1个样本T2T建库测序 进化分析报告

项目名称：

项目编号：

分析人员：

审核人员：

报告日期：

报告单位：西安浩瑞基因技术有限公司

1 项目基本信息

物种名称	{}
物种拉丁学名	*{}*
参与进化分析物种	-

2 分析流程

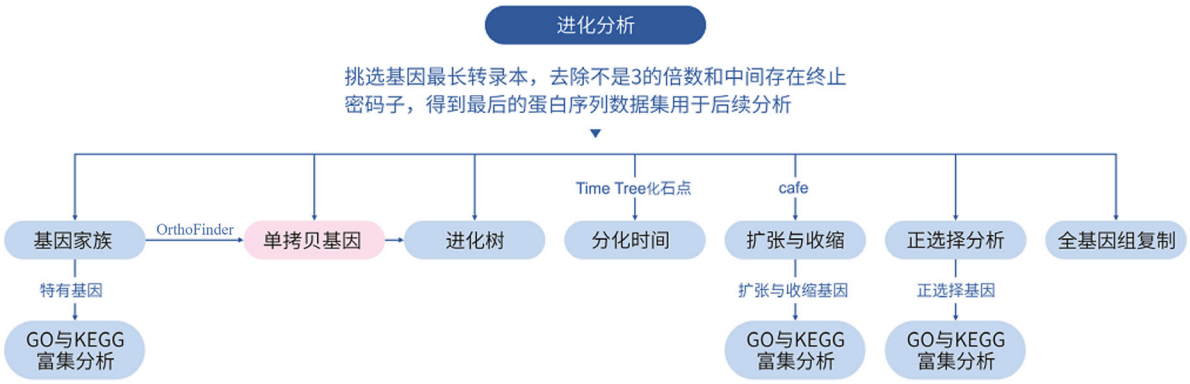


Figure 2-1 基因组进化分析流程图

3 分析结果

3.1 分析数据统计

基于进化分析物种需求，利用下表中的物种数据进行进化分析。相关分析物种注释基因将去除可变剪切基因，保留最长转录本，同时过滤掉那些基因内部存在终止密码子（stop codon）和不是3的倍数的基因用于后续分析。用于进化分析的物种名称等具体信息见下表：

Table 3.1-1 分析物种信息统计

物种简称	物种拉丁文名称	原始基因数	过滤后基因数
-	-	54288	46773
-	-	33078	28165
-	-	33456	27525
-	-	25985	25811
-	-	37969	31807
-	-	28143	25056
-	-	40142	40105
-	-	38478	32701
-	-	28881	24890
-	-	41359	38255
-	-	28400	26600
-	-	60286	39968
-	-	37729	37729

物种简称	物种拉丁文名称	原始基因数	过滤后基因数
-	-	47491	36862
-	-	27433	27433
-	-	41330	33559

3.2 基因家族分析

基因家族是来自一个祖先基因的一组基因。基因家族的鉴定，是进化分析中很重要的内容。通过同源基因的聚类及基因家族的鉴定分析，可以得到单拷贝基因家族和多拷贝基因家族。 这些家族在物种之间都是比较保守的，可用于物种间亲缘关系的分析。我们还可以得到物种特有的基因和家族，它们可能和物种的特异性表型有关。此外还可以从整个基因水平研究物种的进化问题。

3.2.1 基因家族聚类

基于OrthoFinder[1]软件对以下**16**个物种进行基因家族聚类分析，将基因家族分析得到的结果进行分类，并对每个类别所包含的基因进行统计，结果如下： 其中**{}**是本次分析的目标物种，得到其单拷贝基因（Single）**530**个，特有（unique）基因**1,326**个， 未聚类（uncluster）基因**1,762**个，物种特异基因（uncluster和unique基因合并）**3,088**个， 具体结果如表Table 3.2-1（具体分类见Table 3.2-1标注，Single，Multi，Unique，Other为家族分类单元，Unclustered为家族分析中未进行聚类的单元）。

进一步基于Table 3.2-1统计结果进行图形展示，具体结果见Figure 3.2-1。

Table 3.2-1 各物种基因家族类型的基因数量统计

Species	Single	Multi	Unique	Other	Unclustered
-	530	15,643	6,844	12,289	1,556
-	530	14,982	4,177	5,334	2,502
-	530	14,486	709	8,846	319
-	530	15,354	6,876	12,514	2,455
-	530	26,996	1,379	16,218	1,650
-	530	14,709	1,326	7,484	1,762
-	530	20,705	1,988	10,748	4,284
-	530	15,580	3,902	11,864	825
-	530	16,286	1,132	13,146	713
-	530	15,958	6,470	15,900	1,110
-	530	15,604	275	10,174	850
-	530	17,231	1,924	13,176	698
-	530	15,536	293	9,990	251
-	530	16,565	6,824	11,510	4,676
-	530	15,630	263	11,433	309
-	530	13,540	2,505	7,421	1,060

说明:

Species：物种名称；

Single：基因家族中所有物种基因个数都为1的家族；

Mutil：所有物种个数都不为0，但至少有一个物种大于1的家族；

Unique：除自身物种外，其余物种均为0；

Other：自身大于0，其余物种不都为0，但是至少有一个为0的家族；

Unclustered：既没有和其他物种聚类，自身也没有聚类。

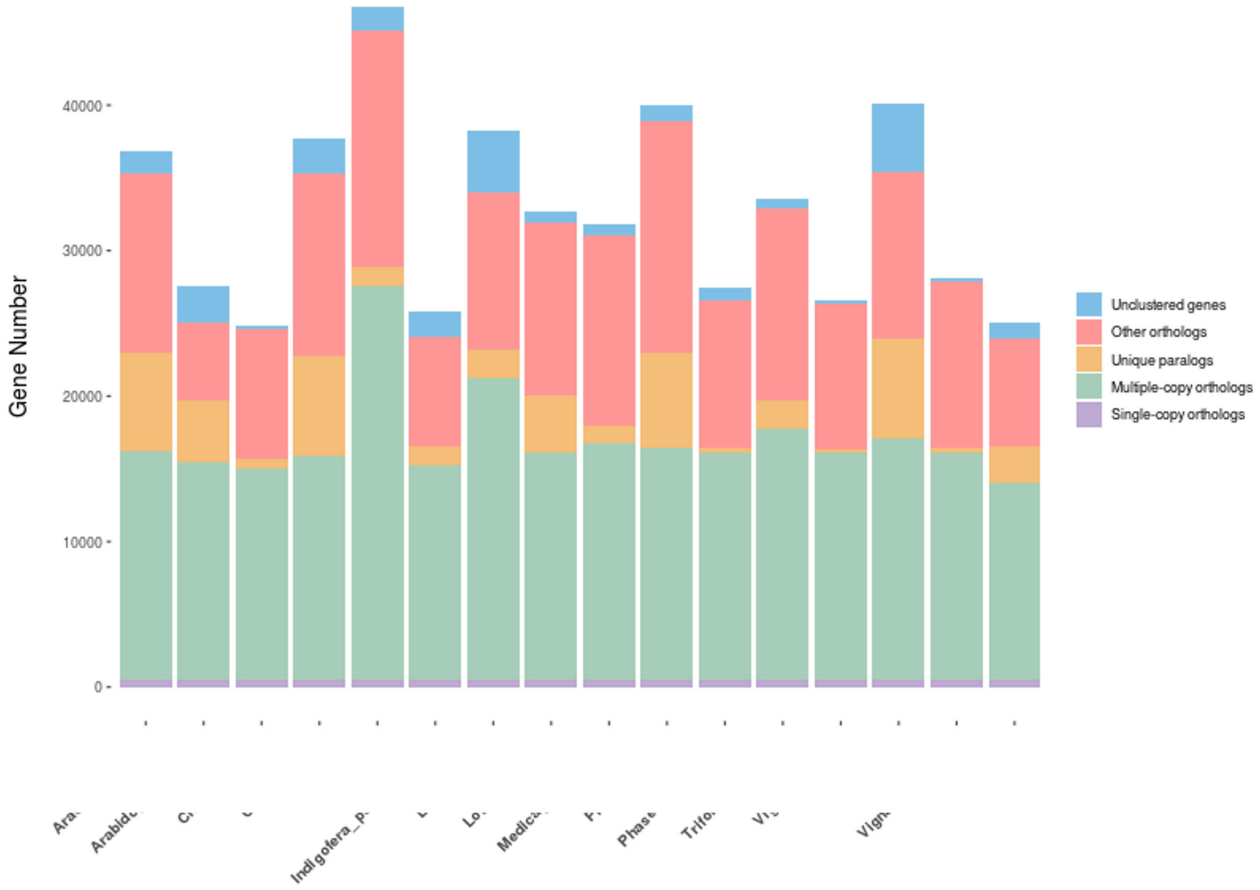


Figure 3.2-1 各样本基因家族类型分类图

基于基因家族分析结果，为了更加清晰地了解基因家族特征，针对基因家族分析结果再次以不同形式统计。目标物种{}， 其共有**25,811**个基因，最终家族聚类共得到**15,641**个基因家族， 其中**399**个基因家族为{}物种特异的基因家族。 {}的基因可以归类于不同基因家族的数量为**24,049**，无法归类的基因数目为**1,762**， 平均每个家族存在**1.54**个基因，详细结果见Table 3.2-2。

Table 3.2-2 基因家族中基因数统计

Species	Genes Number	Genes number in families	Unclustered genes number	Family number	Unique Families number	Average genes number per family
-	36,862	35,306	1,556	16,233	805	2.17
-	27,525	25,023	2,502	14,419	897	1.74
-	24,890	24,571	319	15,794	121	1.56
-	37,729	35,274	2,455	17,063	636	2.07
-	46,773	45,123	1,650	16,797	234	2.69
-	25,811	24,049	1,762	15,641	399	1.54
-	38,255	33,971	4,284	16,682	599	2.04
-	32,701	31,876	825	16,217	401	1.97

Species	Genes Number	Genes number in families	Unclustered genes number	Family number	Unique Families number	Average genes number per family
-	31,807	31,094	713	16,469	271	1.89
-	39,968	38,858	1,110	16,544	640	2.35
-	27,433	26,583	850	16,530	88	1.61
-	33,559	32,861	698	16,565	402	1.98
-	26,600	26,349	251	16,105	59	1.64
-	40,105	35,429	4,676	18,657	1,787	1.90
-	28,165	27,856	309	16,171	83	1.72
-	25,056	23,996	1,060	15,052	413	1.59

说明：

Species：物种名称；

Genes Number：各物种对应总的基因数；

Genes number in families：聚类的基因总数；

Unclustered genes number：未聚类的基因总数；

Family number：基因家族数；

Unique Families number：特有基因家族数；

Average genes number per family：平均每个家族的基因数。

3.2.2 直系同源基因鉴定

基于OrthoFinder[1]的结果进行直系同源基因的鉴定（具体方法见下方材料方法），将鉴定得到的各物种直系同源基因组进行统计，统计结果见下图。

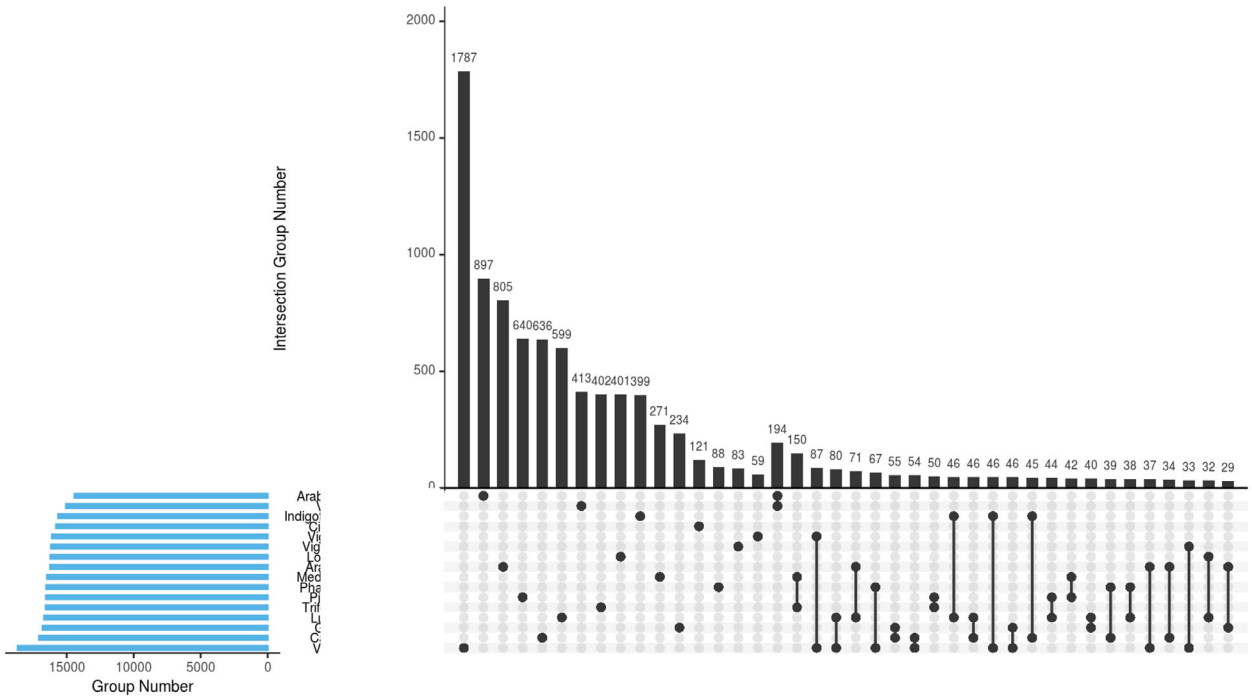


Figure 3.2-2 直系同源基因UpSet图

3.2.3 特异基因的分析

基于统计结果对单拷贝基因组和物种特异基因进行统计绘图，针对目标物种**{}**的基因分类结果中特异的两类结果进行统计， 其中此次分析中共获得了**530**个单拷贝直系同源基因，同时对于**{}**特异基因进行了统计， 共获得了**3,088**个特异基因（这类基因用于后续富集分析）。具体信息见下图。

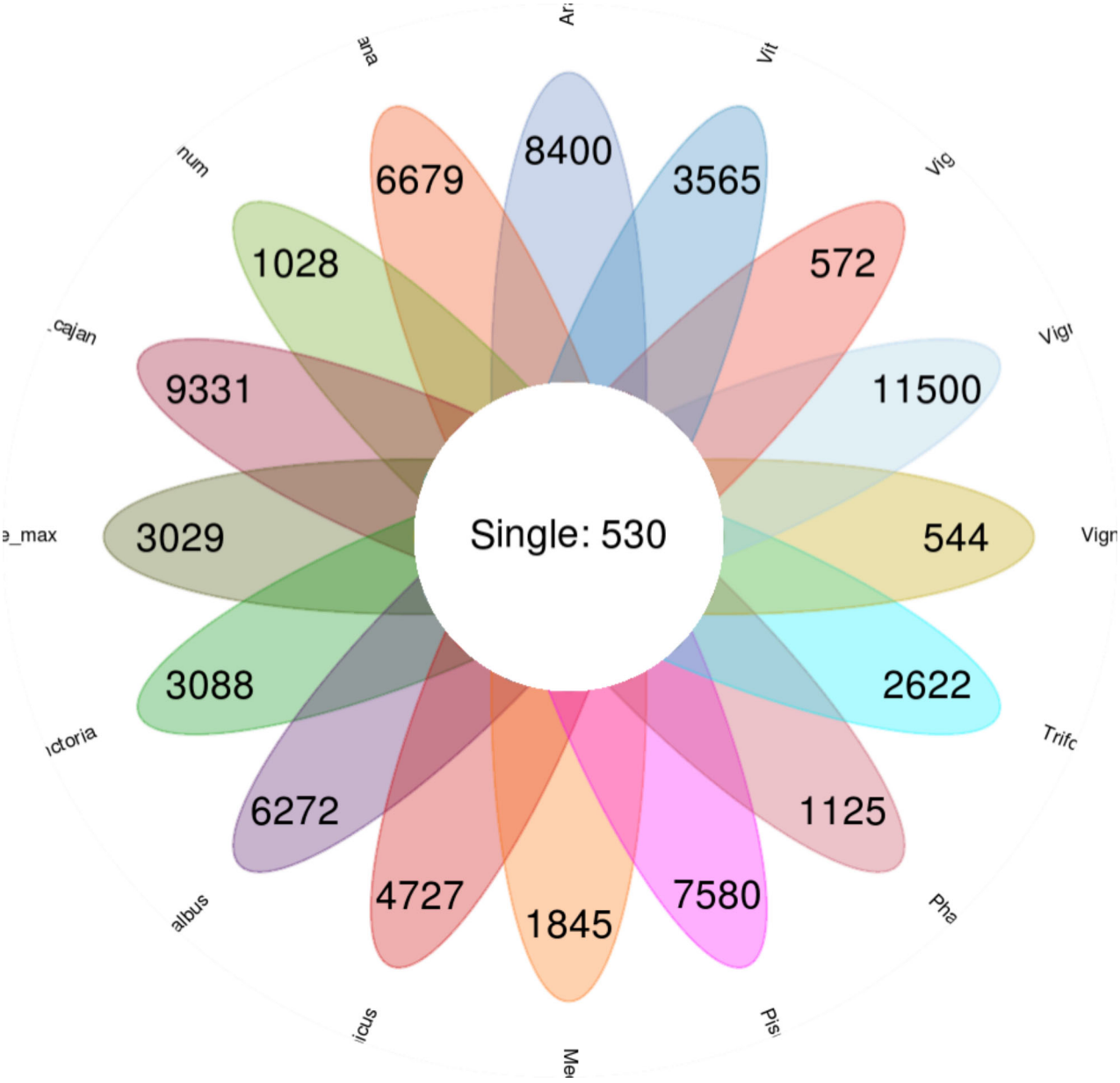


Figure 3.2-3 各样本单拷贝基因与物种本身特有基因花瓣图

说明：Single：单拷贝基因数；花瓣上的数字：表示的是该物种unique和unclustered基因数之和。

3.2.4 物种特异基因富集分析

将基因家族分析结果中，基因归类属于unique gene 和unclustered gene这两类基因视为物种特异基因。针对**{}**特异基因进行统计，具体统计结果见下表。

Table 3.2-3 物种特异基因信息统计

Species	Unique Gene	Unclustered	Unique and Unclustered Gene
{}	1,326	1,762	3,088

3.2.4.1 GO富集分析

基于上述分析方法得到的**{}**特异基因进行GO聚类分析，按照Table 3.2-3的三种类别进行分析，具体结果见下图（取最为显著的前10项的结果做展示，详细结果见对应释放文件)。

1. Unique 基因GO富集分析结果：

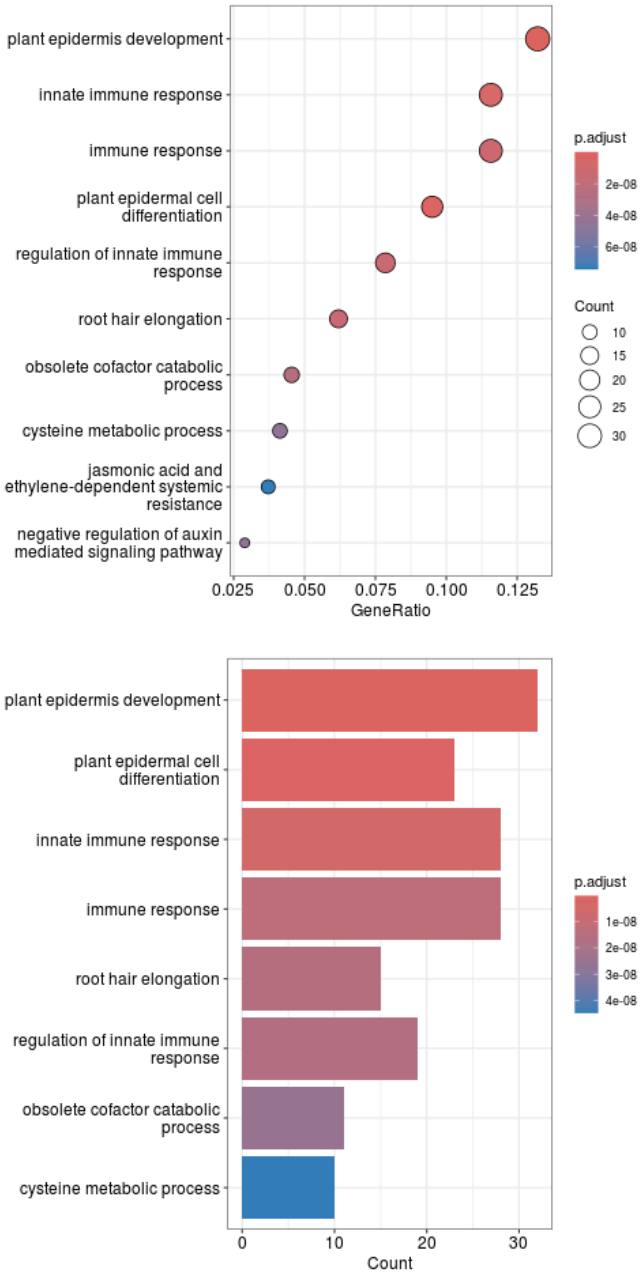


Figure 3.2-4 Unique 基因GO富集图

说明：

散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到GO条目上的基因个数，纵坐标为GO条目。

2. Unclustered 基因GO富集分析结果：

未检测到Unclustered基因的GO富集分析结果（Figure 3.2-5 Unclustered基因GO富集图未生成）。

3. Unique和Unclustered基因GO富集分析结果：

Figure 3.2-6 Unique和Unclustered基因GO富集图

说明：

散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到GO条目上的基因个数，纵坐标为GO条目。

3.2.4.2 KEGG富集分析

基于上述分析方法得到的{}特异基因进行KEGG聚类分析， 按照Table 3.2-3的三种类别进行分析，具体结果见下图（取最为显著的前十项的结果做展示，详细结果见对应释放文件)。

1. Unique基因KEGG富集分析结果：

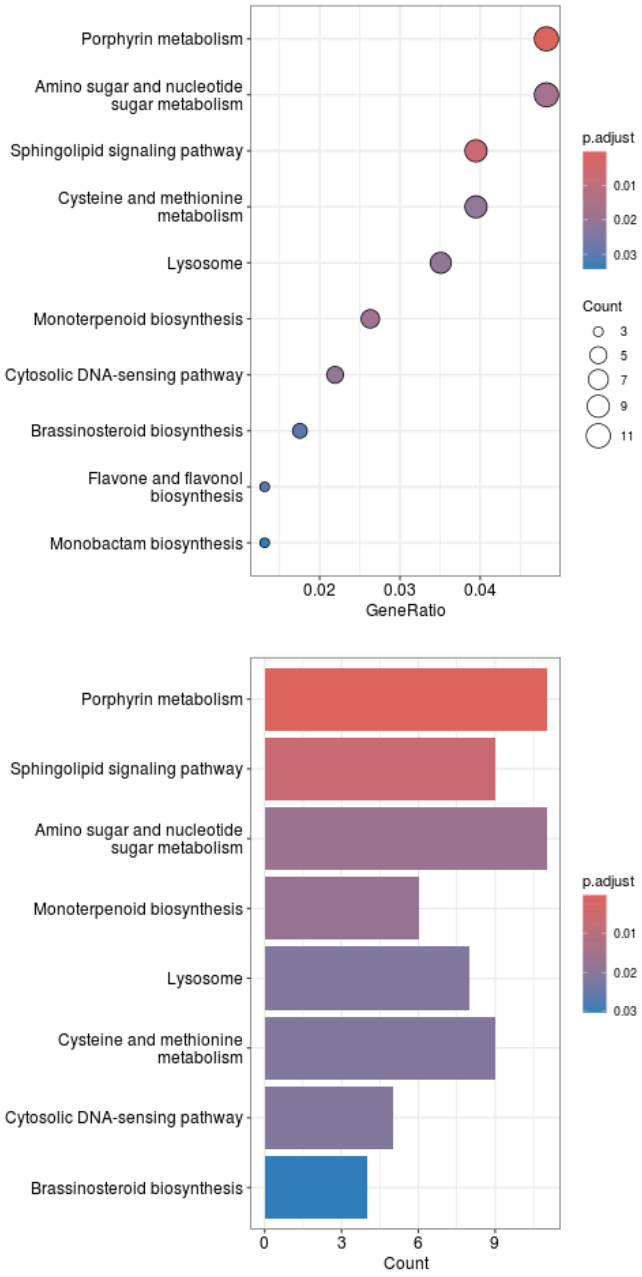


Figure 3.2-7 Unique基因KEGG富集散点图

说明：

散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到pathway的基因个数，纵坐标为KEGG pathway分类。

2. Unclustered基因KEGG富集分析结果：

未检测到Unclustered基因的KEGG富集分析结果（Figure 3.2-8 Unclustered基因KEGG富集图未生成）。

3. Unique和Unclustered基因KEGG富集分析结果：

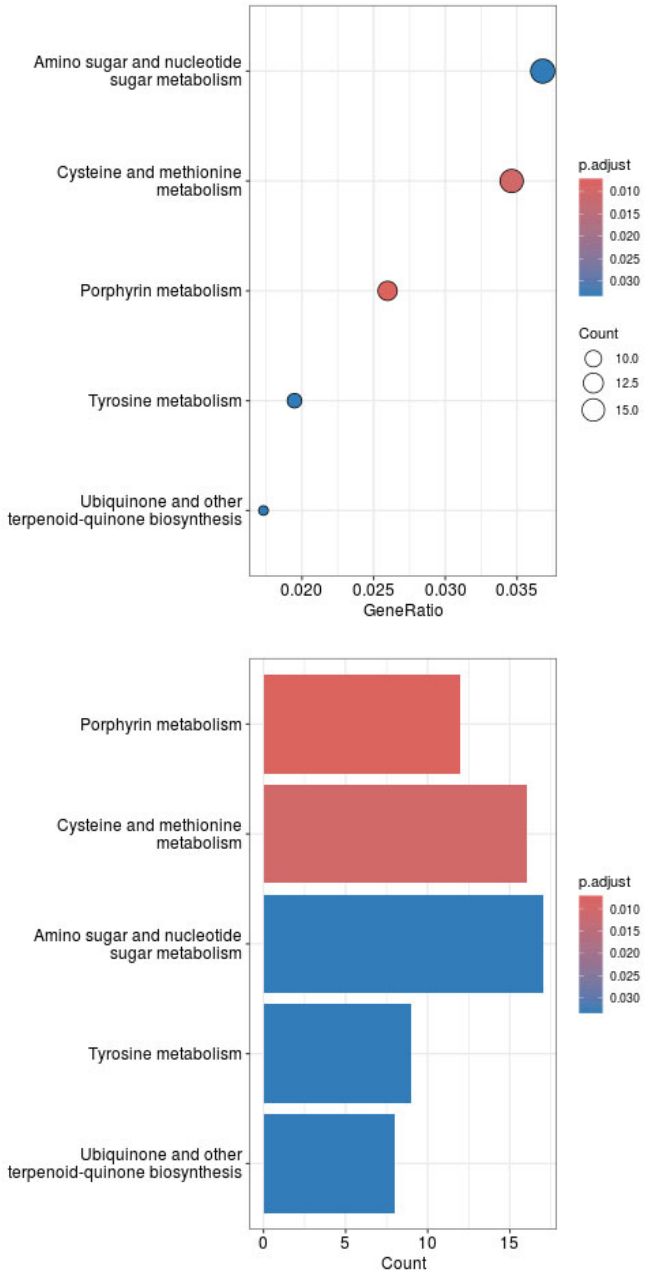


Figure 3.2-9 Unique和Unclustered基因KEGG富集散点图

说明：

散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到pathway的基因个数，纵坐标为KEGG pathway分类。

3.3 构建系统发育树

本分析基于上述基因家族鉴定结果（Table 3.2-2）中的单拷贝直系同源基因。 首先，使用MAFFT[2]软件进行比对得到蛋白多序列比对数据集（protein-MSA）； 其次，利用PAL2NAL[3]软件，得到protein-MSA对应的 CDS 序列比对数据集，即CDS-MSA； 最后，通过IQ-TREE[4]软件获得最佳模型，采用RAxML-NG[5]软件构建系统发育树， 并导入Figtree软件编辑，用*Vitis_vinifera*作为外群reroot进行展示如下：

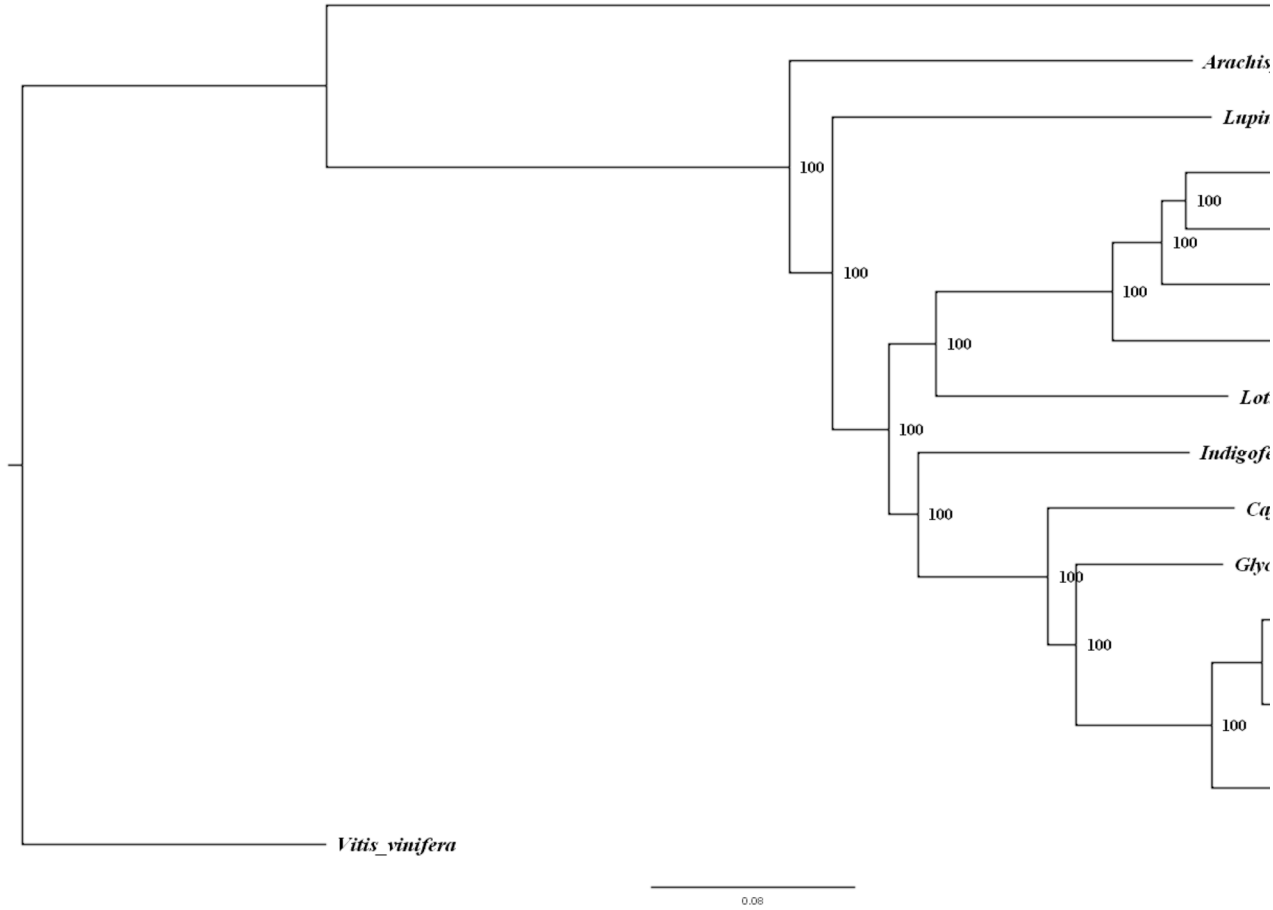


Figure 3.3-1 进化树树形结构

3.4 估算分歧时间

分歧时间估算是指以某一特定类群的化石记录作为参照点，通过基因序列间的分化程度以及分子钟来估算速率恒定分支间分歧时间的方法。通过计算系统发育树上各个节点的发生时间，从而推断相关类群的起源时间和不同类群的分歧时间。

PAML[6]软件中的MCMCTree命令可用于估算系统发育树上各个节点的发生时间。 MCMCTree将比对好的核酸或蛋白序列和经过化石点校准的系统发生树，在不同分子钟模型下，使用贝叶斯法预估其分歧时间，该命令执行时需要程序指令的控制文件（通常称为mcmctree.ctl）。 在本分析中，使用上述系统系统发育树分析中生成的CDS-MSA以及化石时间点标定的系统发育树（化石时间可在TIMETREE(http://www.timetree.org/)网站上查询）， 从而得到物种**{}与Glycine_max在51.09个百万年前发生分歧**。其具体结果如下：

Table 3.4-1 各进化物种的化石标记点

物种1	物种2	化石时间(百万年)(Min)	化石时间(百万年)(Max)
Vitis_vinifera	*Arabidopsis_thaliana*	109.8	124.4

说明：上表为TIMETREE中查询的已知化石标记时间。

Figure 3.4-1 查询的化石标点图

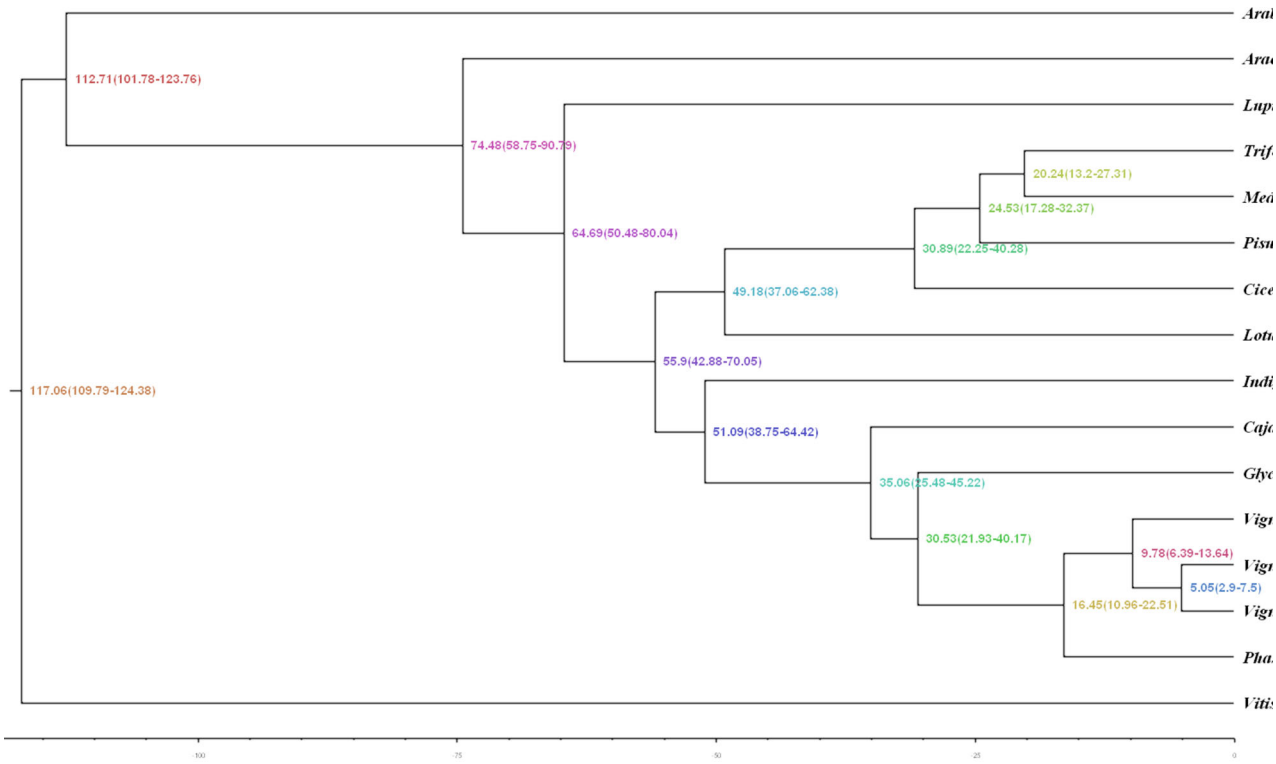


Figure 3.4-2 分歧时间估算图

说明：图中*Vitis_vinifera*为外群；各节点上数值代表距今的分歧时间(单位为百万年Million years ago, Mya)。

3.5 基因家族扩张收缩分析

物种在进化过程中由于受到选择压力，基因会出现扩张和收缩的现象。 通过对基因家族的鉴定以及与祖先数据的比较分析，得到不同物种的基因家族是否发生了大规模的扩张或收缩甚至丢失，从而推断这些物种所受的自然选择压力。 采用CAFE[7]软件，通过**16**个物种在基因家族中的数目文件、MCMCtree得到的超度量树以及软件本身计算的λ值， 从而得到不同物种在各个进化分支上的基因家族扩张收缩情况。结果显示**{}**扩张的group有**1,712**，收缩的group有**8,053**。

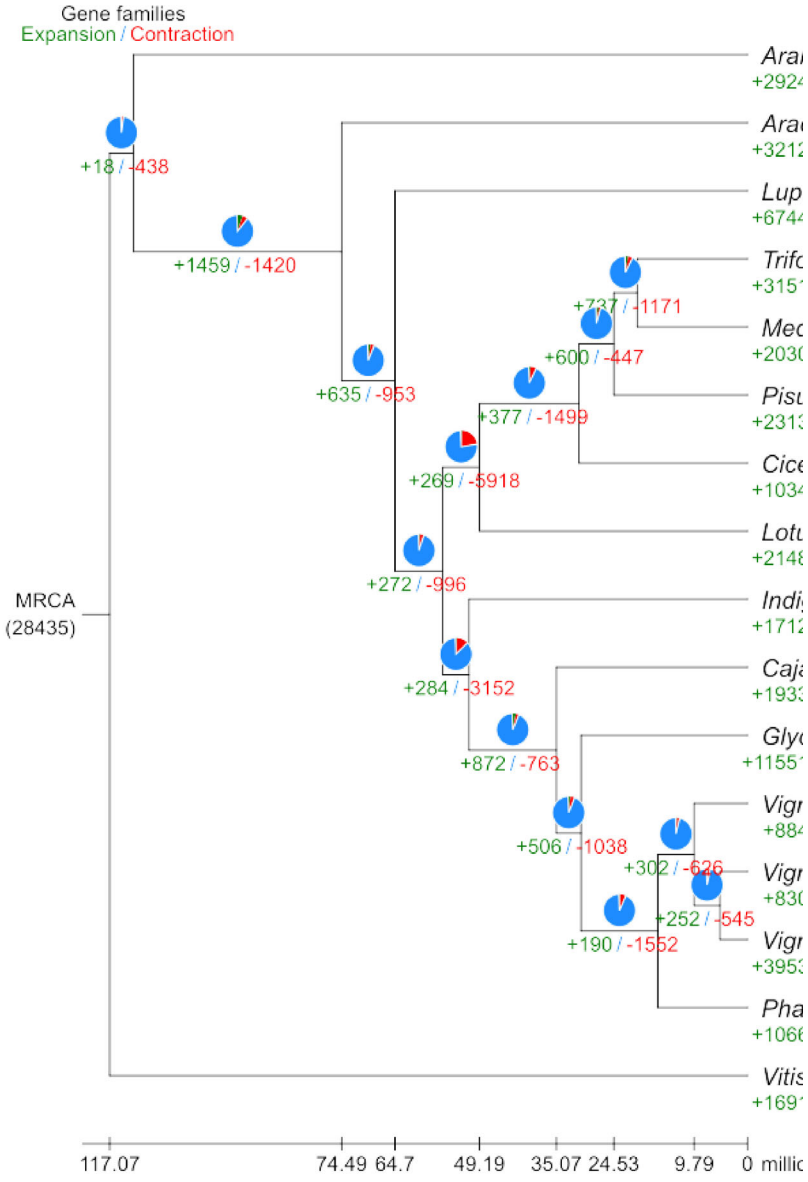
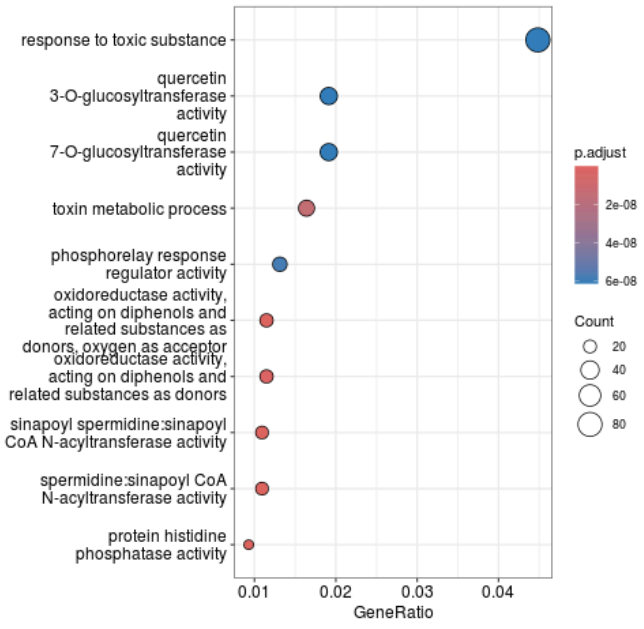


Figure 3.5-1 基因家族扩张收缩情况

3.5.1 扩张收缩家族基因GO富集分析

根据上述扩张收缩的分析结果，将目标物种中显著扩张收缩的基因家族（viterbi p<=0.05且family-wise p<=0.05）中的基因进行GO富集分析。为了更好的展示结果，选择显著富集前10的GO term进行如下绘图（详细结果见对应释放文件）。

1. 显著扩张家族基因GO富集分析：



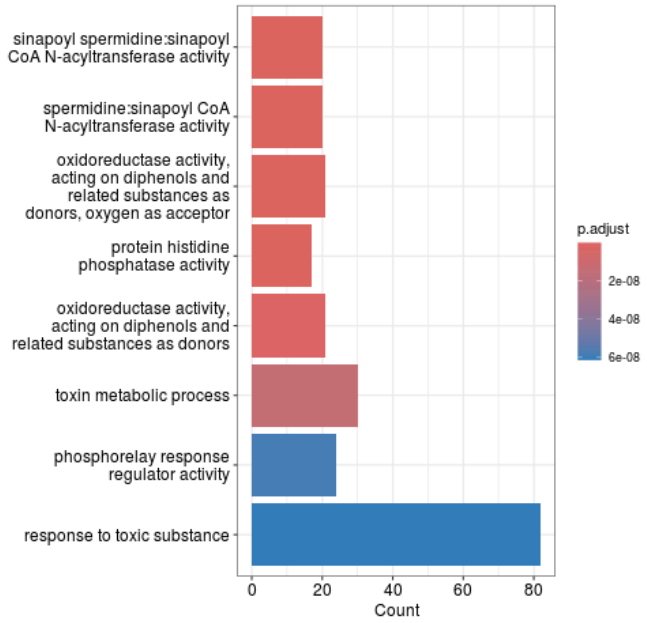


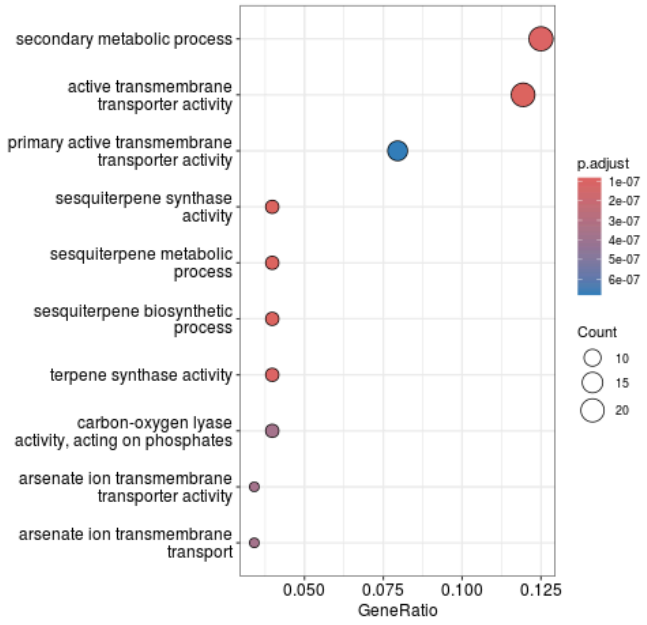
Figure 3.5-2 显著扩张家族基因GO富集散点图

说明：

散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到GO条目上的基因个数，纵坐标为GO条目。

2. 显著收缩家族基因GO富集分析：



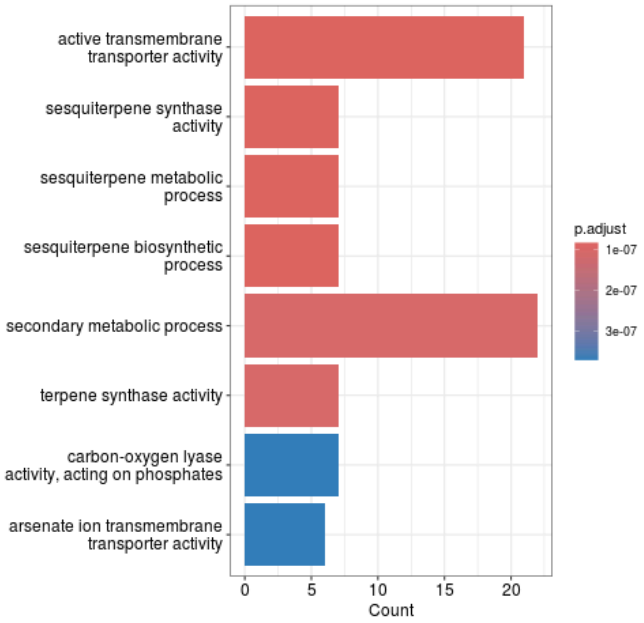


Figure 3.5-3 显著收缩家族基因GO分类柱状图

说明：

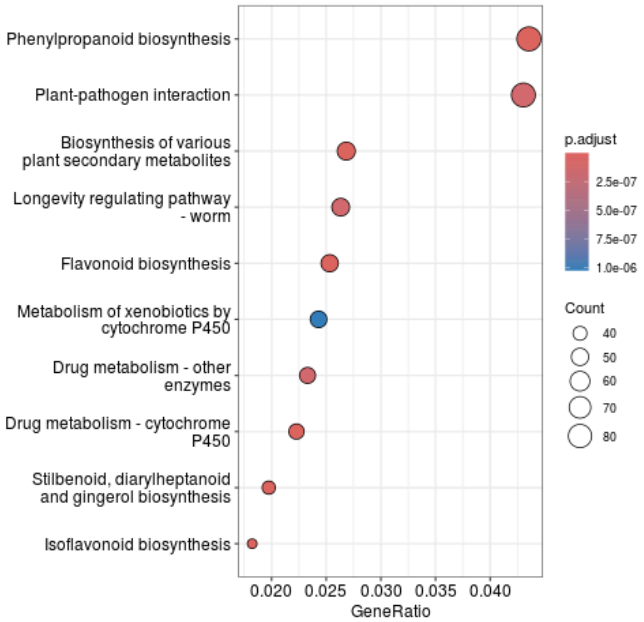
散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到GO条目上的基因个数，纵坐标为GO条目。

3.5.2 扩张收缩家族基因KEGG富集分析

根据上述扩张与收缩的分析结果，将目标物种中出现扩张和收缩的基因家族、且基于不同算法中显著变化的基因家族（viterbi p<=0.05且family-wise p<=0.05）进行KEGG富集分析，为了更好的展示结果，选择显著富集前10的KEGG Pathway进行如下绘图（详细结果见对应释放文件）。

1. 显著扩张家族基因KEGG富集分析结果：



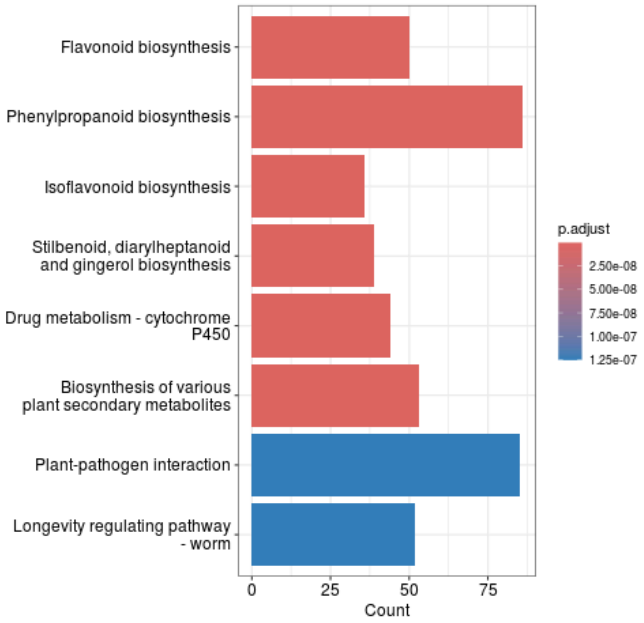


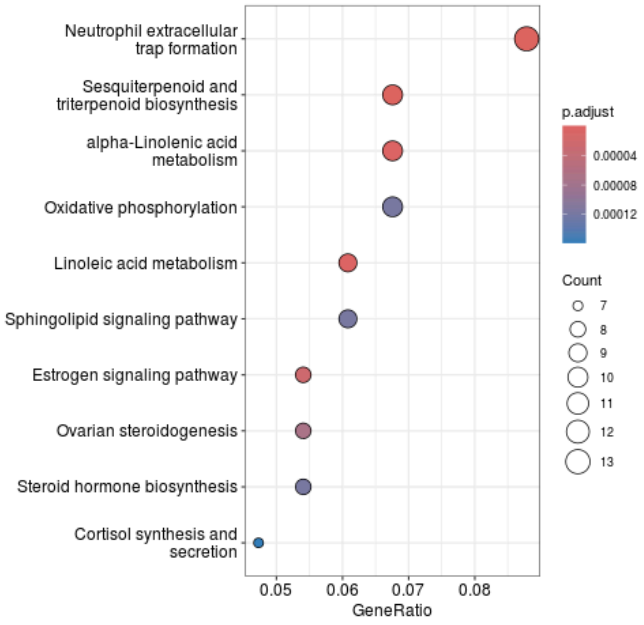
Figure 3.5-4 显著扩张家族基因KEGG富集散点图

说明：

散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到pathway的基因个数，纵坐标为KEGG pathway分类。

2. 显著收缩家族基因KEGG富集分析结果：



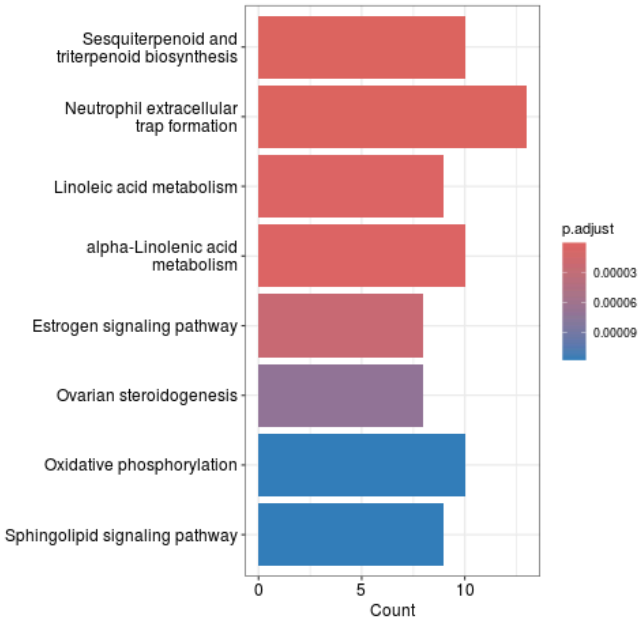


Figure 3.5-5 显著收缩家族基因KEGG分类柱状图

说明：

散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到pathway的基因个数，纵坐标为KEGG pathway分类。

3.6 正选择基因分析

自然界对物种突变的选择可以分为正负两种。当一个群体中出现能够提高个体生存力及育性的突变时，携带该突变基因的个体将比其它个体留下更多的子代， 而该突变基因最终在整个群体中扩散，这种选择称为正选择。而负选择是指群体中出现有害基因时，携带该基因的个体会因为生存力或育性降低而从群体中淘汰，也叫净化选择。

根据自然选择理论（Natural selection）和中性进化论（NeutralTheory of Molecular Evolution），对于编码蛋白的基因， 非同义替换率（nonsynonymous substitution rate , Ka）与同义替换率（synonymous substitution rate , Ks）的比值（Ka/Ks）被广泛地用来检测基因是否受到选择作用。 如果Ka/Ks =1,则认为是中性突变；如果Ka/Ks > 1,则认为此基因的进化过程受到了正向选择；而如果Ka/Ks < 1，则认为是受到了负选择。 用单拷贝基因作为输入，利用 PAML[6]软件中的codeml命令采用分支位点模型（branch-site model）来计算{}分枝上的选择压力， 各基因受选择的显著性采用卡方检验（P-value <=0.05）进行检测。结果显示正选择基因有**5**个，下表展示前5正选择基因结果（详细结果见对应释放文件）：

Table 3.6-1 正选择基因及功能

Group	P-value	Postive site number	Gene	Swissprot function
OG0013983	0.000122108	15	g25706.t1	Putative pentatricopeptide repeat-containing protein At1g74580 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=At1g74580 PE=3 SV=1
OG0014173	0.000137212	6	g10334.t1	Protein PGR OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=F28I16.80 PE=2 SV=1
OG0014357	0.000025712	9	g866.t1	Pentatricopeptide repeat-containing protein At2g17525, mitochondrial OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=At2g17525 PE=2 SV=2
OG0014745	0.000000000	13	g17671.t1	-
OG0014749	0.000008889	7	g7621.t1	-

说明：

Group：直系同源名称；

P-value：p值；

Postive site number：正选择位点数；

Gene：基因名称；

Swissprot function：在Swissprot数据库中的注释结果。

3.6.1 正选择基因GO富集分析

根据上述正选择分析结果，将其正选择基因提取出来进行GO富集分析，为更好的展示结果，选择显著富集前10的GO term进行如下绘图。

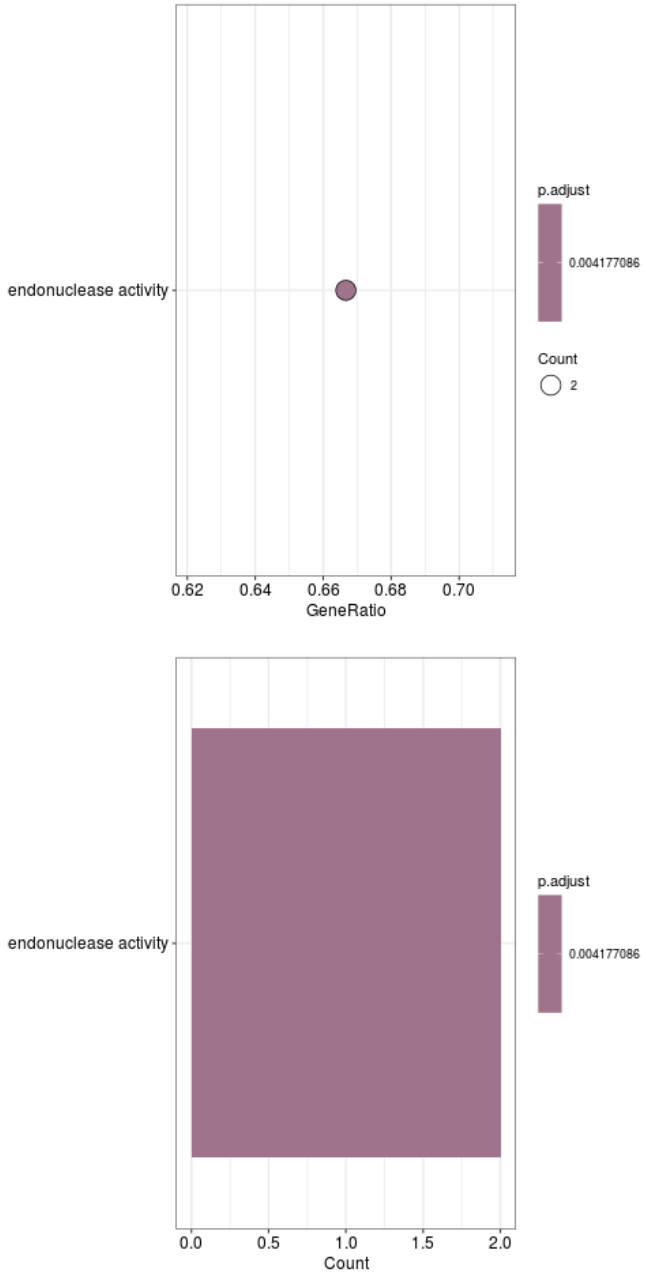


Figure 3.6-1 正选择基因 GO 富集图

说明：

散点图：横坐标为Rich Factor，表示富集上的基因占注释到的基因的百分比；纵坐标表示富集上的条目；点的大小表示富集上的基因个数，颜色表示p.adjust值大小，p.adjust值越低越显著。

柱状图：横坐标为富集到pathway的基因个数，纵坐标为KEGG pathway分类。

3.6.2 正选择基因KEGG富集分析

未检测到正选择基因KEGG富集分析结果（Figure 3.6-2 正选择基因 KEGG 富集图未生成）。

3.7 全基因组复制分析

全基因组加倍（Whole Genome Duplication，WGD，又称全基因组复制或全基因组多倍化）是生物演化史上的重要事件， 在物种起源、基因组扩张等方面有重要意义。长期以来被认为是动物、真菌、和其他生物、尤其是植物一个重要的进化动力。 为鉴定{}基因组的WGD事件，我们首先采用 Blast 软件将物种自身与自身进行蛋白比对， 然后使用 MCScanX[8] 检测自身基因的共线性区域，最后计算出同义突变率 Ks 值， 通过绘制 Ks 值的分布图便可以发现对应全基因组加倍时间的 Ks 值峰[9]。

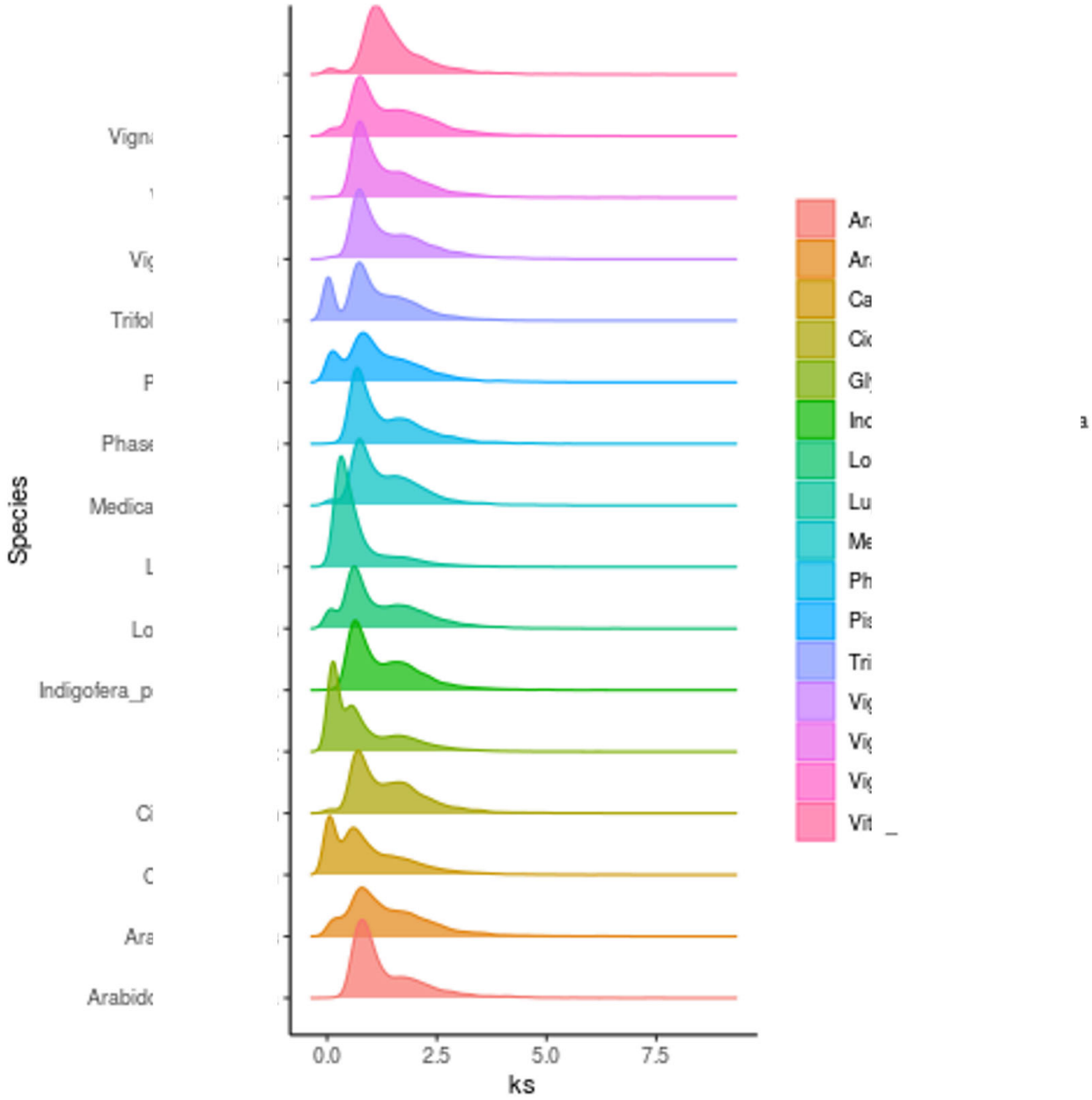


Figure 3.7-1 与其它物种Ks分布图

4 材料方法

4.1 数据处理

进行物种进化分析前，为了获得准确的分析数据，需要对所分析的数据进行预处理。 针对用于进化分析物种的基因组序列文件（FASTA格式）和基因组注释文件（GFF格式）进行预处理。 保留可变剪切信息进行后续分析，在基因家族鉴定中会导致基因家族真实信息无法确定，进而导致后续所有分析无法进行， 因此在分析前需要去掉基因本身可变剪切信息带来的影响，为了保留基因信息的完整性，选择最长的转录本用于后续分析。 从获得的最长转录本中剔除CDS（coding sequence）内部存在终止密码子（stop codon）或CDS长度非3的倍数的序列，将过滤后转录本按照物种对应密码子表（standard genetic code）翻译成蛋白序列， 得到最后的蛋白序列集用于后续基因家族分析。

4.2 基因家族分析

基因之间同源（Homology）关系的鉴定对于我们理解物种的进化和多样性至关重要，不同基因在某个历史节点可能存在共同祖先。 如果两个基因具有共同祖先，我们说它们具有同源性（Homologous）。更准确地说，同源性基因之间的关系可以划分为两种类型： （1）直系同源基因（orthologous gene），是指从同一祖先垂直进化而来的基因。或者说，一个祖先物种分化产生两种新物种，那么这两种新物种共同具有的由这个祖先物种继承下来的基因就称为直系同源基因； （2）旁系同源基因（paralogous gene），是指

由于基因复制而产生的同源基因。 通过不断的物种分化（Speciation）和基因复制事件以及随后的突变，一个基因可以演化成一个基因家族（Gene family）。 更为重要的是，基因家族中基因的缺失与获得往往与某些有趣的生物现象密切相关（比如适应性进化）。因此，进行基因同源性鉴定和基因家族的鉴定是后续分析的前提。

在本研究中，OrthoFinder[1]软件用于基因家族鉴定。 OrthoFinder使用DIAMOND进行搜索，寻找潜在的同源基因，基于基因长度和系统发育距离对得分进行标准化，通过RBNHs确定同源组序列相似度的阈值，构建直系同源组图(orthogroup graph)，最终借助MCL对基因进行聚类，划分直系同源组。

将鉴定出的单拷贝直系同源基因作为下游分析输入（系统进化树构建、估算分歧时间）。

4.3 构建系统发育树

了解不同物种的系统发生关系是进化生物学的核心课题之一，也是开展很多生物学研究的前提。 构建系统发育树的过程实际上就是利用同源性状（homologous trait）（例如MSA中的某一列位点）及同源状态（homologous state）推断祖先状态（anceteral state）的过程。

在此次分析过程中，利用分子数据构建物种系统发育树主要有2个重要步骤： （1）多序列比对（protein multiple sequence alignment，protein-MSA），目的是为了找到同源位点。 首先，使用MAFFT[2]软件对基因家族鉴定结果中单拷贝直系同源基因的蛋白序列进行多序列比对。 其次，利用PAL2NAL软件，将蛋白序列比对转换为对应的CDS序列比对（CDS-MSA）。 然后使用trimAl[10]软件对protein-MSA进行质控，使用Gblock[11]软件对CDS-MSA进行质控，去除比对质量差的位点； （2）系统发育树构建。使用串联基因构树法将单拷贝直系同源基因的比对结果串联在一起，使用IQTREE[4]，RAxML-NG[5]等软件推断物种树（species tree）； 最后，将生成的树形文件用Figtree或MEGA等软件进行编辑。

4.4 估算分歧时间

分歧时间（divergence time）估算是指以某一特定类群的化石记录作为参照点，通过基因序列间的分化程度以及分子钟来估算速率恒定分支间分歧时间的方法。 通过计算系统发育树上各个节点的发生时间，从而推断相关类群的起源时间和不同类群的分歧时间。 物种分歧时间查询工具TimeTree通过汇总大量的原始文献，包括大量的人工核实，估计出不同物种的最近共同祖先，并构建出整个所有可及物种的进化树。 本研究以TimeTree上的分歧时间为参照，使用PAML[6]软件包中MCMCTree程序推断系统发育树上各个节点的分歧时间。

4.5 基因家族扩张与收缩分析

物种在进化过程中由于受到各种选择压力，基因会出现扩张和收缩的现象，通过对基因家族的鉴定以及与祖先数据的比较分析，可以得到不同物种的基因家族是否发生了大规模的扩张或收缩甚至丢失，来推断其所受的自然选择压力，目前用于分析扩张与收缩常用的软件是CAFE[7]。

在本研究中，使用CAFE[7]软件进行基因家族扩张与收缩分析，具体过程： （1）准备输入文件，基因家族在各个物种中的数目，此结果由前面的软件OrthoFinder经统计得到；超度量树，其中枝长表示分歧时间，由前面的MCMCTree得到； （2）根据其软件模拟一个随机的出生和死亡率λ，从而来预测不同物种在各个进化分支上的基因家族进化情况。

4.6 正选择基因分析

自然对物种突变的选择大多可以分为正负两种。 当一个群体中出现能够提高个体生存力及育性的突变时，具有该基因的个体将比其它个体留下更多的子代，而突变基因最终在整个群体中扩散，这种选择称为正选择。 而负选择是指群体中出现有害基因时，携带该基因的个体会因为生存力或育性降低而从群体中淘汰，也叫净化选择。

根据自然选择理论（Natural selection）和中性进化论（NeutralTheory of Molecular Evolution），对于编码蛋白的基因， 非同义替换率（nonsynonymous substitution rate , Ka）与同义替换率（synonymous substitution rate , Ks）的比值（Ka/Ks）被广泛地用来检测基因是否受到选择作用。 如果Ka/Ks = 1,则认为是中性突变；如果Ka/Ks > 1,则认为此基因的进化过程受到了正向选择；而如果Ka/Ks < 1，则认为是受到了负选择。

在本研究中，采用PAML[6]软件中的codeml命令来估算Ka/Ks(omega)各种模型下的值， codeml命令实现了三大类模型： （1）site-Models(位点模型)，假定进化速率在密码子的各碱基中不一样。 （2）branch-Models(分枝模型)，假定进化速率在各进化分支中不一样。 （3）branch-site-Models(分支位点模型)，假定进化速率在各进化分枝和密码子各碱基中都不一样。 在我们的分析中选择branch-site-Models, 因为在物种复杂的分化过程中branch-site-Models可能更符合实际情况。 在branch-site-Models中，目标分化枝具有一个omega值，其它所有分枝具有一个相同的omega值，然后再检测正选择位点。 对下面两种模型进行比较： （1）第一种模型为model = 2，将omega值分成<1、=1、>1的三类，作为alternative model； （2）第二种模型与第一种模型一致，只是将omega固定成1，作为null model。 比较两种模型的似然差异，再利用卡方检验计算出P-value，最后对所有的P-value在全基因组范围内进行FDR校正。筛选出FDR值小于0.05的基因作为最后的候选正选择基因。

4.7 全基因组复制分析

全基因组加倍（Whole Genome Duplication，WGD，又称全基因组复制或全基因组多倍化）是生物演化史上的重要事件，在物种起源、基因组扩张等方面有重要意义。 长期以来被认为是动物、真菌、和其他生物、尤其是植物一个重要的进化动力。 4DTV(four-fold synonymous third-codon transversion)值和同义替换率Ks可以分析物种在进化史中是否发生全基因组复制事件[12]，

其中Ks可以通过程序KaKs_Calculator2.0计算得到[9]。 然后，使用R程序绘制单个物种的4DTV和Ks密度曲线图，从而反映物种在进化史中是否发生全基因组复制事件，并且通过该物种与其它物种分歧时间的比较可以明确不同物种发生全基因组复制相对时间的早晚[13]。

全基因复制分析的具体过程： （1）使用软件Blastp对各物种进行自身蛋白比对； （2）取最好的比对结果，使用软件McScanX[8]得到其共线性区段； （3）使用KaKs_Calculator2.0[9]计算同义替换率（Ks）值； （4）使用自研脚本计算4DTV； （5）使用R程序绘制4DTV和Ks的密度分布图。

4.8 富集分析

在本研究中，将基因家族分析中得到的特有基因、基因家族扩张与收缩分析中扩张和收缩的家族基因以及正选择分析中的正选择基因均进行GO和KEGG富集分析。

4.8.1 GO富集分析

Gene Ontology（简称GO）是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）来全面描述生物体中基因和基因产物的属性。 GO总共有三个ontology（本体），分别描述基因的分子功能（molecular function）、所处的细胞位置（cellular component）、参与的生物过程（biological process）。 GO的基本单位是term（词条、节点），每个term都对应一种或一类特定的功能。

GO功能显著性富集分析给出与基因组背景相比，在差异表达基因中显著富集的GO功能条目，从而给出差异表达基因与哪些生物学功能显著相关。 该分析首先把所有差异表达基因向Gene Ontology数据库（http://www.geneontology.org/）的各个term映射，计算每个term的基因数目， 然后应用卡方检验或者超几何检验，找出与整个基因组背景相比，在差异表达基因中显著富集的GO条目。超几何检验计算公式为:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

其中，N为所有基因中具有GO注释的基因数目；n为N中差异表达基因的数目；M为所有基因中注释为某特定GO term的基因数目；m为注释为某特定GO term的差异表达基因数目。 计算得到的P-value通过Bonferroni校正之后，以q-value<0.05为阈值，满足此条件的GO term定义为在差异表达基因中显著富集的GO term。通过GO功能显著性富集分析能确定差异表达基因行使的主要生物学功能。

4.8.2 KEGG富集分析

KEGG（Kyoto Encyclopedia of Genes and Genomes）是系统分析基因产物在细胞中的代谢途径（Pathway）以及这些基因产物功能的主要公共数据库，利用KEGG可以进一步研究基因在生物学上的复杂行为。

在本研究中，将基因家族分析中所得到的特有基因、收缩与扩张分析中扩张和收缩基因以及正选择分析中的正选择基因均进行GO和KEGG富集分析。

与GO富集类似，我们基于差异分析和KEGG富集的结果，应用超几何检验，找出显著富集的KEGG Pathway。

5 参考英文流程

Identification of homeologous and orthologous gene set

To identify homologous relationships among XX and other species, we downloaded their protein sequences and aligned them using OrthoFinder. Firstly, protein sets were collected from 10 sequenced species and the longest transcripts of each gene were extracted, in which miscoded genes and genes exhibiting premature termination were discarded. Proteins with no homologs in the other 10 genomes were extracted as species-specific genes including XX-specific unique genes and unclustered genes. Functional annotation of species-specific genes and the enrichment tests were performed using information from homologs in the Gene Ontology（http://www.geneontology.org/）and KEGG（Kyoto Encyclopedia of Genes and Genomes）database.

Phylogenetic analyses

On the basis of the identified orthologous gene sets with OrthoFinder, molecular phylogenetic analysis was performed using the shared single-copy genes. Briefly, the coding sequences were extracted from the single-copy families and each ortholog group were multiple aligned using Mafft. Poorly aligned sequences were then eliminated using Gblocks and the RAxML-NG was used for the phylogenetic tree construction with 100 bootstrap replicates. The generated tree file was displayed with Figtree/MEGA. Based on the phylogenetic tree, MCMCtree was utilized to compute the mean substitution rates along each branch and estimate the species divergent time. Three fossil calibration times were obtained from the TimeTree database (http://www.timet ree.org/) as the time control, including divergence times of A. thaliana (148-173 Mya), M. acuminate (90-115 Mya), and O. sativa (40-53 Mya).

Gene family expansion and contraction analysis

Significant expansion or contraction of specific gene families is often associated with adaptive divergence of closely related species. According to the results of OrthoFinder, expansions and contractions of orthologous gene families were then detected using CAFE which uses a birth and death process to model gene gain and loss over a phylogeny.

Genes under positive selection

According to the neutral theory of molecular evolution, the ratio of nonsynonymous substitution rate (Ka) and synonymous substitution rate (Ks) of protein coding genes can be used to identify genes that show signatures of natural selection. We thus calculated average Ka/Ks values and conducted the branch-site likelihood ratio test using Codeml implemented in the PAML package to identify positively selected genes in the XX lineage. Genes with p value <0.05 under the branch-site model were considered positively selected genes.

Screening for whole genome duplication events

Four-fold synonymous third-codon transversion (4DTv) and synonymous substitution rate (Ks) estimation were used to detect whole-genome duplication (WGD) events in the XX genome. Firstly, protein sequences of XX were extracted and all-vs-all paralog analysis were performed using best hits from primary protein sequences by self-BLASTp in these plants. After filtering by identity and coverage, the BLASTP results were then subjected to MCScanX and the respective collinear blocks were thus identified. Finally, the Ks and 4DTV were then calculated for the syntenic blocks gene pairs using KaKs_Calculator and potential WGD events in each genome were evaluated based on their ks and 4DTv distribution.

#此内容为通用流程，如需用于论文发表，请根据实际内容进行修正并注意语言修改

6 软件及参数

Software	Version	Parameter	Website
OrthoFinder	v2.5.5	default	https://github.com/davidemms/OrthoFinder/releases
MAFFT	v7.5.25	default	https://mafft.cbrc.jp/alignment/software/
PAL2NAL	v14	default	https://www.bork.embl.de/pal2nal/
IQ-TREE	v2.3.2	default	http://www.iqtree.org/
RAxML-NG	v1.2.2	default	https://github.com/amkozlov/raxml-ng
PAML	v4.10.7	default	https://github.com/abacus-gene/paml
CAFE	v4.2.1	-p 0.05 -t 10 -r 10000	https://github.com/hahnlab/CAFE
MCScanX	-	default	https://github.com/wyp1125/MCScanX
KaKs_Calculator2.0	v2.0	-m NG	https://sourceforge.net/projects/kakscalculator2
trimAl	v1.4.1	-automated1	https://sourceforge.net/projects/kakscalculator2
Gblocks	0.91b	-t=c -b5=h	http://molevol.cmima.csic.es/castresana/Gblocks.html

7 参考文献

1. Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20:238
2. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-780.
3. Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic acids research, 34(Web Server issue), W609–W612.
4. L. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. and Evol., 32:268-274.

5. Alexey M. Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis (2019) RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics, 35 (21), 4453-4455
6. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.
7. De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22: 1269-1271.
8. Wang Y, Tang H, Debarry JD, Tan X, Li J, et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40: e49.
9. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics 8: 77-80.
10. Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England), 25(15), 1972–1973.
11. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540-552.
12. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120.
13. Guo L, Winzer T, Yang X, Li Y, Ning Z, et al. (2018) The opium poppy genome and morphinan production. Science 362: 343-347

8 联系我们

西安浩瑞基因技术有限公司成立于2019年12月，专注于开发三代单分子测序技术在基因组学研究中的应用以及在精准医学临床中的转化。公司扎根于丝绸之路的起点——西安，在沣东新城"秦创原"创新驱动平台的大力支持下，利用地区优势汇聚和打造了一支专业的技术团队，目标成为涵盖三代测序技术的研发、检测、生产的全产业链闭环平台。目前，公司已引进7台PacBio Sequel II 测序仪和1台最新的PacBio Revio测序仪，并配备了完善的三代测序配套设备。同时，在PacBio公司的支持下，建立了完整的技术体系，即标准化的实验流程、独立的机械设备工程师，独立的技术应用工程师，生物信息专业团队等。团队配置以及硬件设施在国内属于领先水平，尤其在西部地区更是独树一帜，为三代测序平台的稳定运转打下了坚实的基础。

联系方式

热线电话：+86 029-89303503

官方网站：www.xahorizon.cn

邮 箱：project@xahorizon.cn

地 址：陕西省西安市沣东新城中兴深蓝科技产业园A区2号楼3层