

比较基因组学基础理论

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30日



- ◆ 基因组基本特征
- ◆ 基因家族聚类分析
- ◆ 进化树构建
- ◆ 基因家族扩张与收缩
- ◆ 正选择分析
- ◆ 共线性分析
- ◆ 全基因组复制分析



1、四倍体棉基因组特征

A和D两套
亚基因组

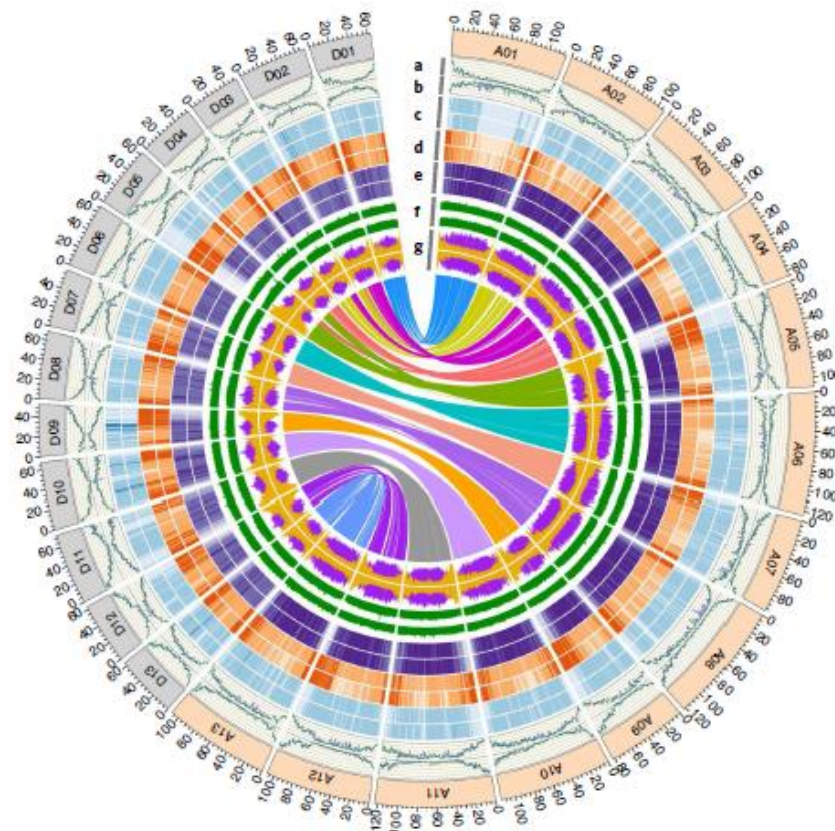


Fig. 1 | Chromosomal features of *G. hirsutum* and *G. barbadense* genomes with integration of genetics and epigenetics data. **a**, Gene density in each chromosome. **b**, Transposable element (TE) content in each chromosome. For **a** and **b**, green lines represent data in *G. hirsutum* and blue lines represent data in *G. barbadense*. **c**, SNP density between *G. hirsutum* and *G. barbadense*. **d**, InDel density between *G. hirsutum* and *G. barbadense*. **e**, 5mC DNA methylation levels. **f**, 6mA DNA methylation levels. **g**, Histone modification levels of H3K4me3 (yellow histogram) and H3K9me2 (purple histogram). For **c-g**, the outer tracks show data in *G. hirsutum* and inner tracks show data in *G. barbadense*. All these data are shown in 1 Mb windows sliding 200 kb. The inner lines show syntenic blocks in homeologous chromosomes between the A and D subgenomes.

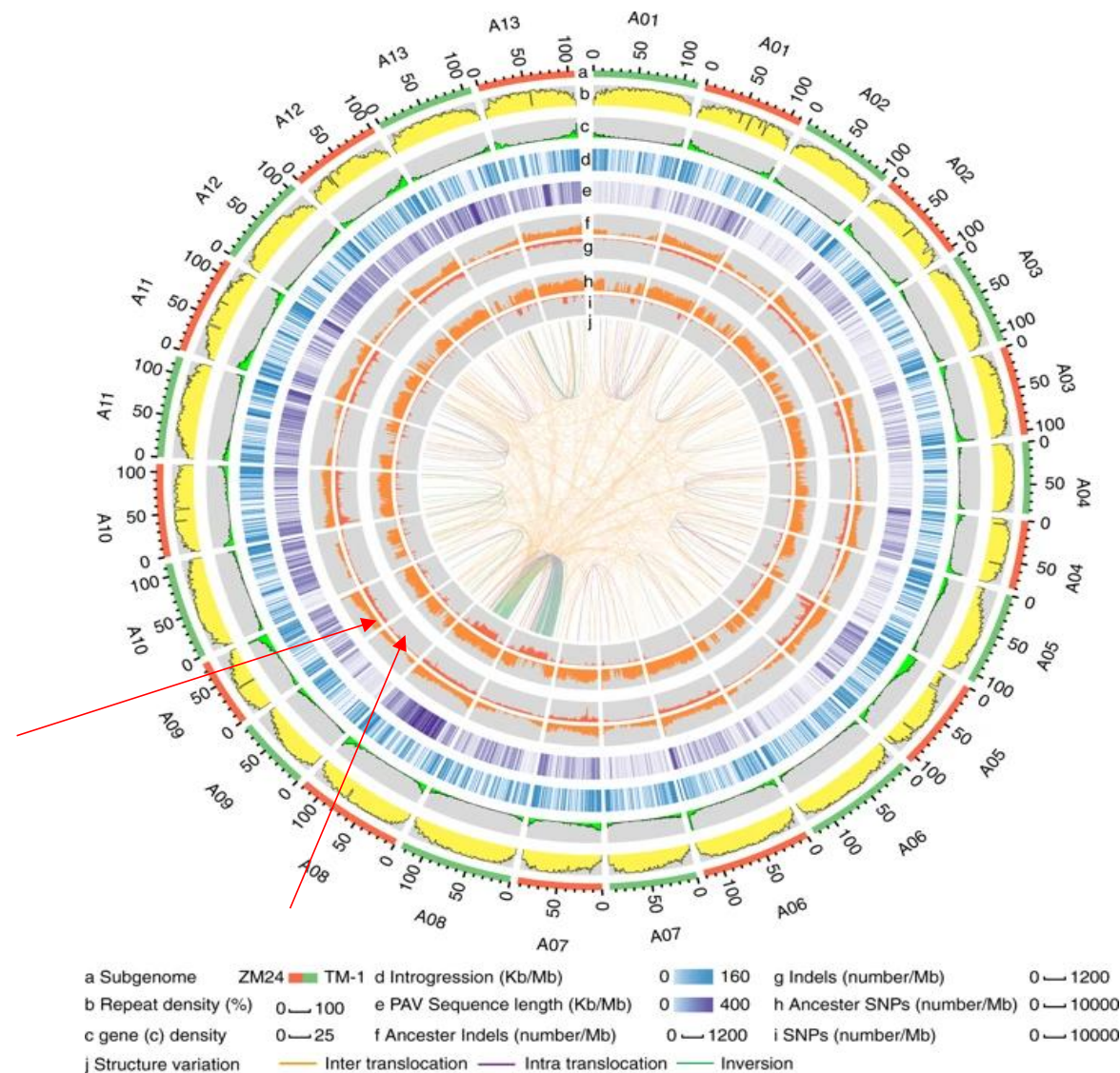
展示基因组的基本特征！

- a、基因密度
 - b: TE密度
 - c: SNP密度
 - d: Indel密度
 - e: 5mC甲基化
 - f: 6mA甲基化
 - g: H3K4me3组蛋白甲基化密度
- 最内测连线为共线性

摘自：Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li, J., ... & Ye, Z. (2019). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nature genetics*, 51(2), 224.

2、两个陆地棉基因特征展示

最内圈结构变异，主要集中在A08中,SNP 密度也很高。



摘自: Yang, Z., Ge, X., Yang, Z., Qin, W., Sun, G., Wang, Z., ... & Lu, L. (2019). Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. Nature communications, 10(1), 2989.



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

第一部分：基因家族聚类分析

一、基因家族简介



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

什么是基因家族？

MYB , bZIP, FAD, WRKY 。 。 。

一、基因家族简介



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

- 定义：基因家族（gene family），来源于多个物种中的，含有两个或更多的拷贝一组同源基因。他们由同一个祖先的一个基因，通过基因复制和物种分歧而产生。
- 研究结果：通过基因家族聚类分析可以得到物种间共有基因家族和共有基因，各物种特有基因家族和特有基因及单拷贝基因家族和单拷贝基因。
- 应用：构建进化树；基因家族扩张与收缩分析。

一、基因家族简介



同源基因：

- 直系同源基因（Orthologues）：两个物种共同祖先的一个基因经过演化产生的分布在这两个物种中的一对基因，是物种形成事件的结果。
- 旁系同源基因（Paralogs）：一个物种内部，通过基因复制产生的基因，是基因复制事件的结果。

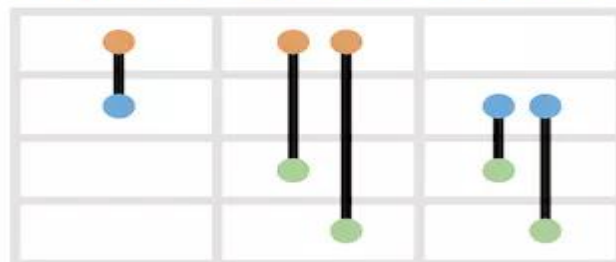
A. Orthogroup



Group of genes descended from single gene in LCA of group of species

B. Orthologues

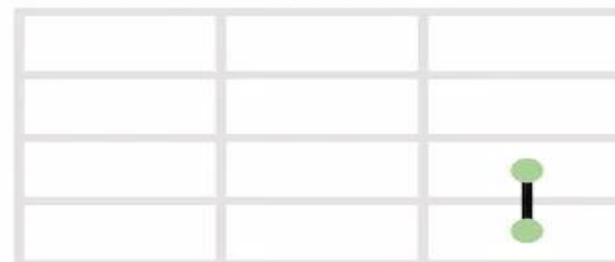
Hu-Mo Hu-Ch Mo-Ch



Pairs of genes descended from single gene in LCA of pair of species

C. Paralogues

Hu-Hu Mo-Mo Ch-Ch



Pairs of genes descended from gene duplication event

二、分析软件介绍



- OrthoMCL，这是2003年发表的一个工具，但在2013年之后其就不在更新了，并且分析需要mysql，多个物种分析占用内存较大，限制了好多人的使用。
- OrthoFinder：2015年开发的一个工具，相比OrthoMCL快速、准确、易于安装，不需要mysql。另外它具有分析结果的自动化统计，避免自己写脚本统计运行的结果。

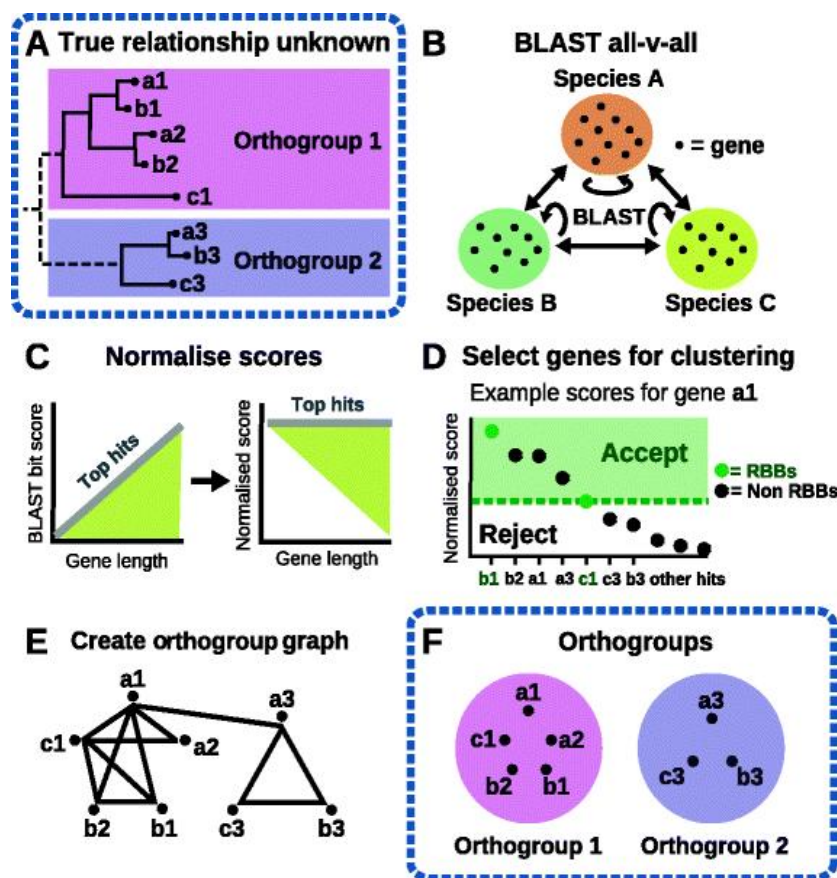
参考文献：

Li, & L. (2003). Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9), 2178-2189.
Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157

二、分析软件介绍



➤ OrthoFinder算法



B、BLAST all-vs-all比对。使用BLASTP进行比对，寻找潜在的同源基因。自身比对和相互比对。

C、基于基因长度和系统发育距离对BLAST bit得分进行标准化。

D、从标准化的BLAST评分中选择假定的同源基因对。

E、直系同源组图的构造，基因是图中的节点，成对的基因通过边缘线连接，边缘权重由归一化比对得分给出

F、用MCL程序对基因进行聚类，划分直系同源组

二、分析软件介绍



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

➤ OrthoFinder安装

1、快捷安装。

有管理员权限，OrthoFinder可以通过conda安装，使用下面命令：

```
conda create -n orthofinder orthofinder=XXX
```

2、无管理员权限，单独安装。单独安装它需要它的三个依赖工具 BLAST+，下载地址如下：

BLAST+: <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST>

OrthoFinder: <https://github.com/davidemms/OrthoFinder>

三、操作步骤



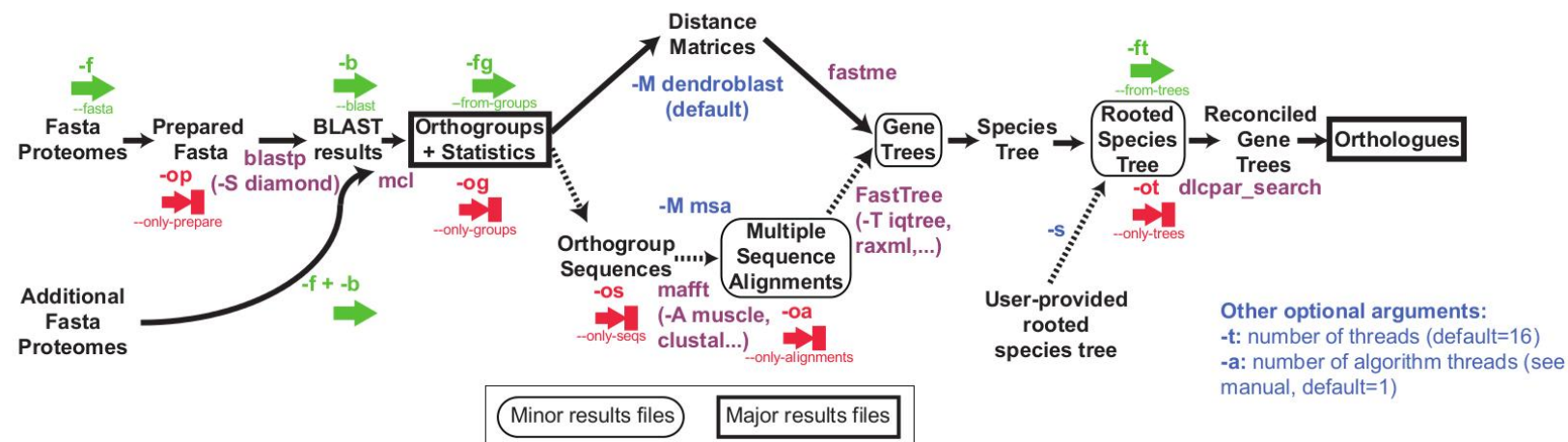
OrthoFinder整体运行控制流程

开始

Controlling the OrthoFinder Analysis

If you just want to run a full analysis automatically,
use: 'orthofinder -f fasta_dir'

Control where OrthoFinder starts
Command-line switch takes directory containing files as argument.
Control where OrthoFinder stops (e.g. -ot = 'only up to & including trees')
Optional arguments
Programs required for each step (alternatives)



结束

三、操作步骤



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

1、数据下载

- 下载常用的数据库：

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

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

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

- 只需一组FASTA格式的蛋白质序列文件。文件末尾必须以.fa或者.fasta结尾！
- 要求不能含有可变剪接，需要提取最长转录本对应的蛋白质序列。
- 对于多倍体必须拆成亚基因组对应的蛋白质序列去做。如陆地棉为AADD，共获得两套亚基因组A和D，所以要把该基因组所有的蛋白分成A和D这两个物种去做。

三、操作步骤



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

2、运行orthofinder

➤ 从头开始: ***orthofinder -f orthofinder/ExampleDataset/ -S blast -t 10 -og***

主要参数介绍:

-f:所有蛋白质序列所在的文件夹

-t: BLASTP运行的CPU

-S:选择比对工具，我们一般选择blast

-op:准备完BLAST运行需要的格式后停止，并输出BLAST命令。

-og:基因家族聚类完成后停止运行



四、结果介绍

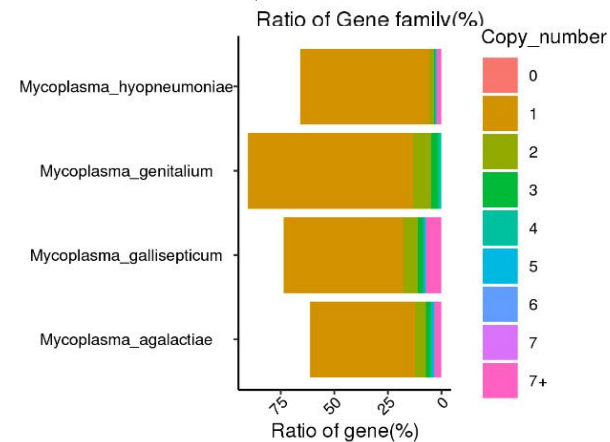
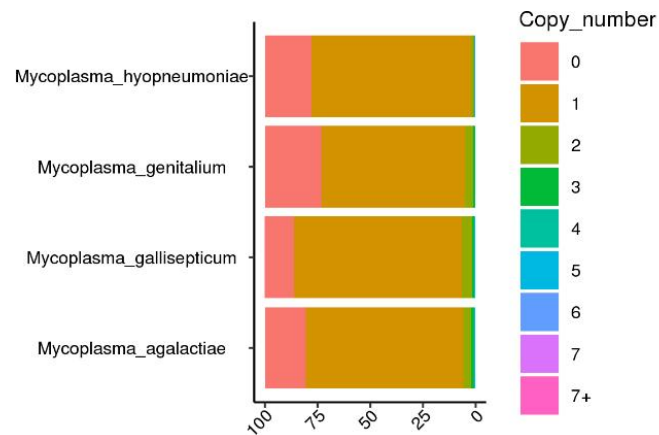
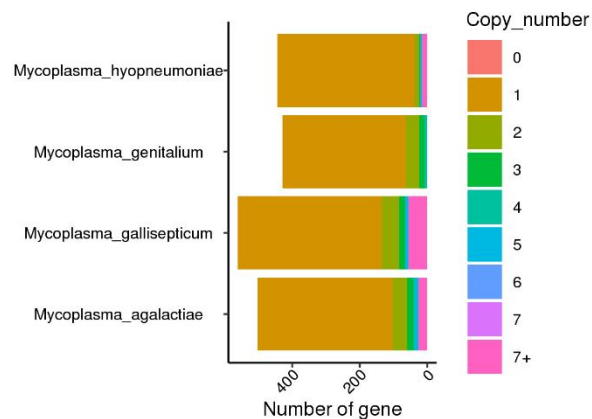
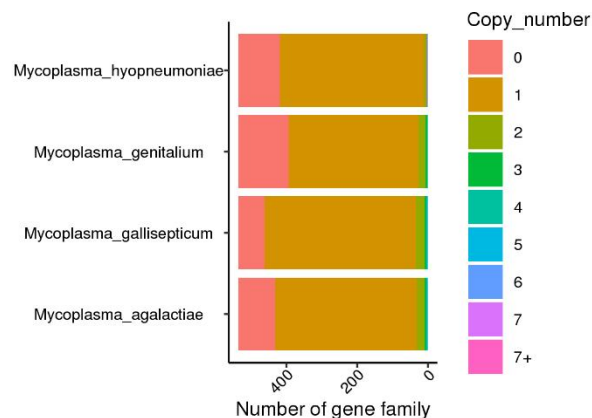
基因家族聚类可能的情况

	A	B	C	D	Species Number	Type
OG0000000	1	1	1	1	4	Single Copy
OG0000001	3	1	3	4	4	All species
OG0000002	2	0	0	0	1	Species-specific orthogroups
OG0000003	1	0	2	0	2	Other
OG0000004	2	0	1	1	3	Other
OG0000005	0	0	0	1	1	Unassigned genes

五、可视化



1、每个物种基因家族分布：Statistics_PerSpecies.tsv



注：

Mycoplasma_agalactiae: 无乳支原体

Mycoplasma_gallisepticum: 鸡败血症支原体

Mycoplasma_genitalium: 生殖支原体

Mycoplasma_hyopneumoniae: 猪肺炎支原体

五、可视化



雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

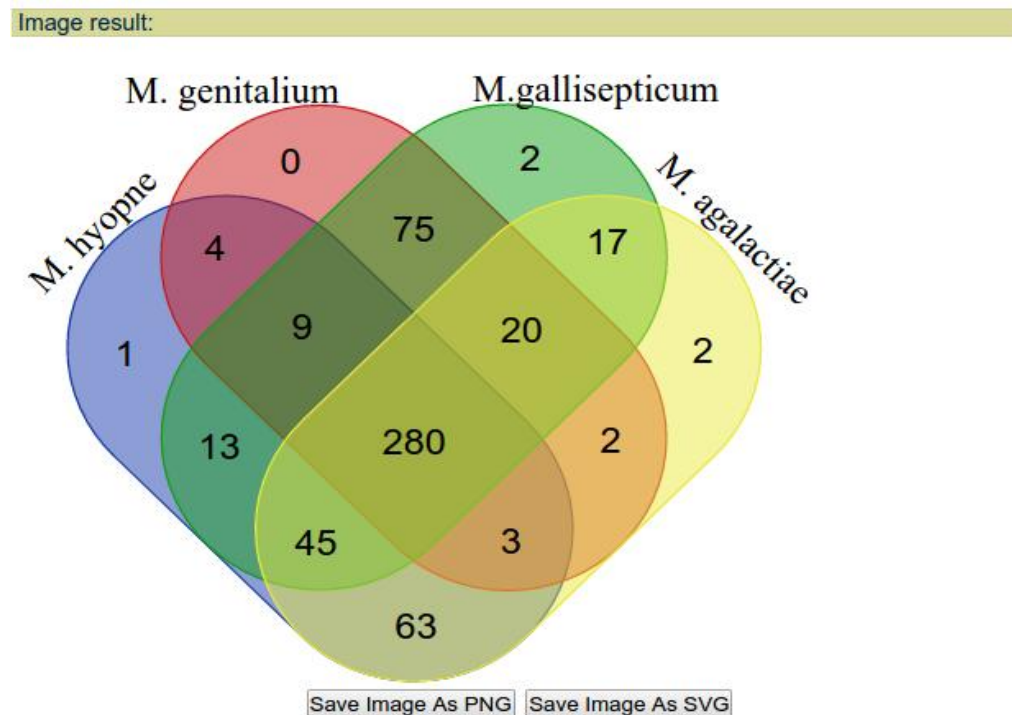
生信帮

2、维恩图

➤ 网址:

<http://bioinformatics.psb.ugent.be/webtools/Venn/>

➤ 不要选择过多，选择3~6个即可。





雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



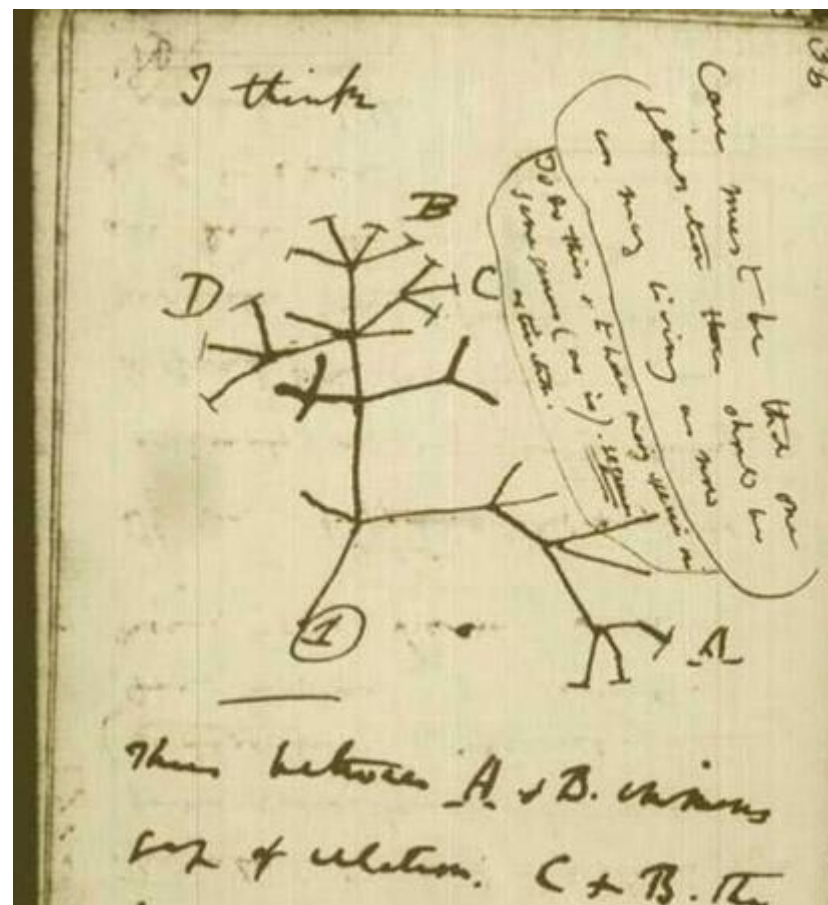
生信帮

第二部分：进化树构建

一、背景简介



- 达尔文在《物种起源》中提到了系统发生的概念。认为地球上的所有物种都是由**同一祖先**经过各种遗传事件进化而来，其终极目标是绘制各物种演化的先后顺序，并用一棵生命之树（即系统进化树）来表示。
- 系统发育树最初构建是通过**表型特征**比对，化石研究等来进行的，但生物性状量化困难，化石数据不易获得，导致系统发育树的构建相当缓慢！



一、背景简介



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

- 一代基因测序技术出现，可以实现少数基因序列的测定，因此可以利用物种间的同源基因序列，采用一定的数学模型，构建**基因树**，以此来推测物种间系统发育关系。如叶绿体的 **rbcL**基因，远缘间及科级以上分类群的系统关系的研究
- 二代和三代高通量测序技术的出现和成本降低，大量物种基因组序列的破译，使得我们我们可以在全基因组层面，进行进化树构建，描述物种间的进化规律变得更加准确。

一、背景简介



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

研究的意义：

- 确立物种之间的亲缘关系。
- 确定物种的进化历史及分歧时间。
- 进化树上关系较近的基因可能具有相似的功能，据此可以挖掘相关功能的基因（基因树）。

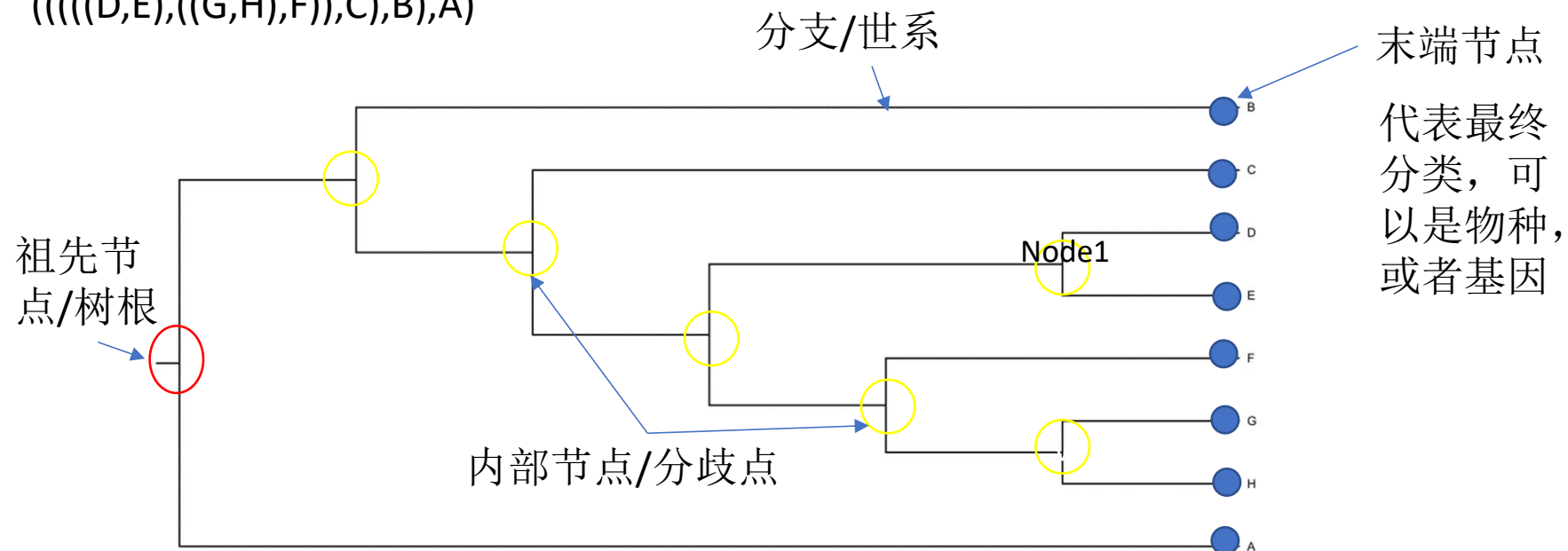
二、进化树基本术语



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

二、进化树基本术语



雲南農業大學
Yunnan Agricultural University

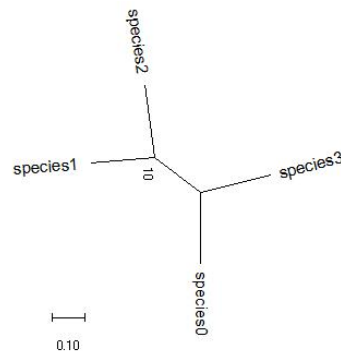
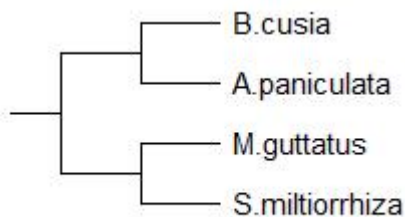


云南省生物大数据重点实验室

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)。
- 外群选择问题：选择一个或多个已知与分析序列关系较远（不能过远）的序列作为外群。外群可以辅助定根；外群序列与其它内群序列之间的差异，必须显著大于其它内群序列之间的差异



二、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、基因树和物种树

- 基于单个同源基因差异构建的系统发育树称为基因树（gene tree），代表的仅仅是单个基因的进化历史。
- 物种树一般是通过比较物种间所有的单拷贝基因数据的分析结果而产生，代表物种的进化历史。借助于化石时间矫正，可以得到有分化时间的物种树。
- 两者存在差异：
 - 来自两不同物种的两个基因的分化（多态性）时间可能早于物种的分化；
 - 基因树的拓扑结构可能与物种树不完全一致（比较近的物种间）
- 由于物种的进化历史不可能再现,所以不可能重建绝对完整的历史,同样也不可能获取绝对的物种树。但是通过**多基因**,大量**DNA**序列的正确分析,可以最大限度地缩小基因树与物种树间的差别。在这种情况下获得的系统树可被接受为物种树

4、Bootstrap值

为进化树可靠性评估指标。对于构建物种树，一般采用500次就可以了，单个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。

二、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

5、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)

三、进化树的构建方法



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			慢慢

三、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



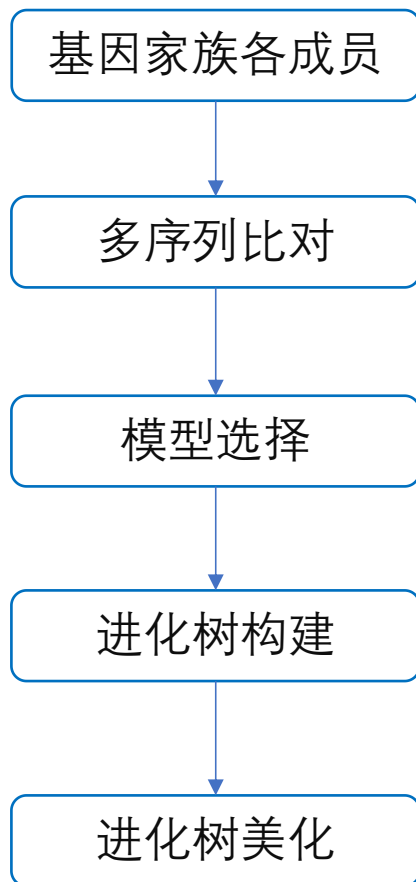
生信帮

- 目前做物种树的文章中应用比较多，认可度比较高的方法是**ML方法**，**IQ-TREE**软件也逐步被人所接受。在进化模型确定的情况下，**ML法**是与进化事实最吻合的建树算法。
- 对于单个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。
- 分析时间：距离法 < 最大似然法。

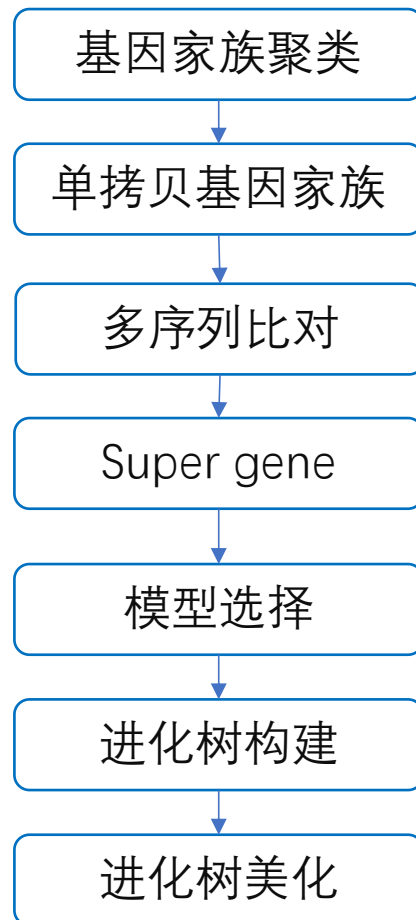
四、进化树构建步骤



单个基因家族基因树构建 (MEGA)



物种系统发育进化树构建



五、单个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

1、研究意义

- 主要针对比较关注的单个的与某个功能相关的基因家族或者基因家族发生显著扩张与收缩的基因家族中的所有成员的进行精准的进化分析。
- 通过分析我们首先可以实现对基因家族的进一步分类，如对四倍体栽培花生及两个祖先二倍体花生的ARF基因家族的进化分析，可以将其划分成9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。

五、单个基因家族进化树的构建

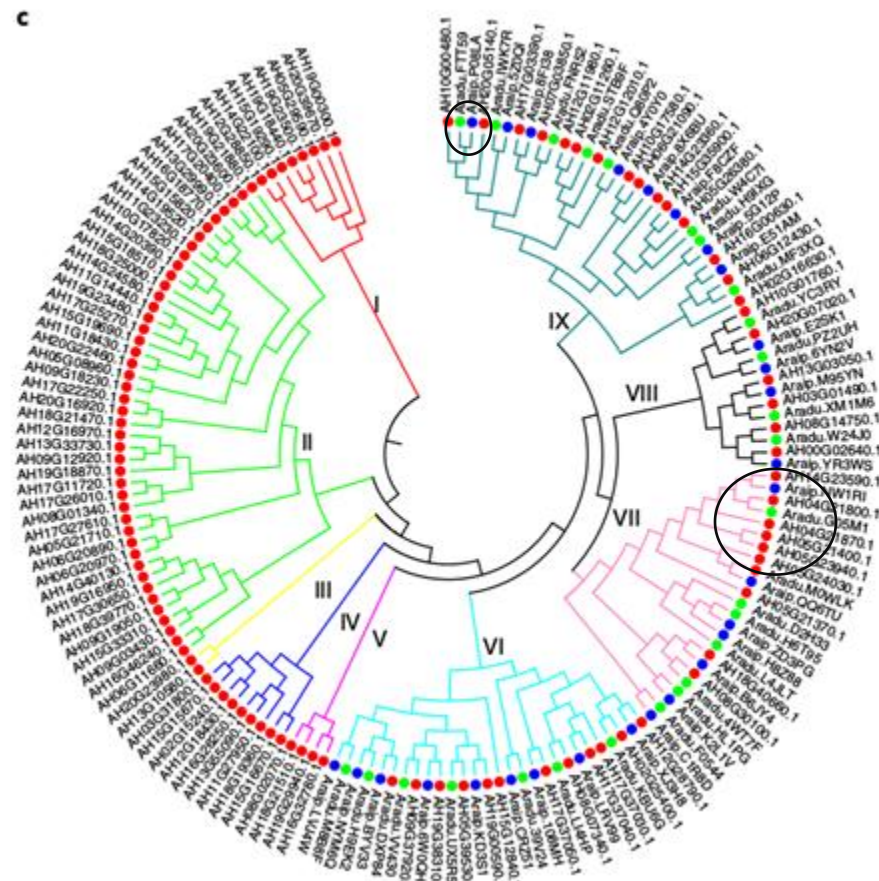


1、研究意义

四倍体栽培花生及两个祖先二倍体花生的ARF基因家族的进化分析

- 分类
- 亲缘关系
- 基因复制

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.)



五、单个基因家族进化树的构建



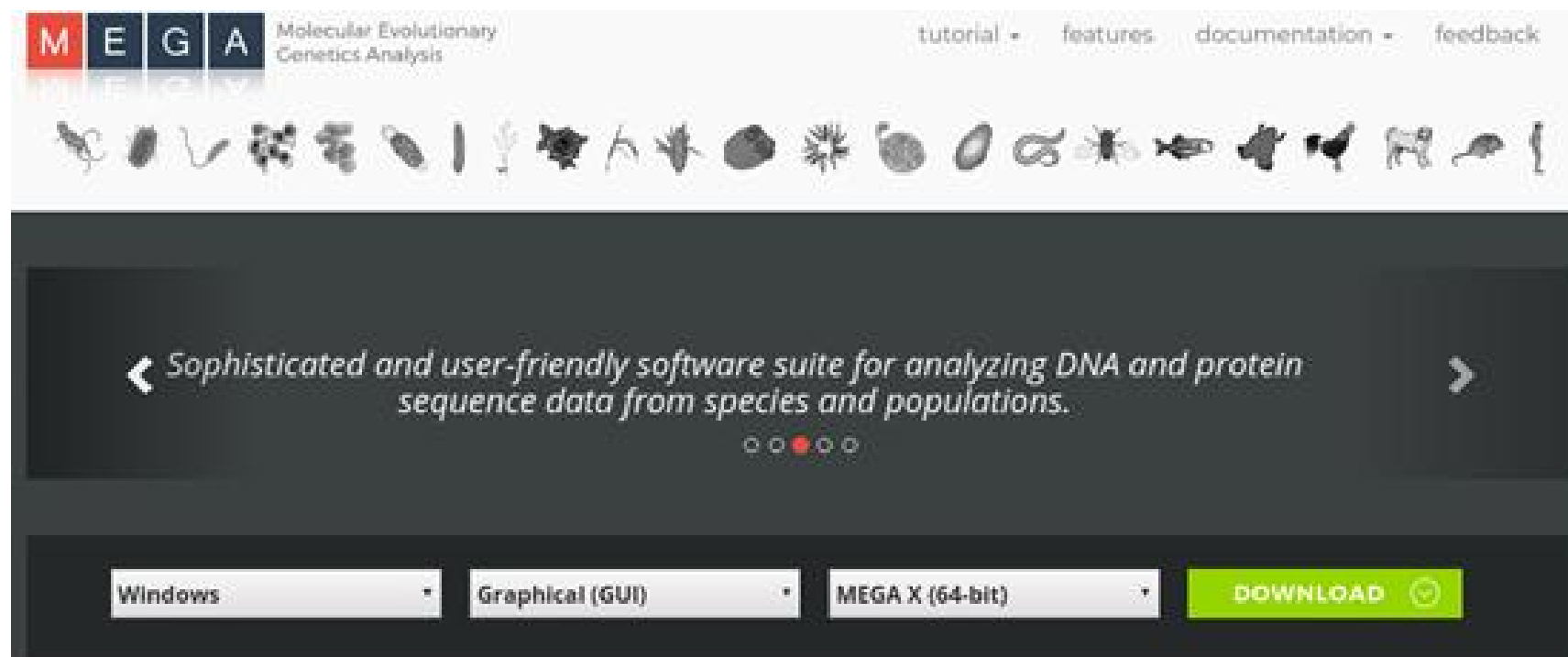
雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

2、分析流程



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.

六、系统发育进化树构建步骤



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

1、利用单拷贝基因家族序列进行多序列比对。

在基因家族聚类的结果中提取单拷贝基因家族的基因，然后利用mafft进行多序列比对。

✓ 为啥选用单拷贝？

- 单拷贝基因进化方向单一。
- 多拷贝基因：一般是物种分歧后来通过基因复制产生的，不能够判定哪一个成员是从最初的同一个祖先继承来的。

六、系统发育进化树构建步骤



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

2、将比对结果转换为密码子比对结果

mafft比对之后需要将比对好的蛋白(protein alignment)转换为codon alignment，主要使用pal2nal程序。

3、比对结果优化（Alignment refinement）

主要是针对比对产生的GAP的处理，采用了Gblocks软件。

六、系统发育进化树构建步骤



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

4、连接形成Super gene。

将每个物种的比对结果首尾相连连成super gene，同时整理成Phylip格式

5、替代模型的选择。

进行最大似然（ML）法构建进化树不可获缺的一个重要阶段。目前常用IQ-Tree软件中的ModelFinder，其速度和性能及灵活性均优于之前的 jModelTest/ProtTest等工具

参考文献：Kalyaanamoorthy S, et al, 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14:587-589.

六、系统发育进化树构建步骤



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

6、进化树构建。

IQ-TREE

7、进化树美化。

MEGA和Evolview软件

六、系统发育进化树构建步骤



8、构建有分化时间的进化树

物种出现的大致时间表：

中生代	白垩纪	145.5 ± 4	恐龙繁荣后灭绝
			白垩纪 - 第三纪灭绝事件，地球上45%生物灭绝
	侏罗纪	199.6 ± 0.6	有胎盘的哺乳动物出现
			有袋类哺乳动物出现
			鸟类出现
			裸子植物繁荣
			被子植物出现
	三叠纪	251.0 ± 0.7	恐龙出现 卵生哺乳动物出现

古生代	二叠纪	299.0 ± 0.8	二叠纪灭绝事件，地球上95%生物灭绝 盘古大陆形成 昆虫繁荣
	石炭纪	359.2 ± 2.5	爬行动物出现 煤炭森林
			裸子植物出现
			鱼类繁荣
	泥盆纪	416.0 ± 2.8	两栖动物出现 昆虫出现
			种子植物出现
			石松类植物出现
	志留纪	443.7 ± 1.5	陆生裸蕨植物出现
	奥陶纪	488.3 ± 1.7	鱼类出现；海生藻类繁盛
	寒武纪	542.0 ± 1.0	寒武纪生命大爆炸

六、系统发育进化树构建步骤



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

8、构建有分化时间的进化树。

使用**PAML**软件包中的**mcmctree** 进行分歧时间估计，同时利用**timetree**(<http://www.timetree.org/>)网站提供的化石时间，对分歧时间进行校正。一般要运行两次，看两次的一致性来确定结果的好坏。

计算步骤：

- 获取最近物种与最远物种之间的化石时间：Panicum miliaceum (Pm)-foxtail millet (Si): (9.4 - 23.7 MYA);

Panicum miliaceum (Pm)-rice(Os): (42 - 52 MYA)

- 准备树文件：((((Si,Pm)')>9.4<23.7',(Sb,Zm)),Bd),Os)')>42<52';

- 估计分歧时间所需的参数gradient 和Hessian

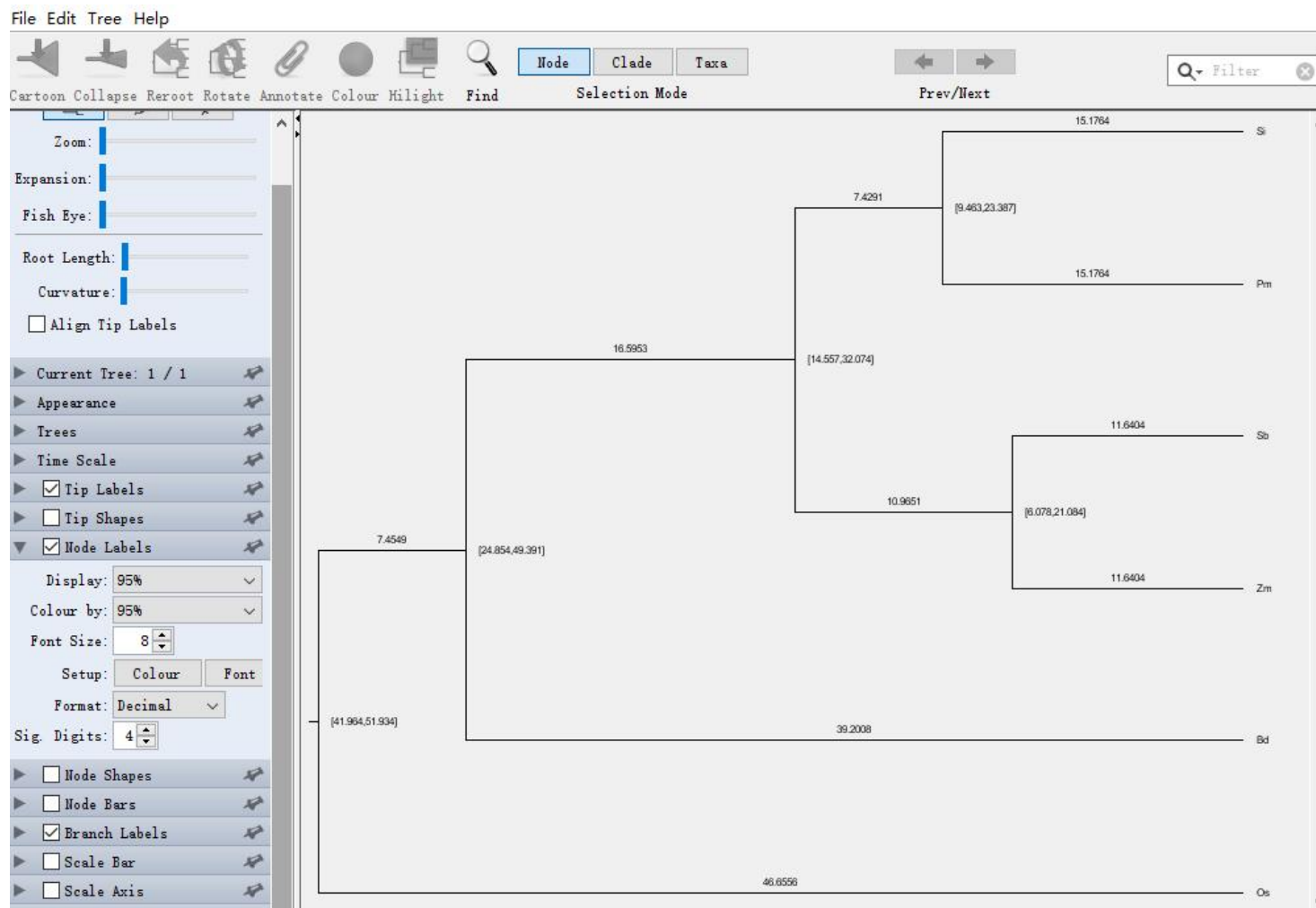
- 最大似然法估计分歧时间

- FigTree.tre文件

六、系统发育进化树构建步骤



8、构建有分化时间的进化树。





第三部分：基因家族扩张与收缩分析

一、基因家族扩张与收缩简介



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

- 传统意义上，分析哪个基因家族成员数目收缩和扩张，可以**直接比较**两个近缘物种对应基因家族成员的数目多少即可。但直接比较的只能得到相对的比较结果，无法确定扩张与收缩情况。而真实情况应该是**与祖先比较**，阐明物种在进化过程中发生的基因家族成员的扩张与收缩情况。
- `café` (Computational Analysis of gene Family Evolution) 软件，能够利用带有**分歧时间的进化树**，通过出生死亡率模型估计每个分枝的祖先的基因家族个数，从而预测出研究物种的基因家族相对于祖先的收缩和扩张情况。这样在共同的条件下，得到的结果才更有意义。
- 扩张的基因家族对于物种特殊的功能有折至关重要的意义。

二、基因家族扩张与收缩分析步骤



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

根据基因家族的聚类分析结果，使用CAFÉ软件进行基因家族扩张和收缩分析；需要下面两个输入：

- 相同枝长的带有分化时间的进化树（自己根据树制备）：

(((Si:14,Pm:14):7,(Sb:8,Zm:8):13):16,Bd:37):9,Os:46)

- 基因家族聚类分析结果：Orthogroups.GeneCout.tsv。

过滤高拷贝数基因家族并转换为需要的格式文件：[cafe_gene_family_format.txt](#)。

运行命令：cafe cafe.sh

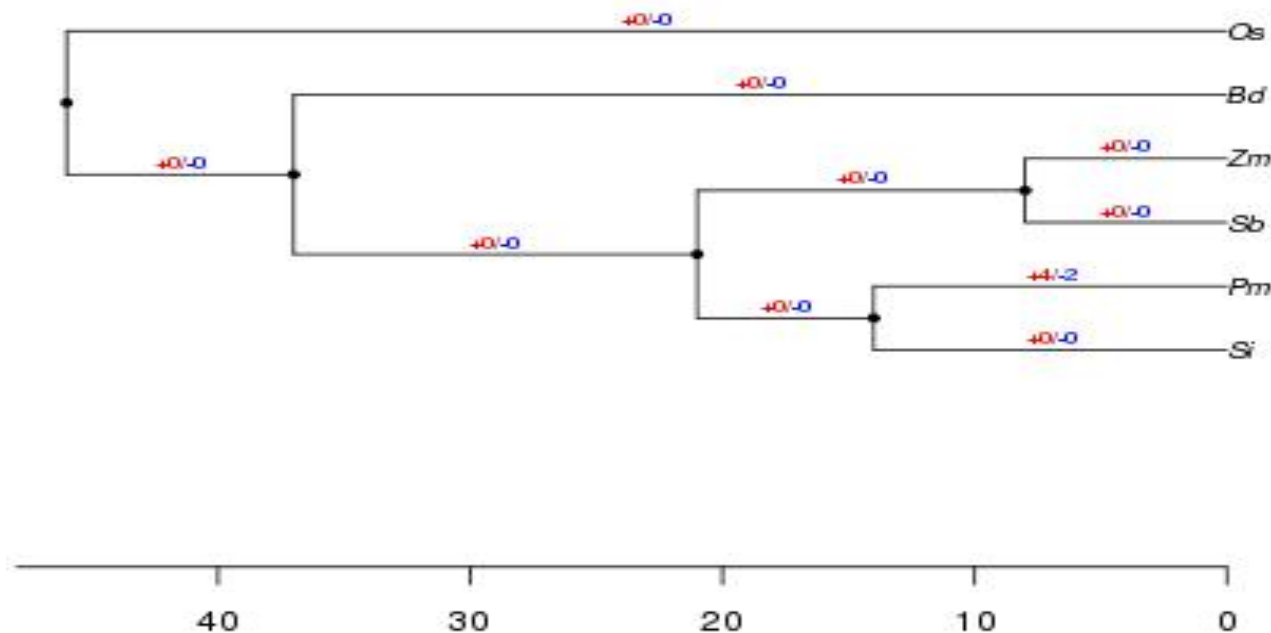
命名为cafe.sh，其包含下面内容：

```
date
version
load -i cafe_gene_family_format.txt -t 1 -l cafe.log -p 0.01
tree (((Si:14,Pm:14):7,(Sb:8,Zm:8):13):16,Bd:37):9,Os:46)
lambda -s
report result
```

三、基因家族扩张与收缩分析结果



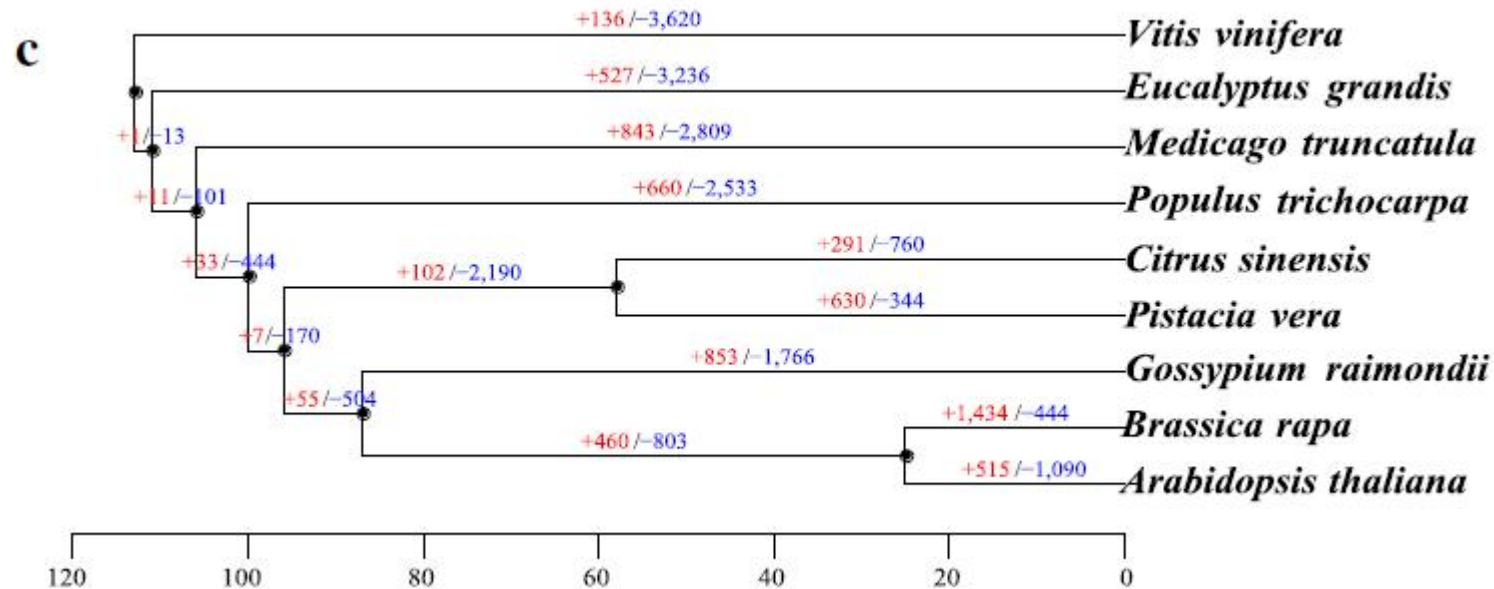
```
1 Tree:(((Si:14,Pm:14):7,(Sb:8,Zm:8):13):16,Bd:37):9,Os:46)
2 Lambda: 0.000251445
3 # IDs of nodes:(((Si<0>,Pm<2><1>,(Sb<4>,Zm<6><5><3>,Bd<8><7>,Os<10><9>
4 # Output format for: 'Average Expansion', 'Expansions', 'No Change', 'Contractions', and 'Branch-specific P-values' = (node ID, node ID): (0,2) (1,5) (4,6) (
5 # Output format for 'Branch cutting P-values' and 'Likelihood Ratio Test': (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
6 Average Expansion: (0,0.030303) (0,0) (0,0) (0,0) (0,0)
7 Expansion : (0,4) (0,0) (0,0) (0,0) (0,0)
8 nRemain : (99,92) (99,99) (99,99) (99,99) (99,99)
9 nDecrease : (0,3) (0,0) (0,0) (0,0) (0,0)
10 'ID' 'Newick' 'Family-wide P-value' 'Viterbi P-values' 'cut P-value' 'Likelihood Ratio'
1 OG00000000 (((Si_3:14,Pm_3:14)_3:7,(Sb_3:8,Zm_3:8)_3:13):16,Bd_3:37):9,Os_3:46)_3 0.6015 ((-,-),(-,-),(-,-),(-,-),(-,-))
2 OG00000001 (((Si_2:14,Pm_2:14)_2:7,(Sb_2:8,Zm_2:8)_2:13):16,Bd_2:37):2:9,Os_2:46)_2 0.573 ((-,-),(-,-),(-,-),(-,-),(-,-))
3 OG00000002 (((Si_2:14,Pm_2:14)_2:7,(Sb_2:8,Zm_2:8)_2:13):16,Bd_2:37):2:9,Os_2:46)_2 0.573 ((-,-),(-,-),(-,-),(-,-),(-,-))
4 OG00000003 (((Si_2:14,Pm_2:14)_2:7,(Sb_2:8,Zm_2:8)_2:13):16,Bd_2:37):2:9,Os_2:46)_2 0.573 ((-,-),(-,-),(-,-),(-,-),(-,-))
5 OG00000004 (((Si_2:14,Pm_1:14)_2:7,(Sb_2:8,Zm_2:8)_2:13):16,Bd_2:37):2:9,Os_2:46)_2 0.0725 ((-,-),(-,-),(-,-),(-,-),(-,-))
6 OG00000005 (((Si_1:14,Pm_4:14)_1:7,(Sb_1:8,Zm_1:8)_1:13):16,Bd_1:37):1:9,Os_1:46)_1 0 ((0.503502,2.15824e-08),(0.501755,0.503253),(0.502006,
```



二、构建进化树和基因家族扩张与收缩分析



4、案例-开心果基因组基因家族扩张情况



摘自：Zeng, L., Tu, X. L., Dai, H., Han, F. M., Lu, B. S., Wang, M. S., ... & Ji, L. L. (2019). Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome biology*, 20(1), 79.



第四部分：正选择分析

一、正选择分析简介



1、概念。

通过自然选择某些有利的基因（**突变产生新功能**）被固定下来。
受到了正选择的基因，往往在该物种对环境的适应性方面有重要作用，全基因组测序能充分挖掘。

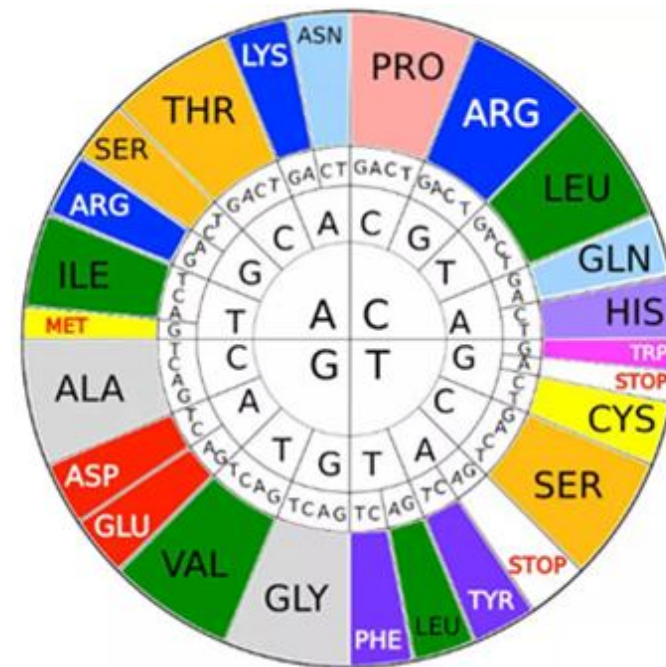
K_a =非同义替换的SNP位点数/非同义替换位点数

K_s =同义替换的SNP位点数/同义替换位点数

$K_a > K_s \rightarrow K_a/K_s > 1$ ，基因受正选择(positive selection)

$K_a = K_s \rightarrow K_a/K_s = 1$ ，基因中性进化(neutral evolution)

$K_a < K_s \rightarrow K_a/K_s < 1$ ，基因受纯化选择(purify selection)



一、正选择分析简介



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

- 同义突变：表示基因序列改变，但氨基酸序列没有改变。其原因是密码子的简并性造成的，一种氨基酸可有多种密码子编码。在遗传学上我们一般用同义突变频率 (K_s 或者 dS)表示。不受自然选择影响。
- 非同义突变表示基因序列突变后，导致氨基酸序列也发生了改变。在遗传学上我们一般用非同义突变频率(K_a 或者 dN)表示。受到自然选择。
- K_a/K_s 或者 dN/dS 表示的是非同义替换 (K_a) 和同义替换 (K_s) 之间的比例。这个比例大小可以用来判断是否有选择压力作用于这个蛋白质编码基因。

$K_a > K_s$ 或者 $K_a/K_s > 1$ ，基因受正选择(positive selection)

$K_a = K_s$ 或者 $K_a/K_s = 1$ ，基因中性进化(neutral evolution)

$K_a < K_s$ 或者 $K_a/K_s < 1$ ，基因受纯化选择(purify selection)

- 受到正选择的基因，往往与生物特殊的功能密切相关！

一、正选择分析简介



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

需要注意的是： Ka/Ks 计算不仅要考虑基因序列差异，同时也可要考虑**基因的不同位点**以及该基因所属的物种在**进化枝系**中的进化速率来综合决定。

- 因为不同碱基之间发生替换的速率的不同，比如转换（transition，嘌呤 $\rightleftharpoons$ 嘌呤，嘧啶 $\rightleftharpoons$ 嘧啶）发生的概率要高于颠换（transversion，嘌呤 $\rightleftharpoons$ 嘧啶，嘧啶 $\rightleftharpoons$ 嘌呤）；基因序列不同位点选择压力也是不同的。
- 物种在不同进化支系的进化速率也存在差异；
- 采用一些替换速率模型对 Ka/Ks 进行修正

一、正选择分析简介



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

PAML软件中枝点模型 (Branch- site model)

枝点模型其实是Branch model 和site model的合集，在该model下，不仅假定不同位点间的 ω (ka/Ks) 值是不同的，同时也假定不同支系间的 ω 值也是不同的。

备择假设(alternative hypothesis): 通常将研究者想收集证据予以支持的假设称为备择假设。

原假设 (null hypothesis): 通常将研究者想收集证据予以反对的假设称为原假设。

- Model A(备择 Model) 的 fix_omega 设成0，即指定了背景枝 $\omega_0 < 1$ ，因此前景枝 $\omega_2 > 1$,均处于正选择。
- null Model (原Model) 中fix_omega=1 和 omega=1，即使 $\omega_2=1$ 。指不允许任何位点的 ω 值大于1；如果nullmodel 不成立,则 ω 大于1，基因受到正选择。
- 对Model A和 null Model卡方检验，可以直接用来鉴定前景枝中的positive selection 位点。

二、正选择分析步骤



采用codeml进行分析。

➤ 输入1：单拷贝基因家族的基因树，研究物种用#1标记前景枝。

(rna17311:0.10,((Zm00001d038685_P002:0.03,SbRio.09G205700.1:0.022)100:0.0393,(Seita.3G181600.1:0.029,C2845_PM05G32370_1_mrna#1:0.032)100:0.0188)100:0.0681,LOC_Os05g43790.1:0.098);

➤ 输入2：单拷贝基因家族的多序列比对文件

➤ 输入3：两个模型配置文件：

模型A（Alternative model）的配置文件，fix_omega设成0

对比模型(null model)的配置文件，注意omega和fix_omega均是1。

➤ 主要参数：

- Model命令主要是对系统发育树中不同支系的omega值进行限定。2表示指定的前景枝和背景枝之间的omega值不同。
- NSsites是在Site Model和Branch-site Model中进行设置的参数，选择2，为进行正选择分析的参数。

运行：

codeml 模型A配置文件

codeml 对比模型配置文件

三、正选择分析结果



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

正选择基因及相应的受选择位点

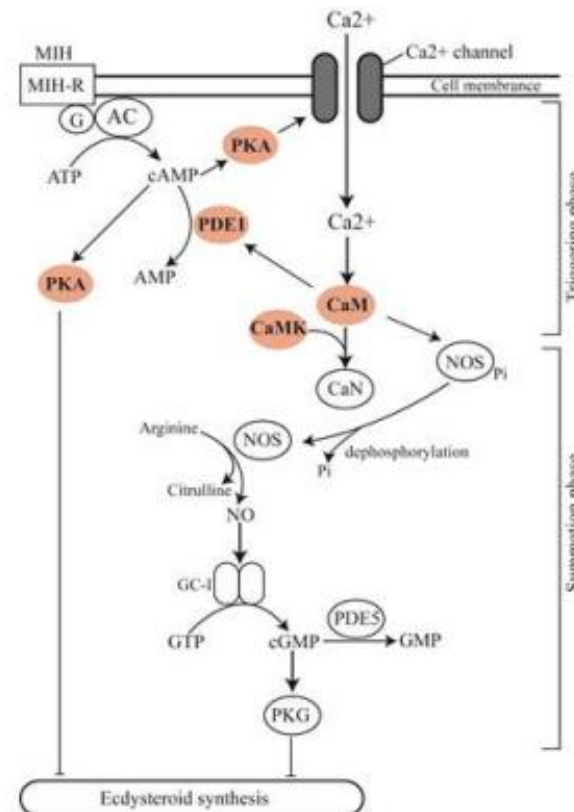
```
#gene    branch_pvalue    sites  
C2845_PM03G23850_1_mrna 0.002767979    528,R,0.975*
```

四、正选择分析



3、案例-南美对虾MIH信号通路正选择分析

MIH受体介导的甲壳动物蜕皮调控研究
五个正选择基因分布此通路上！



Supplementary Figure 22. Positively selected genes related to MIH signaling pathway in *L. vannamei*.

The circles with red backgrounds represent the positively selected genes in *L. vannamei*. Modified from Chang and Mykles², and Covi³.

Zhang, X., Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., ... & Ma, K. Y. (2019). Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nature communications*, 10(1), 356.



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

第五部分：共线性分析

一、共线性分析简介



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

- 分为两种方式，一种基因水平共线性，另外一种基因组水平的共线性。对于亲缘关系较远一些的一般采用基因共线性。
- 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；
- 基因组共线性分析软件MUMMER。一般应用于泛基因组中不同品种的基因组的比较，包括祖先二倍体种与多倍体种之间的比较，也应用于不同的基因组版本之间的比较。

一、共线性分析简介



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

其主要的意义:

- 1、结构变异的挖掘。不同品种之间，祖先二倍体种与四倍体种之间等，基因顺序一致的，基因组结构变异分析。
- 2、基因组组装准确性验证。
- 3、观察全基因组复制事件。
- 4、计算Ka/Ks。
- 5、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

一、共线性分析简介



共线性示例：

Duplication depth	Reference chromosome	Collinear blocks
1	C2845_PM02G00010_1_mrna	C2845_PM01G29780_1_mrna
1	C2845_PM02G00020_1_mrna	C2845_PM01G29790_1_mrna
1	C2845_PM02G00030_1_mrna	C2845_PM01G29800_1_mrna
2	C2845_PM02G00040_1_mrna	C2845_PM01G29820_1_mrna
3	C2845_PM02G00050_1_mrna	C2845_PM01G29840_1_mrna
7	C2845_PM02G00060_1_mrna	C2845_PM01G29850_1_mrna
8	C2845_PM02G00070_1_mrna	C2845_PM01G29860_1_mrna
8	C2845_PM02G00080_1_mrna	C2845_PM01G29870_1_mrna
8	C2845_PM02G00090_1_mrna	
8	C2845_PM02G00100_1_mrna	C2845_PM01G29880_1_mrna
8	C2845_PM02G00110_1_mrna	C2845_PM01G29890_1_mrna
8	C2845_PM02G00120_1_mrna	
8	C2845_PM02G00130_1_mrna	C2845_PM01G29900_1_mrna
13	C2845_PM02G00140_1_mrna	C2845_PM01G29920_1_mrna
13	C2845_PM02G00150_1_mrna	C2845_PM01G29930_1_mrna
13	C2845_PM02G00160_1_mrna	
13	C2845_PM02G00170_1_mrna	
13	C2845_PM02G00180_1_mrna	C2845_PM01G29940_1_mrna
13	C2845_PM02G00190_1_mrna	C2845_PM01G29960_1_mrna
13	C2845_PM02G00200_1_mrna	C2845_PM01G29970_1_mrna
13	C2845_PM02G00210_1_mrna	C2845_PM01G29980_1_mrna
13	C2845_PM02G00220_1_mrna	C2845_PM01G29990_1_mrna

序列相似
位置相邻

二、共线性分析步骤



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

基因共线性:

- 1、BLAST。寻找序列相似的基因对，确认哪些基因可以比对上。
- 2、GFF文件。确定每个基因在染色体上的位置。
- 3、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。

基因组共线性(MUMMER):

比对: **nucmer** --prefix=ref_qry Sl.Chr1.genome.fa Pm.Chr1.genome.fa

过滤(可选): **delta-filter** -l 1000 -i 90 ref_qry.delta > ref_qry.filter.delta

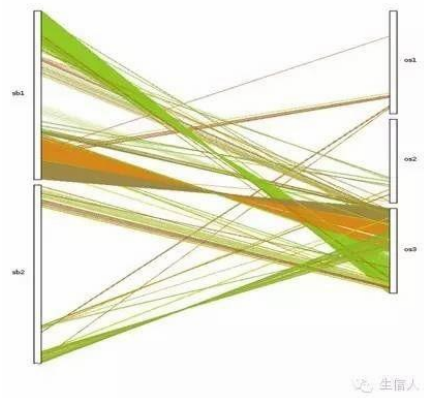
结果可读: **show-coords** -rclT ref_qry.filter.delta > ref_qry.filter.coords

可视化: **mummerplot** ref_qry.filter.delta --png --filter -p test

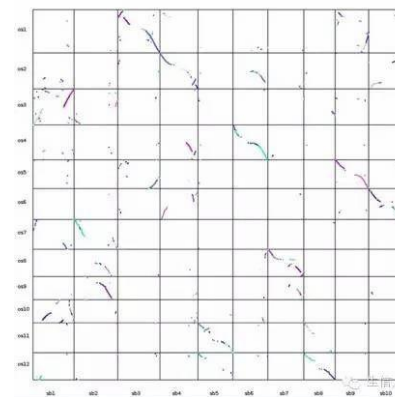
三、共线性分析结果



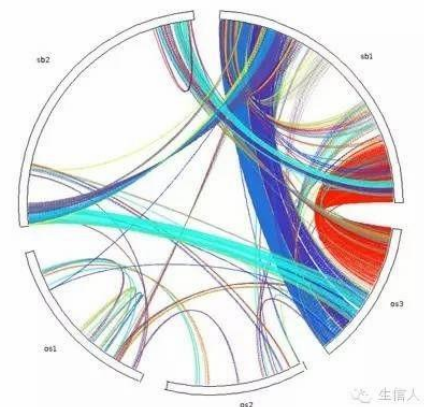
相互连线图



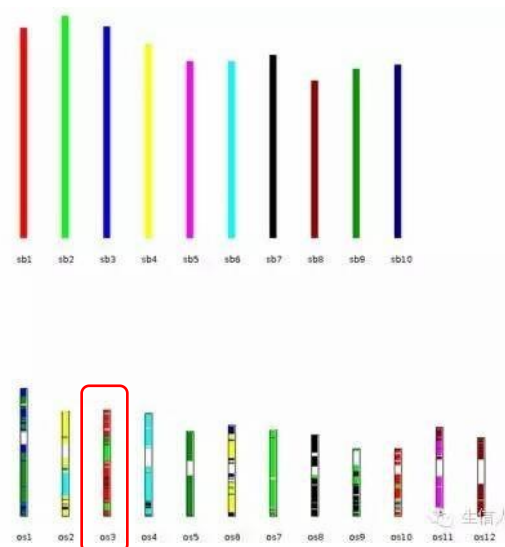
点图



圈图



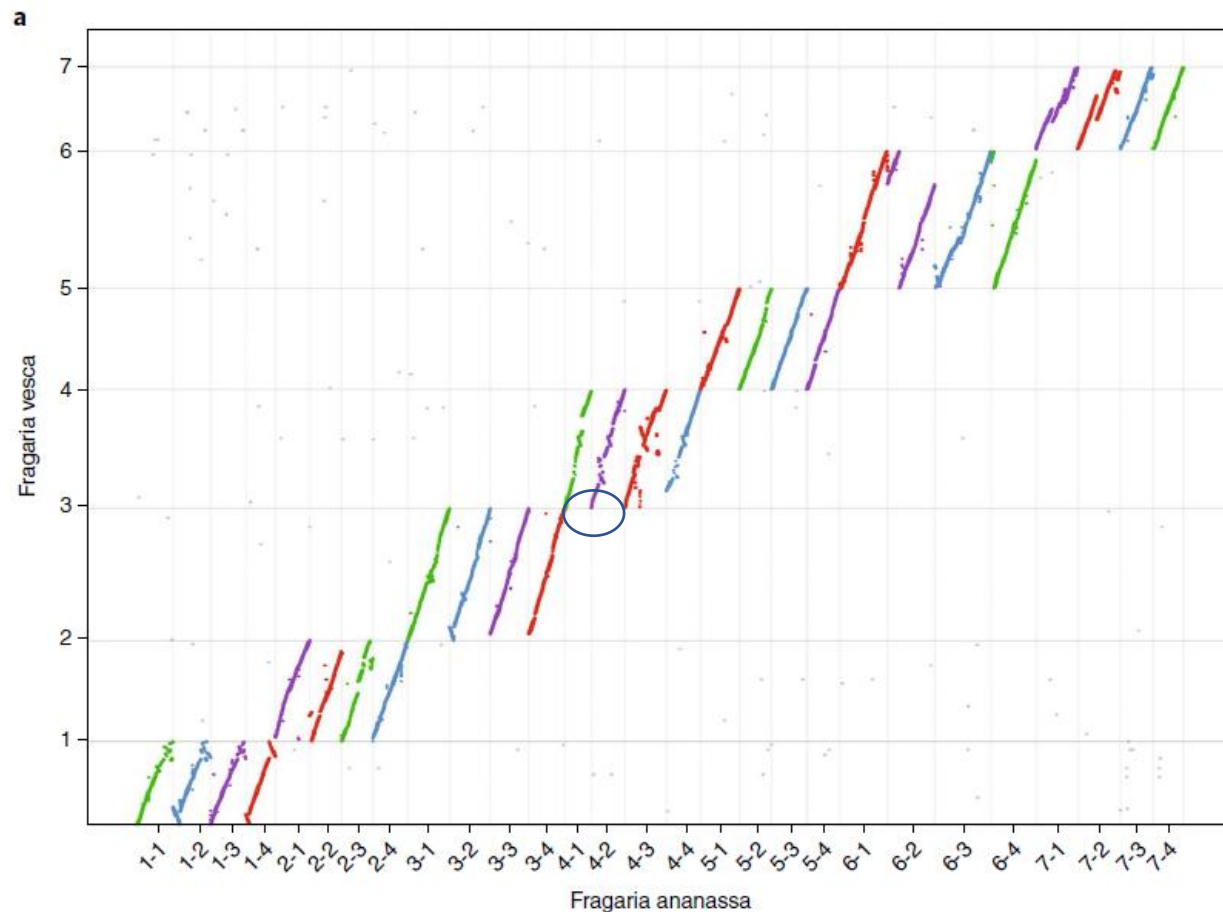
柱形图



五、共线性分析



案例-八倍体草莓与二倍体草莓基因组之间共线性分析



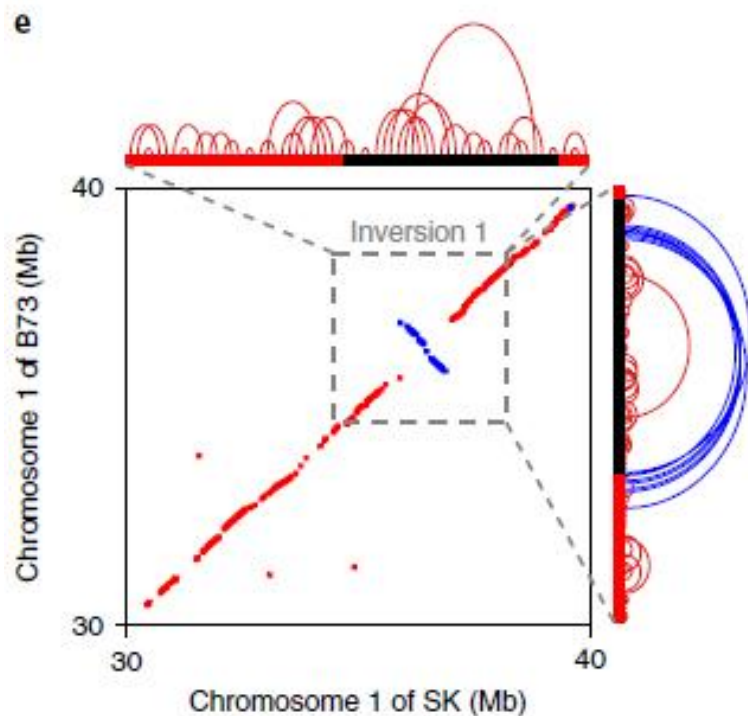
横坐标：八倍草
莓四套亚基因组

摘自：Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., ... & Alger, E. I. (2019). Origin and evolution of the octoploid strawberry genome. *Nature genetics*, 51(3), 541.

五、共线性分析



案例-两个玉米之间染色体1的基因组



中间区段存在倒位!

摘自: Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L., ... & Wang, Y. (2019). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nature genetics*, 51(6), 1052.



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

第六部分：全基因组复制事件分析

一、全基因组复制事件简介



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

- 全基因组复制事件 (Whole genome duplication, WGD) , 亦称古多倍化, 是基因组整体复制加倍的过程。主要发生在**被子植物**和**鱼类**等物种中, 对于物种起源和基因组扩张中有着重要的意义。
- whole-genome triplication(WGT): 全基因组三倍化事件(古六倍化)

一、全基因组复制事件简介



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

研究进展:

- 1997 年，在酿酒酵母(*Saccharomyces cerevisiae*)基因组中证实了 WGD 假想。
- 2000年，首例植物全基因组草图测序完成，解析了拟南芥基因组进化历史，提出拟南芥可能有一个四倍化祖先。
- 2007年，**葡萄基因组**测序完成，古基因组学研究表 明葡萄的祖先基因组可能是古六倍体。对葡萄基因组重复事件的研究奠定了双子叶植物祖先基因组共享六倍化的理论基础。
- 2011年，卷柏基因组测序完成，没有发现古多倍化证据。
- 2013年，**无油樟基因组**测序草图公布，WGD分析表明其分化发生在古六倍化前
- 2015年，**菠萝基因组**测序完成，古基因组学研究表明其基因组发生了 2 次 WGD事件。但没有发生谷物 ρ WGD
- 2016年，银杏基因组草图完成，古基因组学研究表明银杏发生了2 次WGD 事件，其中较近 1 次是银杏特异性WGD。

一、全基因组复制事件简介



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

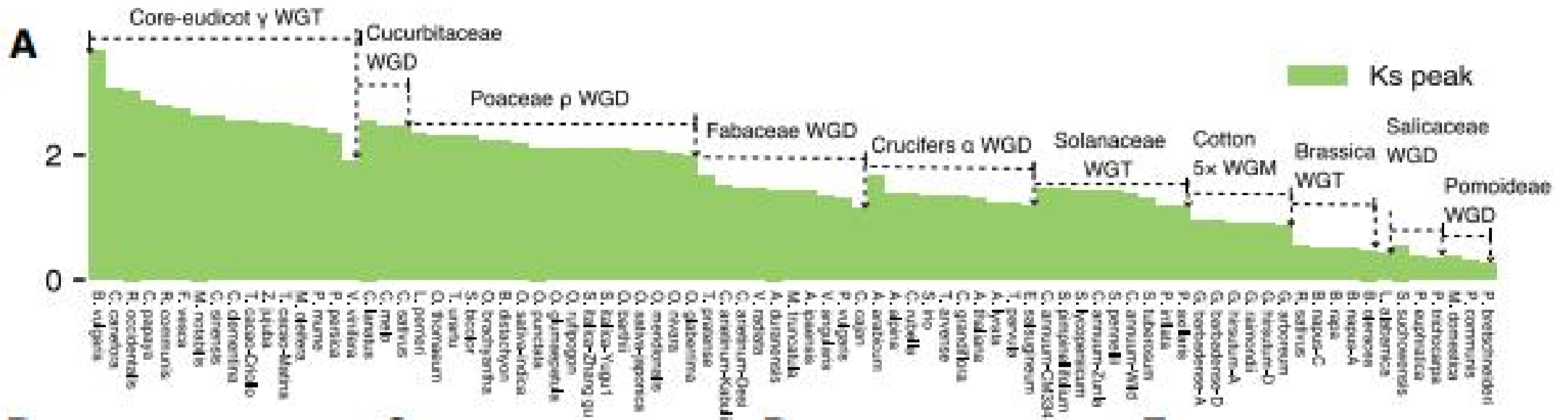
动物全基因组复制发生时间

- 脊椎动物的祖先被认为在至少450百万年前经历了两轮古WGD（定义为1r和2r），约20–30%人类的旁系同源基因被认为是由于这两次WGD产生的。
- 在约320百万年前，硬骨鱼的共同祖先发生了一次WGD(定义为3r)；
- 在约80百万年前，鲑鱼，包括虹鳟和大西洋鲑鱼又发生了一次全基因组复制事件（称为4R）。
- 最新的鲤鱼基因组，在约8.2百万年，发生了WGD（4R）。

一、全基因组复制事件简介

植物全基因组复制发生时间

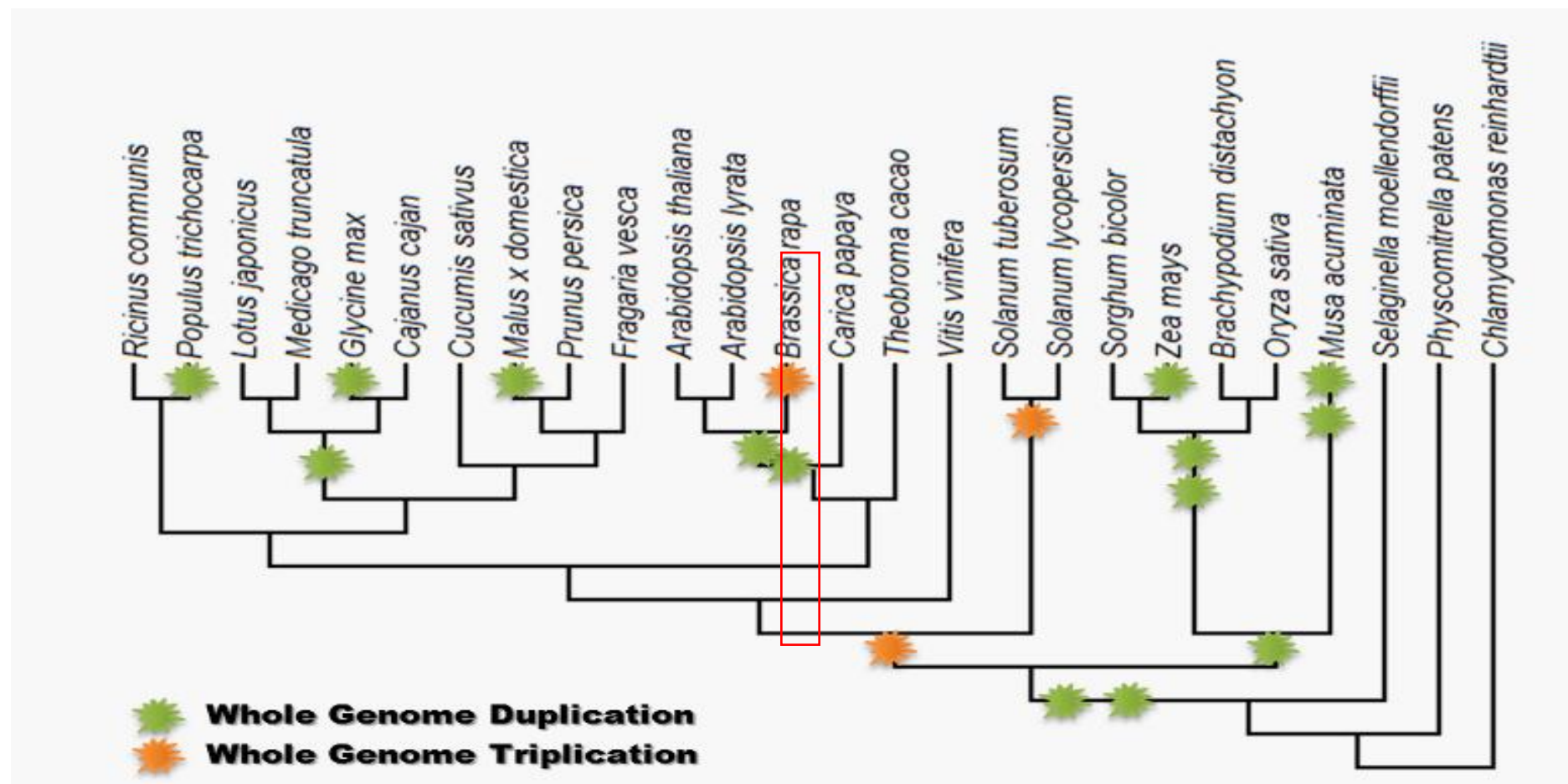
- 被子植物基因组经历了二倍化(α)、四倍化(β)和六倍化(γ)3次古WGD事件， γ 发生在**130~150 MYA**百万年为WGT。但只有一个物种作为特例：无油樟（被认为是被子植物中最古老的种）。
- 禾本目谷类植物基因组经历了 τ ， σ 和 ρ 3次WGD事件。现代谷物的祖先大约在**70**百万年前发生第一次WGD事件，被命名为“ **ρ** ”。菠萝没有发生 ρ 这次事件。



一、全基因组复制事件简介



PGDD数据库检测的全基因组复制事件



一、全基因组复制事件简介



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

对于经历过多次全基因组复制事件的物种来说，后来的加倍事件会加速上一次加倍事件的基因丢失，这决定了基因和基因组不会无限制的增加！

全基因组加倍事件结果：可以增加一个物种所有的基因拷贝，但在自然选择的作用下，倍增后的基因会经历不同的命运：

- 部分拷贝丢失；
- 失去功能（假基因化）；
- 部分拷贝获得新的功能（新功能化）；
- 或者各自行使祖先基因的部分功能（亚功能化）。

二、全基因组复制事件鉴定



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

- 共线性方法。
共线性区块。
- Ks方法。
Ks分布图。
直系同源基因对之间的峰值代表物种分歧；
旁系同源基因对之间的峰值代表本物种的全基因组复制事件
- 4DTv方法。
4DTv分布图。
密码子的第三个位点无论是哪种核苷酸，均编码同样的氨基酸。

二、全基因组复制事件鉴定



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

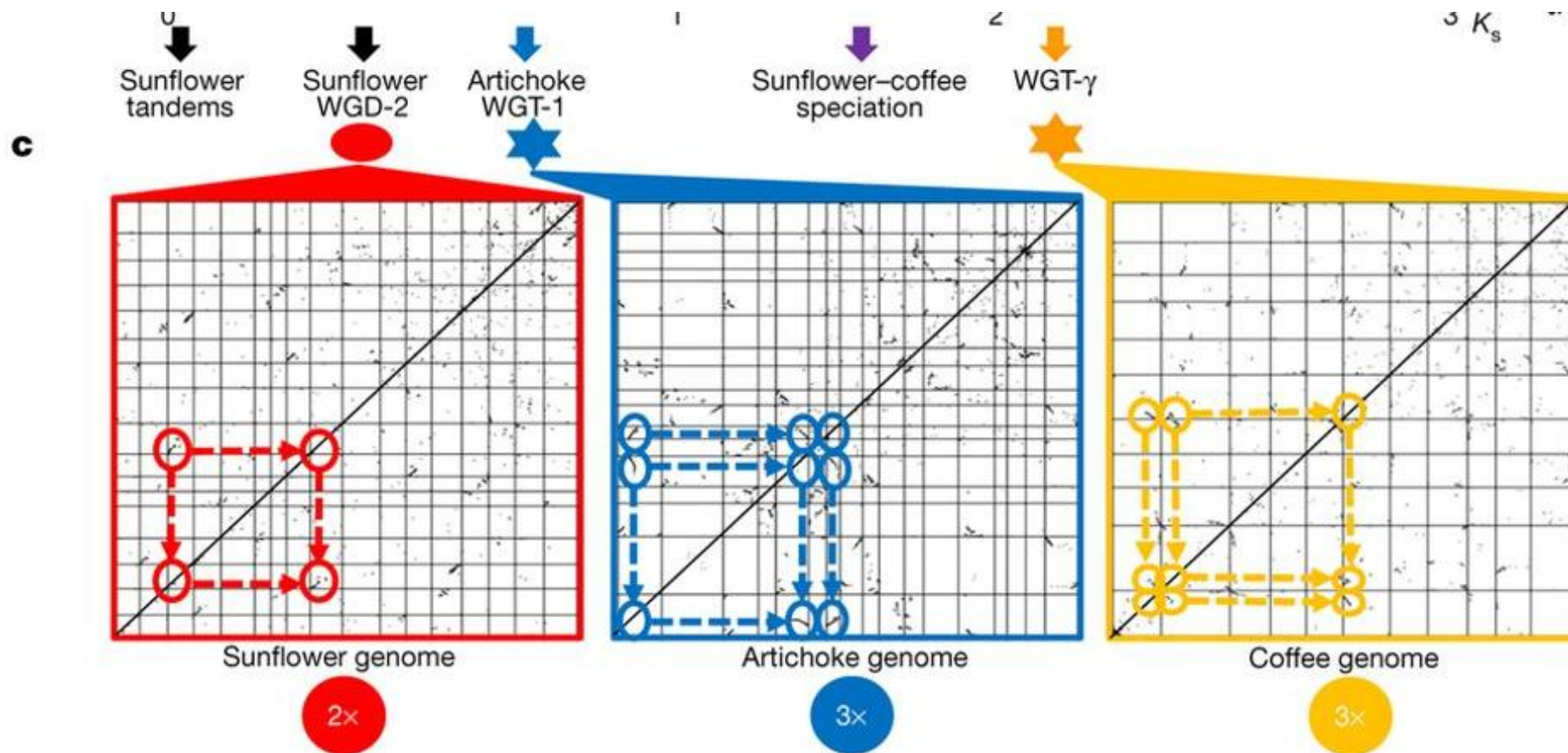
1、共线性方法

- 比较两个物种（比较的对象，未发生**新的**全基因组复制事件的，如葡萄、无油樟）的基因序列，如果研究物种多个片段对应比较物种的**1**个片段，则证明该物种相对于另外一个物种发生了新的全基因组复制事件。
- 对于单个物种，以绘制基因组内部的基因共线性的点状图，如果发现大量的基因组的某一区间可以匹配到其他多个不同的区间中，这就暗示了该物种经历全基因组的复制事件。
- 对于经历过多次全基因组复制事件的物种来说，后来的加倍事件会加速上一次加倍事件的基因丢失，使共线性变得模糊

二、全基因组复制事件鉴定



1、共线性方法



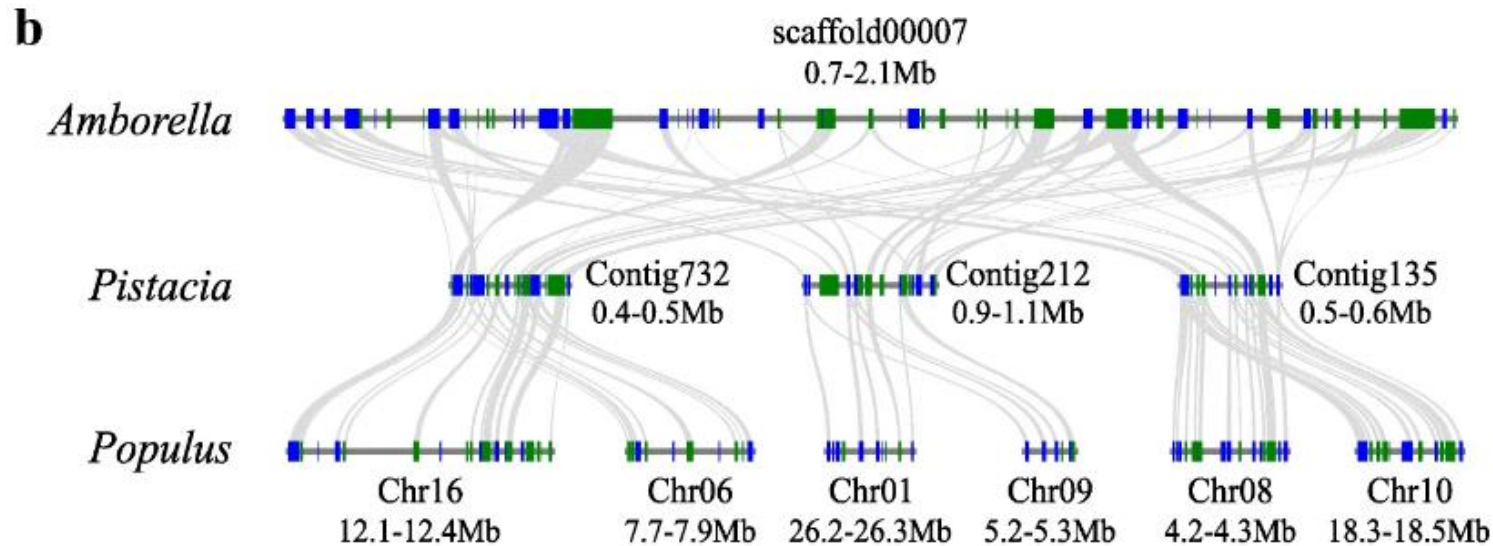
The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution[J]. Nature, 2017.

六、全基因组复制事件分析



案例-开心果基因组基因共线性

将开心果基因组与基础被子植物无油樟基因组进行共线性分析，表明每个无油樟区域最多有三个开心果区域，而每个开心果区域最多有两个毛果杨区域。开心果中没有发生谱系特异性基因组复制，但它们与真双子叶植物中发生的 γ 复制相同，而毛果杨经历了谱系特异性基因组复制事件。



摘自：Zeng, L., Tu, X. L., Dai, H., Han, F. M., Lu, B. S., Wang, M. S., ... & Ji, L. L. (2019). Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome biology*, 20(1), 79.

二、全基因组复制事件鉴定



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

2、Ks方法。

- 同义突变率Ks方法，这也是目前比较流行的方法。这种方法的背景是认为Ks值在某种程度上反映了同源基因的产生时间。而全基因组加倍事件会产生大量的同源基因，反映在Ks值上便是会有大量的**Ks值接近的同源基因对**的产生，这样通过绘制Ks值的分布图便可以发现**明显的Ks值峰**，而这些峰的存在也就对应了全基因组加倍事件的发生，对应的Ks值可以用于计算发生的时间。
- 通常选用多个物种组合进行全基因组复制事件的比较分析。物种内组合代表全基因组复制事件，物种间组合代表物种分歧。
- 这种方法是基于两点假设：**1.基因的突变频率是稳定的；2.同义突变（Ks）不会影响物种适应性，因为并不会造成氨基酸序列的变化。**

二、全基因组复制事件鉴定



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

2、Ks方法。

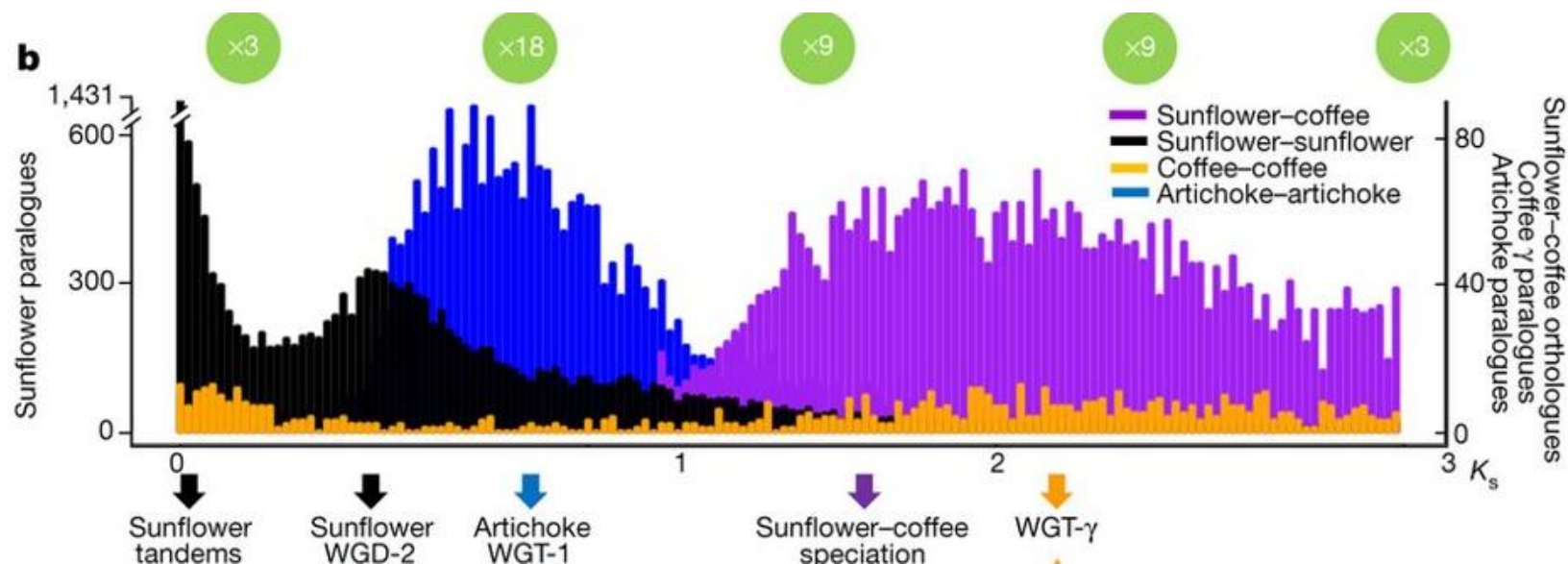
- 同义突变率Ks方法，这也是目前比较流行的方法。这种方法的背景是认为Ks值在某种程度上反映了同源基因的产生时间。而全基因组加倍事件会产生大量的同源基因，反映在Ks值上便是会有大量的**Ks值接近的同源基因**对的产生，这样通过绘制Ks值的分布图便可以发现**明显的Ks值峰**，而这些峰的存在也就对应了全基因组加倍事件的发生，对应的Ks值可以用于计算发生的时间。
- 通常选用多个物种组合进行全基因组复制事件的比较分析。物种内组合代表全基因组复制事件，物种间组合代表物种分歧。
- 这种方法是基于两点假设：**1.基因的突变频率是稳定的；2.同义突变（Ks）不会影响物种适应性**，因为并不会造成氨基酸序列的变化。

二、全基因组复制事件鉴定



2、Ks方法。

横坐标是同源基因对的Ks值的分布，纵坐标为同源基因的数量，不同颜色的柱子代表不同的物种组合，比如黑色就是向日葵基因组内部的Ks分布。每一个峰都对应一次全基因组加倍事件，比如向日葵的WGD-2，因为发生的较晚，所以峰所在的位置Ks值较小，而且峰比较明显。而最下面橙色的峰，是咖啡发生的全基因组三倍化事件，峰值所在位置Ks值较大，但峰已经很不明显。同样，蓝色的峰为洋葱的WGT-1事件，紫色的峰代表向日葵和咖啡的分化事件。



二、全基因组复制事件鉴定



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

3、4DTv方法。

4DTv(four-fold synonymous (degenerative) third-codon transversion)一个遗传密码子通常由三个核苷酸构成，从左到右依次为第一个位点、第二个位点、第三个位点。如果密码子的第三个位点无论是哪种核苷酸，均编码同样的氨基酸，则称这个位点为4倍简并位点。例如甘氨酸密码子（GGA, GGG, GGC, GGU）的第三个位点就是一个4倍简并位点。4倍简并位点存在使基因更加耐受点突变。

Seq1: CTT CCT ACT ATT GTG

Seq2: CTA CCC ACA ATC GTA

L P T I V

4DTv=?

六、全基因组复制事件分析



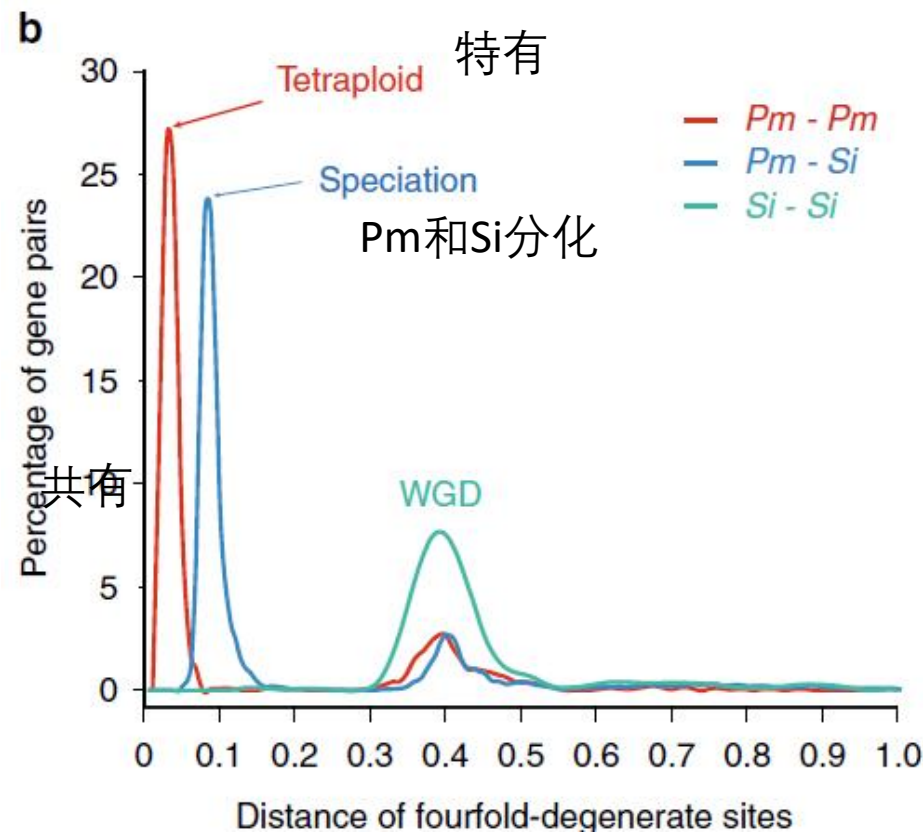
案例糜子基因组全基因组复制。

4DTV:

four-fold synonymous (degenerative) third-codon transversion。

一个遗传密码子通常由三个核苷酸构成，从左到右依次为第一个位点、第二个位点、第三个位点。如果密码子的第三个位点上无论是哪种核苷酸，均编码同样的氨基酸，则称这个位点为4倍简并位点。

阐明糜子基因组全基因组复制事件的发生。





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