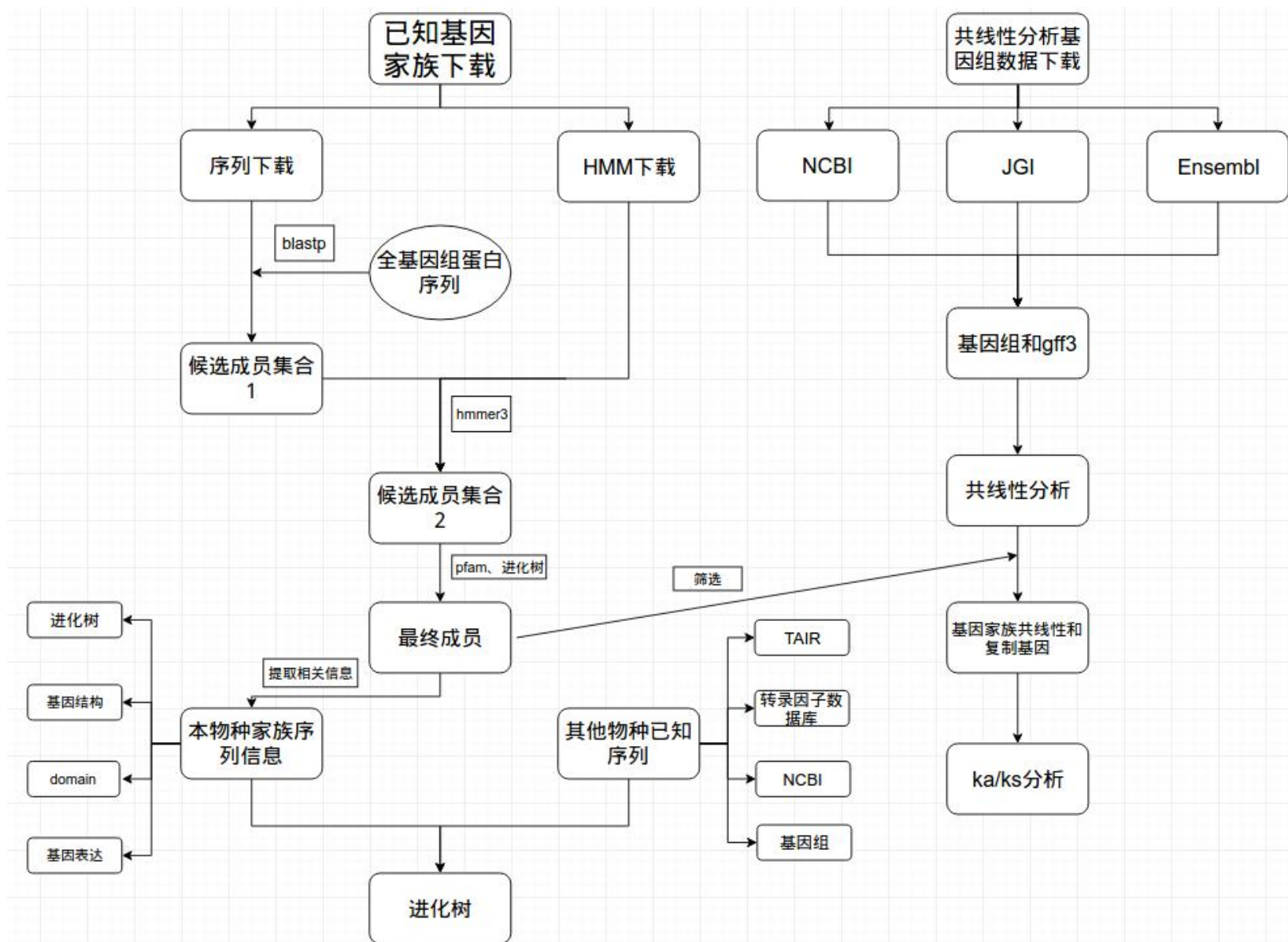


云南农业大学
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮





雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

第一章 基因家族成员鉴定方法

1、已知基因家族成员下载



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

- 根据ID或者基因家族名称在线网站检索下载：NCBI等
- 根据发表文献中的ID获取：从基因组中直接提取序列
- 基因组数据下载：获得FASTA格式的蛋白质序列文件，并提取最长转录本对应的蛋白质序列。

NCBI: <https://www.ncbi.nlm.nih.gov/>

ENSEMBL: <http://asia.ensembl.org/index.html>

Phyzone: <https://phytozome.jgi.doe.gov/pz/portal.html#>

基因组序列：作启动子分析使用等，比如说截取基因上下游2Kbp序列作启动子分析

GFF文件：里面包含基因的外显子和内含子等信息，方便基因结构的分析

Protein序列：为了鉴定基因家族成员

2、基因家族成员鉴定



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

- blastp使用：根据已知序列寻找与该序列同源性较高的序列。
- hmmer使用：根据蛋白保守结构域来寻找与该结构域相似的序列。
- 采用其中一种方法或者两种方法相互印证。

3、基因家族成员过滤



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



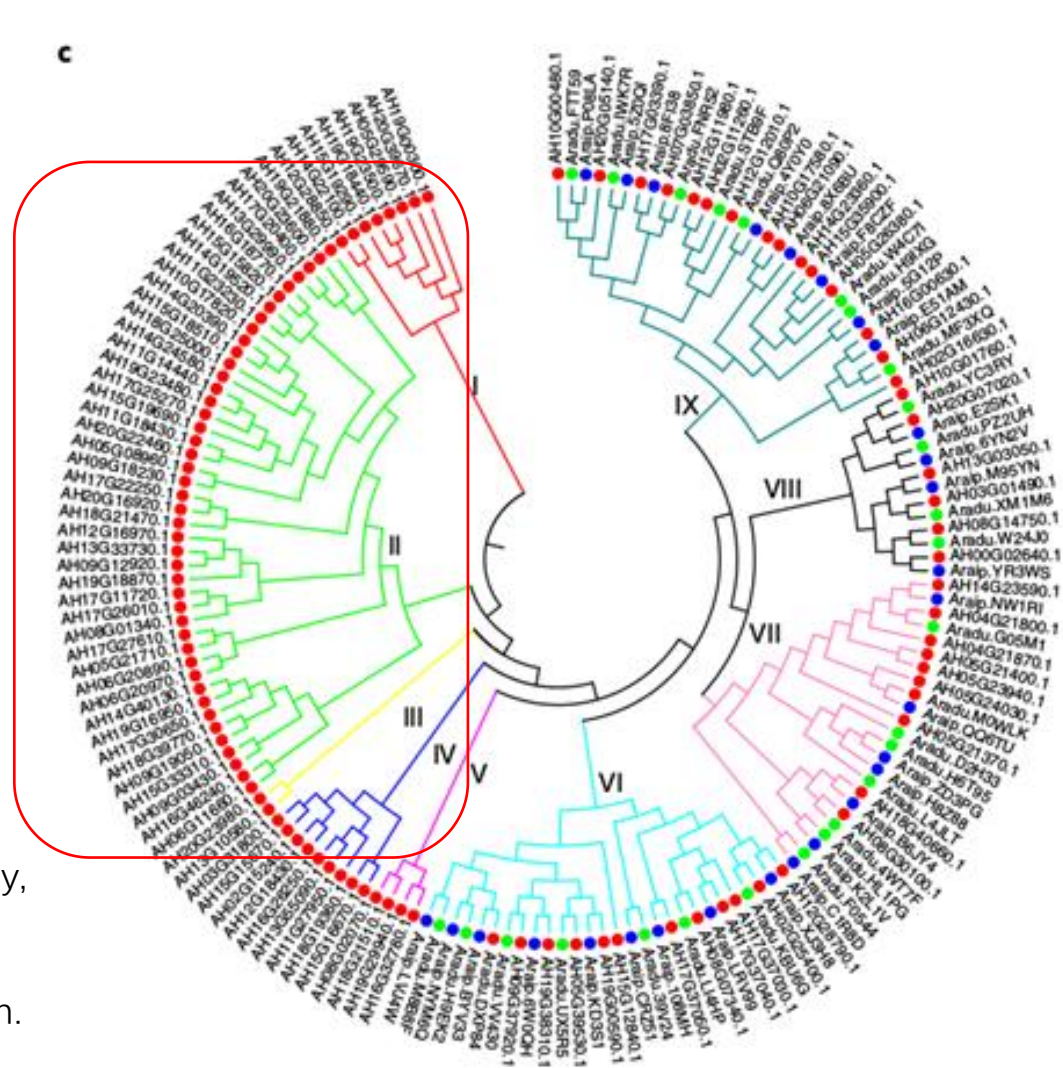
生信帮

- Pfam、CDD等数据库注释：必须含有关键结构域。
- 进化树：无独有的异常分支
- 其他特征，如序列长度、相似度，跨膜结构域个数等。

3、基因家族成员过滤



I-V分支是否可靠？



引自：Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., ... & Garg, V. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nature genetics, 51(5), 865.)

注：内实圈一种颜色代表一个物种

3、基因家族成员过滤



Identification of DTX/MATE protein family members in flax genome

The sequences of 56 MATE/DTX family members in *A. thaliana* were downloaded from the TAIR database (<http://www.arabidopsis.org/>) (Lamesch et al. 2011). BLASTP searches were performed against sequenced genomes of flax in Phytozome (v12.1; <http://www.phytozome.net/>) (Goodstein et al. 2011) using *Arabidopsis* MATE/DTX proteins as queries. All retrieved non-redundant MATE sequences (E-value was $\leq 1e^{-7}$) were collected and were preliminarily screened and filtered for the presence of conserved MATE domain (Pfam: PF01554) using the Pfam (<http://pfam.xfam.org/>) (El-Gebali et al.

2018), the simple modular architecture research tool (SMART, <http://smart.embl-heidelberg.de/smart/batch.pl>) (Letunic et al. 2014), and conserved domain database (CDD; <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) (Marchler-Bauer et al. 2014). The sequences were assessed against the expected features of the *MATE* transporters in plants (e.g., containing *MATE/DTX* domains, 8-12 transmembrane domains, ≥ 400 amino acid residues, high similarity to plant *MATEs* in comprehensive databases). Redundant and partial genes were manually removed. Physical parameters of

(2020). Genome-wide identification and expression analysis of detoxification efflux carriers (dtx) genes family under abiotic stresses in flax. *Physiologia Plantarum*.

4、命名



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

基因家族命名

- 借助于模式生物（参考物种）的同源基因命名（NCBI比对）
- 按照染色体顺序和位置。

4、命名



借助于拟南芥同源基因命名

Table 1. Information of longan *Sm* family genes.

Gene ID	Gene name	Comments to members of <i>Arabidopsis</i>
Dlo_023124.1	<i>DlLSm1a</i>	<i>AT3G14080 AtLSm1b</i>
Dlo_023126.1	<i>DlLSm1b</i>	<i>AT3G14080 AtLSm1b</i>
Dlo_024845.1	<i>DlLSm2</i>	<i>AT1G03330 AtLSm2</i>
Dlo_009273.1	<i>DlLSm3</i>	<i>AT1G76860 AtLSm3b</i>
Dlo_014195.1	<i>DlLSm4a</i>	<i>AT5G27720 AtLSm4</i>
Dlo_021351.1	<i>DlLSm4b</i>	<i>AT5G27720 AtLSm4</i>
Dlo_031761.1	<i>DlLSm4c</i>	<i>AT5G27720 AtLSm4</i>
Dlo_023673.1	<i>DlLSm5a</i>	<i>AT5G48870 AtLSm5</i>
Dlo_012057.1	<i>DlLSm5b</i>	<i>AT5G48870 AtLSm5</i>

Li Xue., Chen Yan., Zhang Shuting., Su Liyao., Xu Xiaoping., Chen Xiaohui., Lai Zhongxiong., Lin Yuling.(2020). Genome-wide identification and expression analyses of *Sm* genes reveal their involvement in early somatic embryogenesis in *Dimocarpus longan* Lour. PLoS One, 15(4), e0230795. doi:10.1371/journal.pone.0230795

案例-玉米Hsf鉴定



雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮

Identification and bioinformatic analysis of Hsfs in maize. The maize reference genome (Maize B73 RefGen_V4) was downloaded from the MaizeGDB database on the website ftp://ftp.ensemblgenomes.org/pub/release-41/plants/fasta/zea_mays/dna/. 47 non-redundant Hsfs amino acid sequences were downloaded from PlantTFDB and blasted against the amino acid sequences retrieved from the maize reference genome. Additionally, the Hsf-type DBD domain (Pfam: PF00447) was used as a query in BLASTP ($P = 0.001$) search for probable Hsf protein sequences in the maize genome reference data base. 58 maize Hsf proteins were collected after homologous comparison and HMM research. Results from the two searches described above were blasted, respectively, against the UniprotKB/Swiss-prot data base on the NCBI blastp suite and the NCBI batch CDD (<https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>). Redundant sequences were discarded. Using SMART, sequences without a DBD domain or HR-A/B domain were eliminated. Motif elicitation of Hsf proteins was completed using the MEME suite (<http://meme-suite.org/tools/meme>). The pI and MW of identified Hsf proteins were calculated using the Expasy website (https://web.expasy.org/compute_pi/). WoLFPSORT was used to predict subcellular localization motifs in amino acid sequences of the identified Hsf proteins (<https://wolfpsort.hgc.jp/>).

Zhang Huaning., Li Guoliang., Fu Cai., Duan Shuonan., Hu Dong., Guo Xiulin.(2020). Genome-wide identification, transcriptome analysis and alternative splicing events of Hsf family genes in maize. Sci Rep, 10(1), 8073. doi:10.1038/s41598-020-65068-z

案例-竹子KXK家族鉴定



雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮

Identification of HXKs in Moso Bamboo

To identify HXK genes of moso bamboo, we used *Arabidopsis* HXK protein sequences (TAIR¹) as query sequences for a BLAST search of the moso bamboo database (Zhao et al., 2018²) using a threshold P -value $< 10^{-5}$. Fifteen candidate HXK sequences were obtained. Three sequences were discarded as a result of a search for conserved domains within the amino acid sequences of the National Center for Biotechnology Information (NCBI) Conserved Domains database³ and the EMBL-EBI Pfam database⁴. The number of amino acids and length of the coding sequence (CDS) were characterized, and the theoretical isoelectric point (pI) and molecular weight of amino acids were calculated using the ProtParam online tool⁵ (Gasteiger et al., 2003).

Zheng Wenqing., Zhang Yuan., Zhang Qian., Wu Ruihua., Wang Xinwei., Feng Shengnian., Chen Shaoliang., Lu Cunfu., Du Liang.(2020). *Phyllostachys edulis* Genome-Wide Identification and Characterization of Hexokinase Genes in Moso Bamboo (). *Front Plant Sci*, 11(undefined), 600.
doi:10.3389/fpls.2020.00600



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

第二章 基因家族特征分析

基因家族特征分析基本工具汇总



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

- ✓ 工具计算蛋白质序列的理论等电点 (PI) 和分子量 (MW) :

Compute pI / Mw (https://web.expasy.org/compute_pi/)

- ✓ 亚细胞定位: WoLF PSORT (<https://wolfpsort.hgc.jp/>)

- ✓ 染色体定位: http://mg2c.iask.in/mg2c_v2.1/

- ✓ Motif: MEME (<http://meme-suite.org/tools/meme>)

- ✓ Domain:

Pfam (<http://pfam.xfam.org/search#tabview=tab1>)

CDD (<https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>)

SMART (<http://smart.embl.de/>)

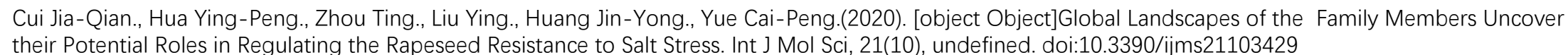
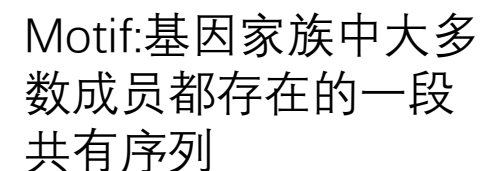
- ✓ 基因结构: GSDS (<http://gsds.gao-lab.org/>)

- ✓ 启动子: PlantCARE (<http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>)

- ✓ 联合绘图: Evolview (<https://www.evolgenius.info/evolview-v3/#login>)

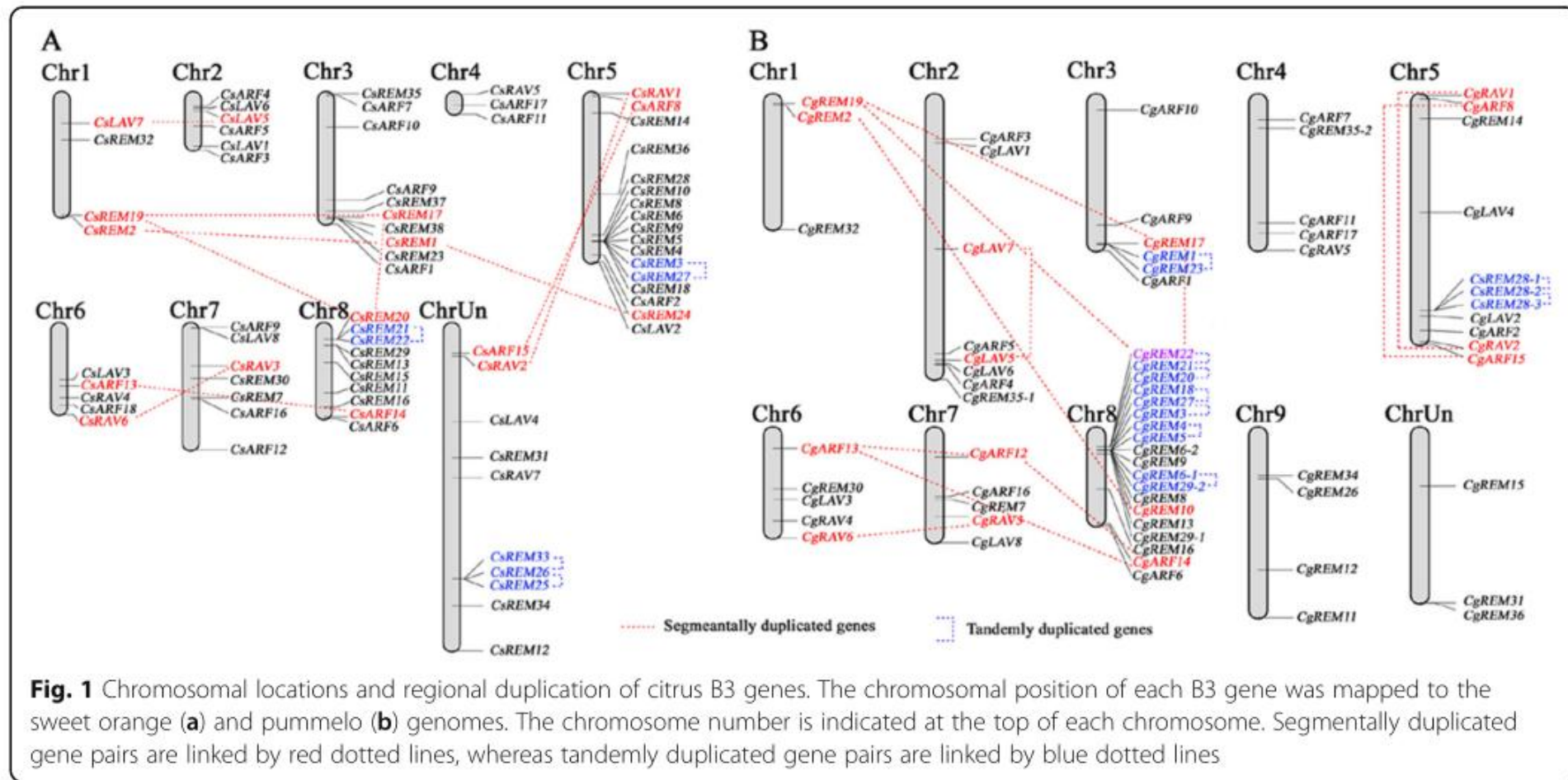


生信帮



Diao Donghui., Hu Xiao., Guan Dan., Wang Wei., Yang Haiqing., Liu Yueping.(2020). Genome-wide identification of the ARF (auxin response factor) gene family in peach and their expression analysis. *Mol Biol Rep*, 47(6), 4331-4344. doi:10.1007/s11033-020-05525-0

3、染色体分布展示



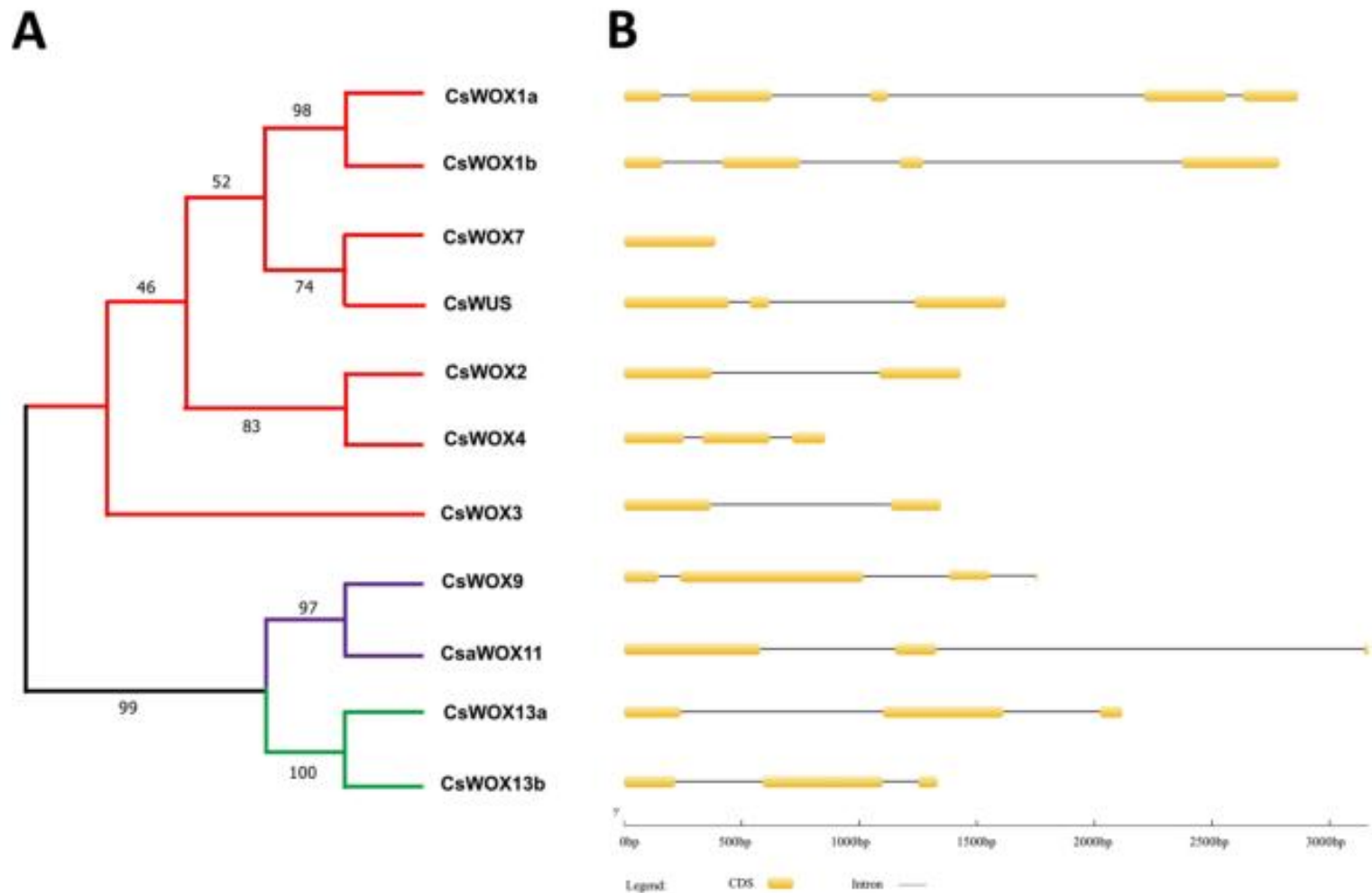
注：基因复制基因
连线连接

Liu Zheng., Ge Xiao-Xia., Wu Xiao-Meng., Xu Qiang., Atkinson Ross G., Guo Wen-Wu.(2020). Genome-wide analysis of the citrus B3 superfamily and their association with somatic embryogenesis. BMC Genomics, 21(1), 305. doi:10.1186/s12864-020-6715-9

4、基因结构



展示内含子、外显子和UTR等结构

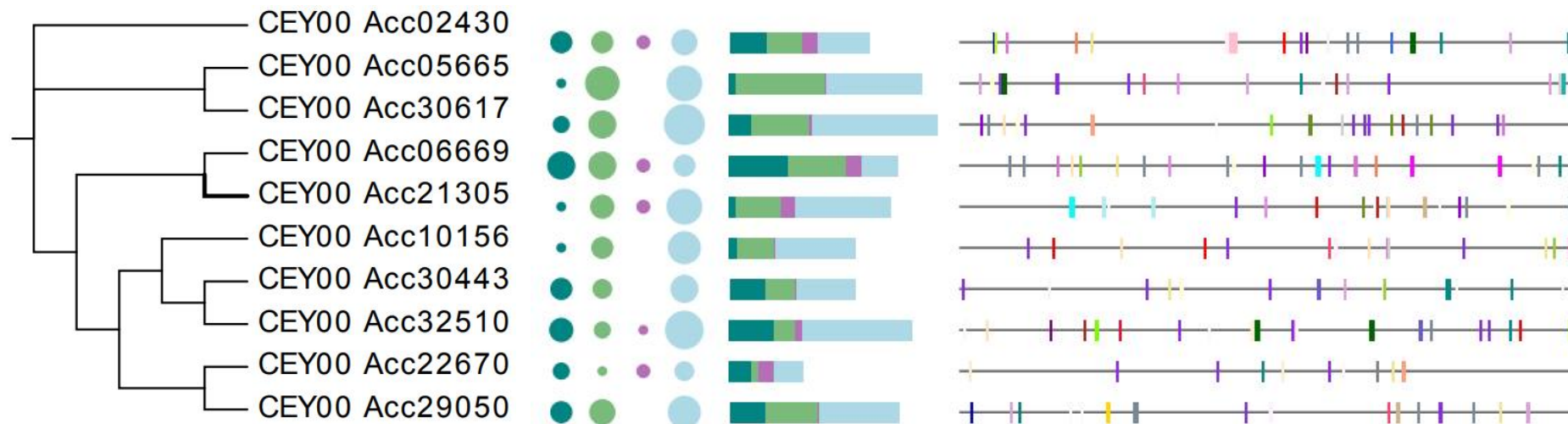


5、启动子序列

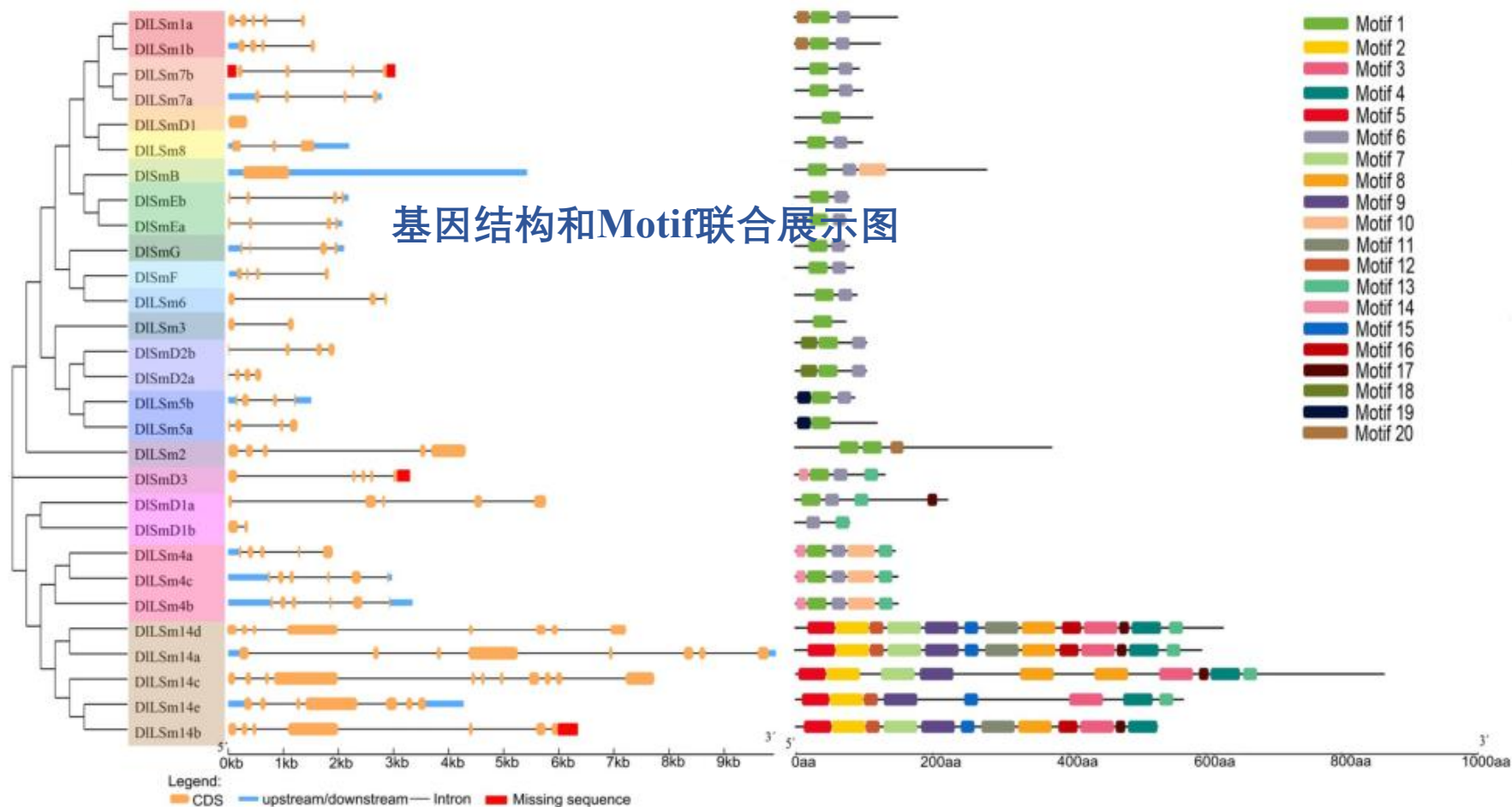


启动子元件的分布及大类划分：

- hormone-response elements
- plant growth related elements
- environmental stress
- responsive to light



6、联合绘图



Li Xue., Chen Yan., Zhang Shuting., Su Liyao., Xu Xiaoping., Chen Xiaohui., Lai Zhongxiong., Lin Yuling.(2020). Genome-wide identification and expression analyses of Sm genes reveal their involvement in early somatic embryogenesis in *Dimocarpus longan* Lour. PLoS One, 15(4), e0230795. doi:10.1371/journal.pone.0230795

6、联合绘图

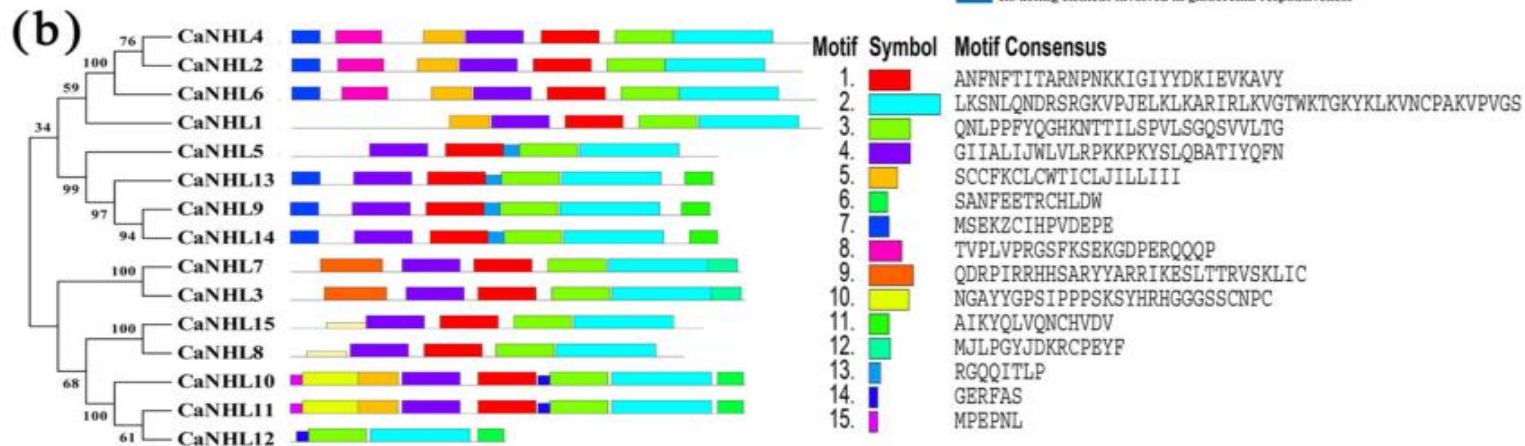
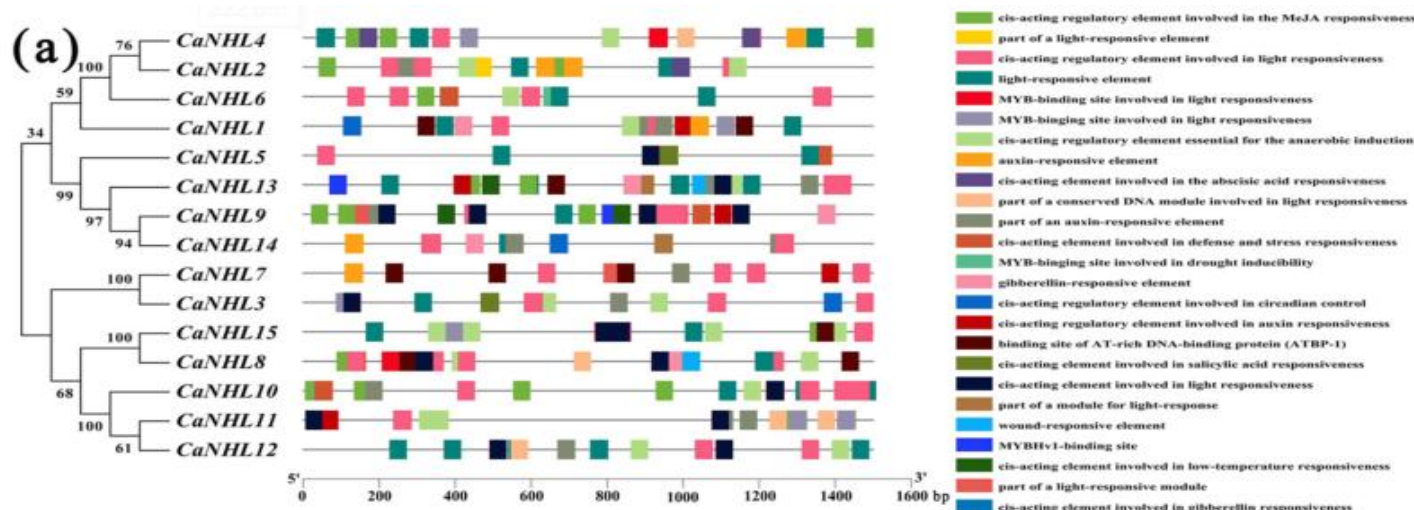


雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮



启动子+Motif

Liu Changyun., Peang Haoren., Li Xinyu., Liu Chaolong., Lv Xing., Wei Xuefeng., Zou Aihong., Zhang Jian., Fan Guangjin., Ma Guanhua., Ma Lisong., Sun Xianchao.(2020). Capsicum annuumGenome-wide analysis of NDR1/HIN1-like genes in pepper (L.) and functional characterization of under biotic and abiotic stresses. Hortic Res, 7(undefined), 93. doi:10.1038/s41438-020-0318-0



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

第三章 基因家族进化分析方法

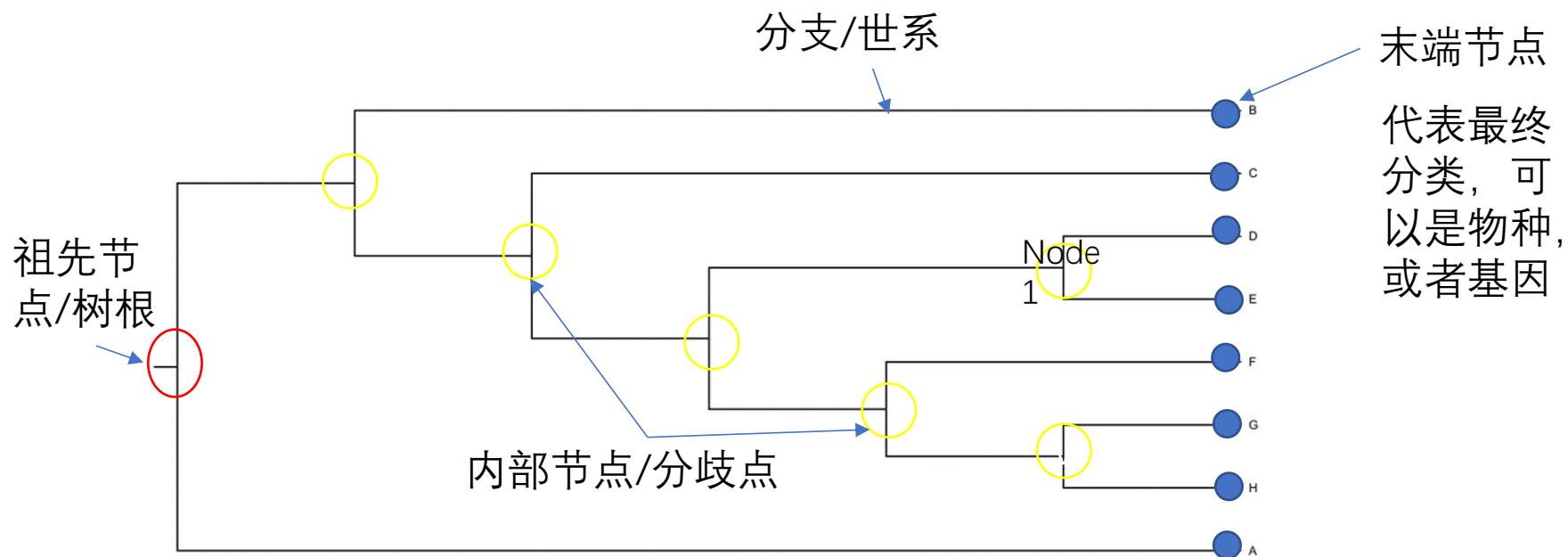
一、进化树基本术语



1、进化树

确定物种之间的进化关系，通常我们用树型结构来表示，即系统发育树或进化树。

((((D,E),((G,H),F)),C),B),A)



父节点: Node1,其对应的子节点是D和E, D和E为同胞节点

一、进化树基本术语



雲南農業大學
Yunnan Agricultural University

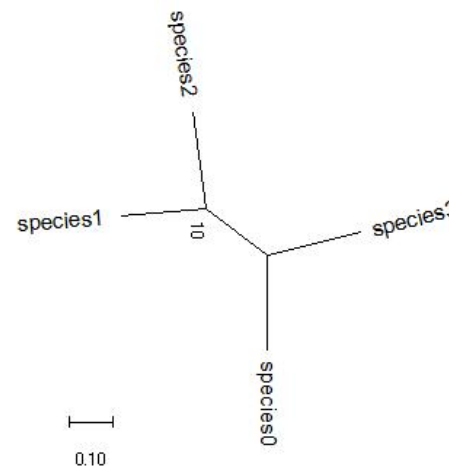
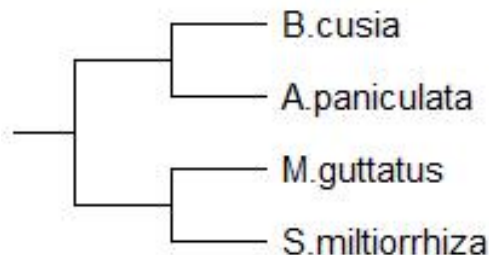


云南省生物大数据重点实验室

生信帮

2、有根树和无根树

- 有根树反映树上物种（基因）演化的先后**时间**顺序,而无根树则只反映要分类的物种（基因）之间的**距离**，并不涉及谁是谁的祖先问题。
- 无根树(不指定外群时，软件产生的树)在引入一个外部远缘类群（**外群**）后可以转化为一棵有根树。也可以采用MEGA中点生根(Mid-point rooted tree)。



一、进化树的基本术语



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

3、Bootstrap值

为进化树可靠性评估指标。基因家族分析一般采用1000次。

Bootstrap检测的具体方法为：经过多序列比对后，序列长度为 L 列，我们从这 L 列中进行 L 次有返回抽样，构成一组新的多序列比对结果，然后对其进行似然估计，构建进化树，如此重复 N 次，如果有 M 次与我们之前构建的进化树的拓扑结构一致，则我们就可以得到反映进化树可信度的bootstrap值，即 $M/N \times 100$ 。一般我们认为bootstrap > 70 即为比较可靠。

一、进化树的基本术语



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

4、Newick格式介绍

Newick 是最常见的进化树文件格式，其为一个文本文件。采用**圆括号**将**同胞节点**括起来，多个节点之间用逗号相连分支长度信息，如分化时间，遗传距离等。对于分支的信息，我们用节点的属性来表示，与节点的名称之间用**冒号**分割

(A:0.1,(B:0.1,(C:0.1,((D:0.1,E:0.1),(F:0.1,(G:0.1,H:0.1):0.1):0.2):0.1):0.3):0.1)

二、进化树的构建方法



雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮

方法	软件	替代模型	数据	序列要求	速度
距离法	MEGAx	Kimura 2-parameter/p-distance	snp数据、基因序列、线粒体序列等	具有较高相似性	最快
最大似然法	RAxML	基于GTR的系列模型		任何相关序列集合	较快
	Phyml	JC69、HKY85、GTR等			中等
	iqtree	替代模型		任何相关序列集合	快
贝叶斯法	Mrbayes	单参数、双参数、GTR等		任何相关序列集合	慢
	Exabayes	GTR			慢

二、进化树的构建方法



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

- 基因家族的进化分析，由于序列相似度较高，可以采用NJ法进行构树，常用的软件为MEGA软件。也可以采用ML方法，如IQ-TREE
- 分析时间：距离法 < 最大似然法。

三、进化树构建步骤

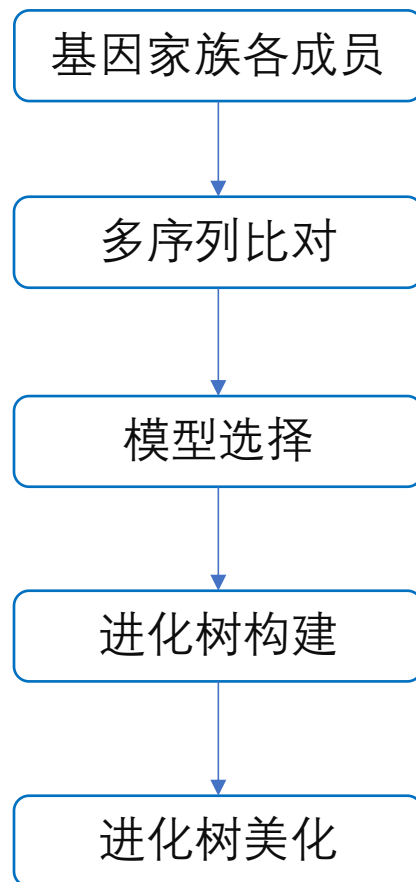


雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮



四、基因家族进化树的构建



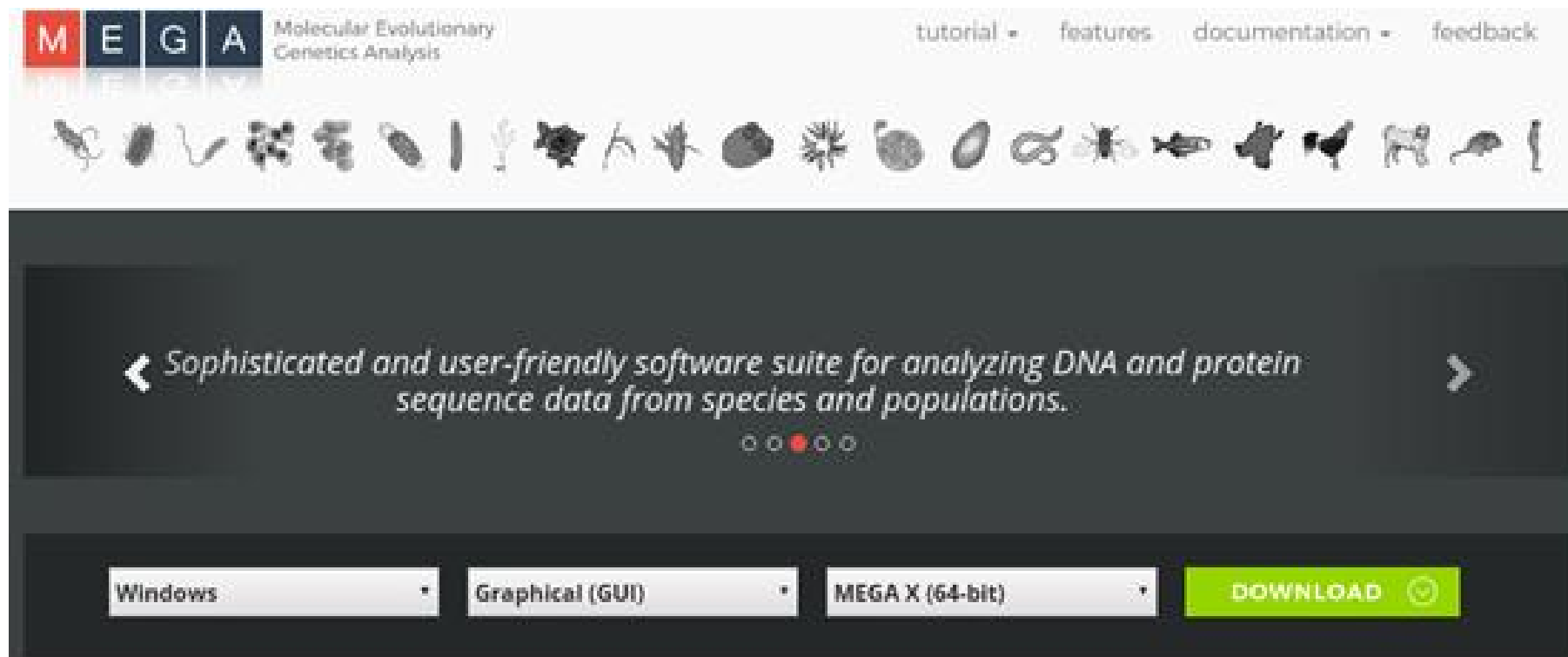
雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

网站: <https://www.megasoftware.net>



参考文献: Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

四、基因家族进化树的构建



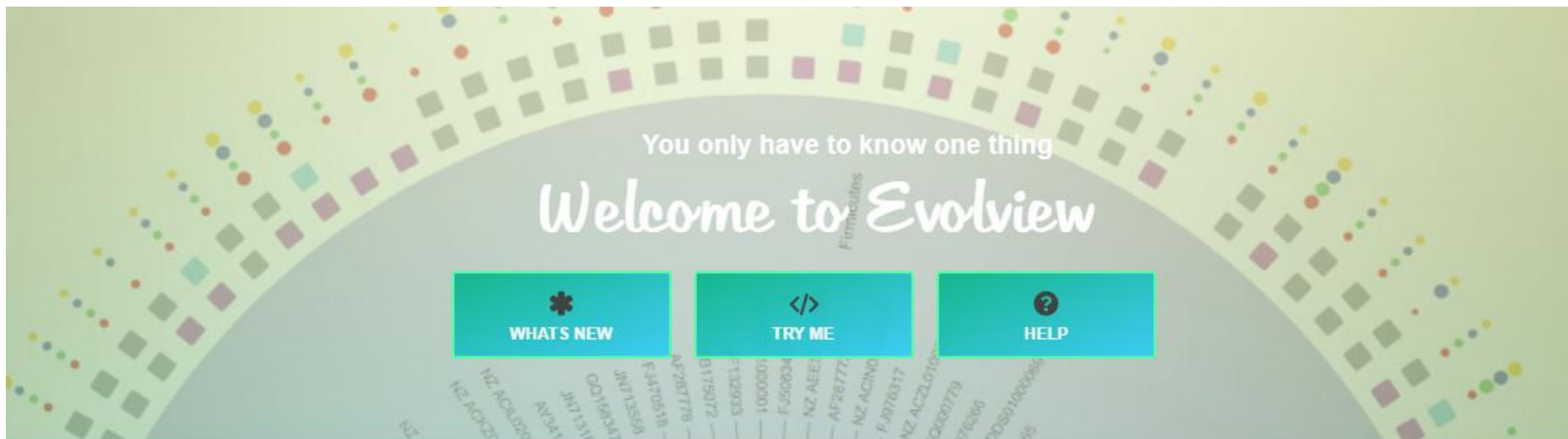
雲南農業大學
Yunnan Agricultural University



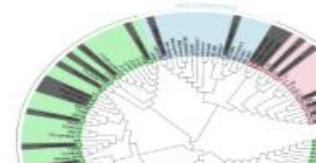
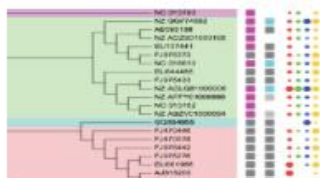
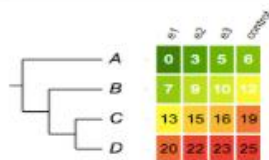
云南省生物大数据重点实验室

生信帮

- 美化和多组数据联合展示
- Evolview网址: <https://www.evolgenius.info/evolview-v3/#login> (提前注册)



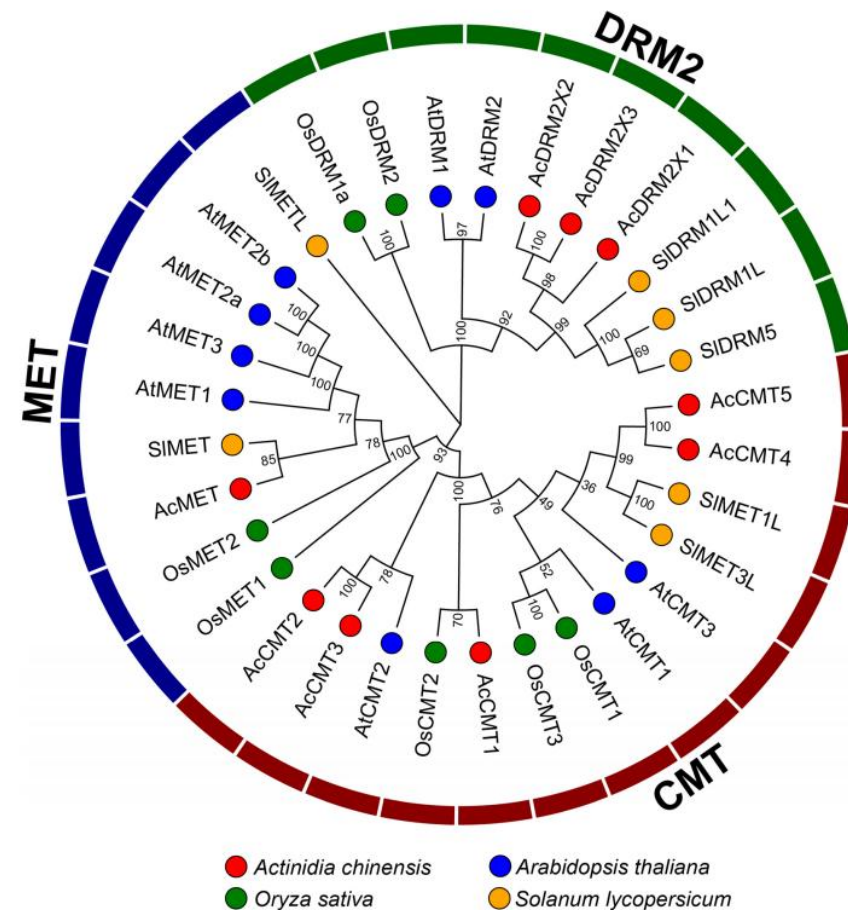
FEATURES



五、进化树分析内容



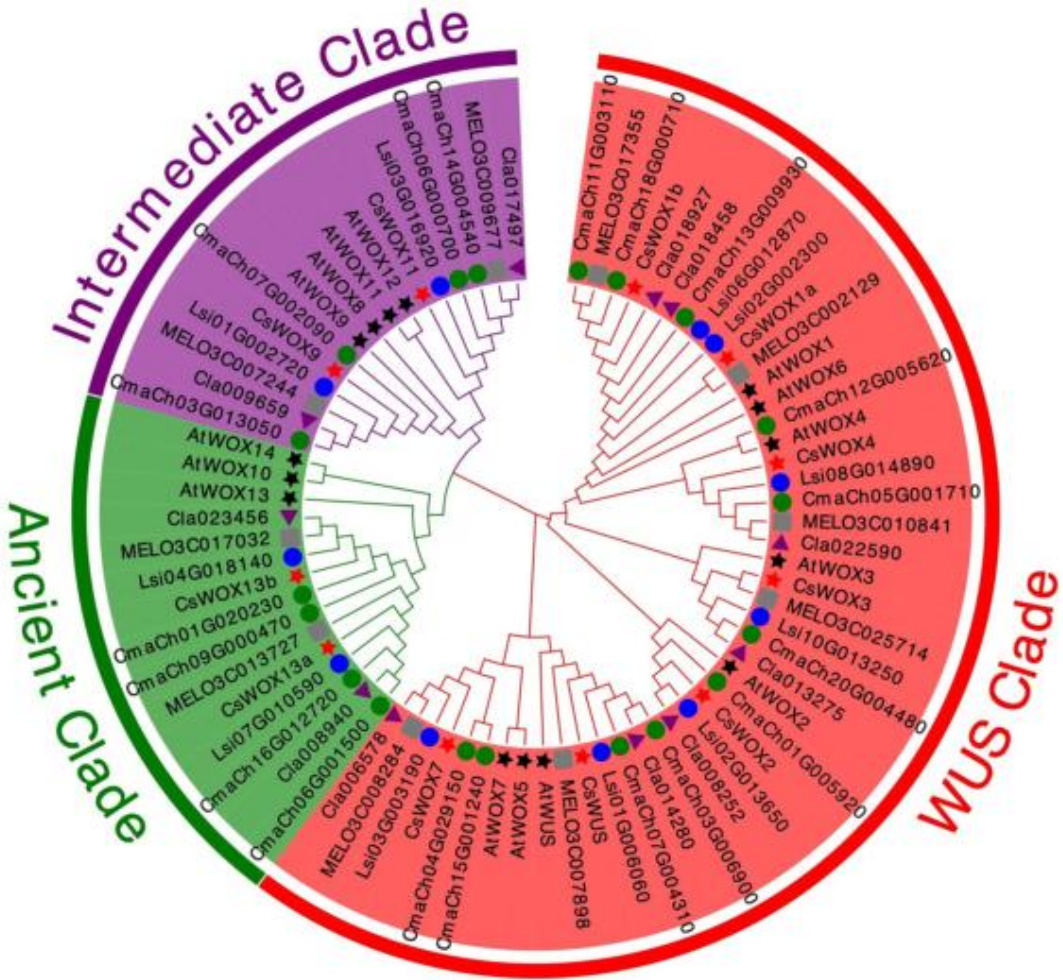
- 通过分析我们首先可以实现对基因家族的进一步分类，如可以将DNA甲基化酶基因家族划分成3个亚组，每个亚组可能发挥不同的功能。
- 通过进化树分析，位于同一分枝的成员之间往往具有更近的亲缘关系，因而功能也更加相近。
- 能比较直观地发现基因家族各成员的扩增情况。



具瓜-WOX基四家族进化树



云南省生物大数据重点实验室



Gu Ran., Song Xiaofei., Liu Xiaofeng., Yan Liying., Zhou Zhaoyang., Zhang Xiaolan.(2020). Genome-wide analysis of CsWOX transcription factor gene family in cucumber (*Cucumis sativus* L.). *Sci Rep*, 10(1), 6216. doi:10.1038/s41598-020-63197-z



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

第四章 共线性分析

1、简介



- 目前主要采用是McscanX，进行基因水平共线性分析的较为常用的软件，使用此软件可以得到共线性区块（block），然后可以进行进一步共线性作图；

- 其主要的意义：

计算Ka/Ks。

寻找复制基因。

WGD分析。

功能基因组学研究。一般共线性基因更倾向于有相同的生物学功能

1	Solyc01g087900.3			
2	Solyc01g087910.3	CEY00_Acc09282		
3	Solyc01g087970.3	CEY00_Acc09281	CEY00_Acc17054	
3	Solyc01g087980.3	CEY00_Acc09280	CEY00_Acc17052	
3	Solyc01g087990.3	CEY00_Acc09279	CEY00_Acc17051	
3	Solyc01g087995.1			
3	Solyc01g088000.3		CEY00_Acc17050	
3	Solyc01g088005.1			
3	Solyc01g088010.1			
3	Solyc01g088020.3	CEY00_Acc09278	CEY00_Acc17049	CEY00_Acc07779
3	Solyc01g088030.3			
3	Solyc01g088040.3			
3	Solyc01g088050.3			CEY00_Acc07780
3	Solyc01g088060.3			
3	Solyc01g088080.3			
3	Solyc01g088090.3			CEY00_Acc07781
2	Solyc01g088100.3			
2	Solyc01g088110.1			
2	Solyc01g088120.1			
2	Solyc01g088130.2			
2	Solyc01g088140.3	CEY00_Acc09274	CEY00_Acc17046	
2	Solyc01g088150.3			
2	Solyc01g088160.3	CEY00_Acc09272	CEY00_Acc17044	
2	Solyc01g088170.3			



生信帮



Yves, Van De Peer, , S. Maere , and A. Meyer . "The evolutionary significance of ancient genome duplications." *Nature Reviews Genetics* 10.10(2009):725-732.

2、共线性分析步骤



- BLAST。寻找序列相似的基因对，确认哪些基因可以比对上。
- GFF文件。确定每个基因在染色体上的位置。
- 共线性区块寻找。 McscanX，寻找在染色体上至少有五对基因在染色体不同位置**连续**比对上。

3、物种内共线性分析



背景是两个物种的共线性，标红颜色是所研究基因家族共线性

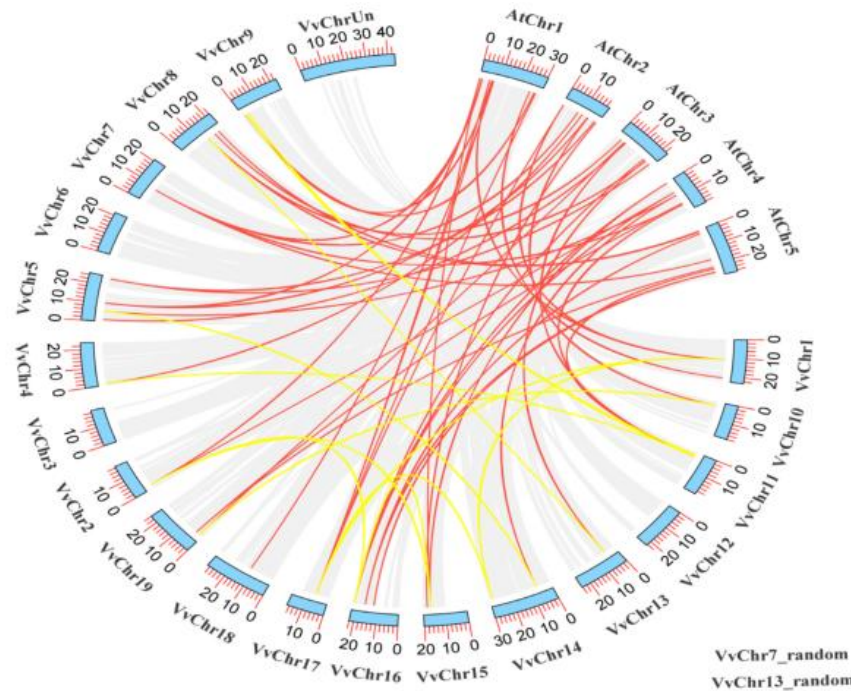
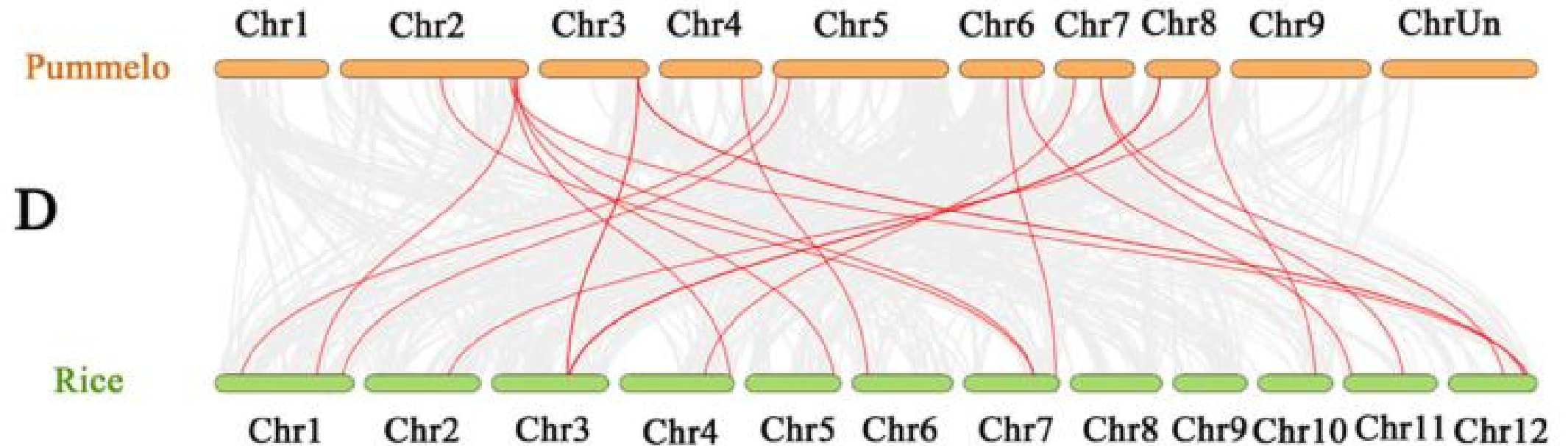


Figure 2. The collinear correlation of *VvIQDs* between grapevines and *Arabidopsis* (indicated with red colored lines) and within grapevine (indicated with yellow colored lines). The two random genes in grapevine genes are also given at the bottom.

4、物种间共线性分析



背景是两个物种的共线性，标红颜色是所研究基因家族共线性



Liu Zheng., Ge Xiao-Xia., Wu Xiao-Meng., Xu Qiang., Atkinson Ross G., Guo Wen-Wu.(2020). Genome-wide analysis of the citrus B3 superfamily and their association with somatic embryogenesis. BMC Genomics, 21(1), 305. doi:10.1186/s12864-020-6715-9

5、Ks计算



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

计算步骤：

- 1、获取复制基因对。包括自身共线性基因对以及与去其他要分析物种的共线性基因对。
- 2、对复制基因对进行mafft比对（蛋白质序列）。
- 3、反转CDS。
- 4、使用KaKs_Calculator中计算Ka/Ks。

5、Ks计算



Table 1. The Ka/Ks values of *Theobroma cacao* tandem repeat sequences.

Tandem Repeat Sequence	Ka	Ks	Ka/Ks
TcNAC048/TcNAC047	0.069652	0.204593	0.34044
TcNAC055/TcNAC057	0.022392	0.046771	0.47876
TcNAC055/TcNAC056	0.031369	0.066627	0.47082
TcNAC056/TcNAC057	0.020842	0.045468	0.45839
TcNAC047/TcNAC046	0.071078	0.056378	1.26074
TcNAC085/TcNAC084	0.057197	0.128216	0.44609
TcNAC085/TcNAC086	0.051726	0.110318	0.46888
TcNAC003/TcNAC004	0.028539	0.129133	0.22101
TcNAC100/TcNAC101	0.014445	0.046379	0.33115
TcNAC084/TcNAC086	0.053422	0.090511	0.59023
TcNAC063/TcNAC005	0.111896	0.215804	0.51851
TcNAC048/TcNAC046	0.077084	0.210968	0.36538

Shen Shiya., Zhang Qianru., Shi Yu., Sun Zhenmei., Zhang Qianqian., Hou Sijia., Wu Rongling., Jiang Libo., Zhao Xiyang., Guo Yunqian.(2019).
Theobroma cacaoGenome-Wide Analysis of the NAC Domain Transcription Factor Gene Family in . Genes (Basel), 11(1), undefined.
doi:10.3390/genes11010035



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室

生信帮，答你所问