植物学报 *Chinese Bulletin of Botany* 2021, **56** (6): 699–714, www.chinbullbotany.com
doi: 10.11983/CBB21100

·研究报告·

基于转录组数据揭示4种兜兰的全基因组复制历史

王蒙[1], 王婷[1, 2], 夏增强[1, 3], 李廷章[1], 金效华[4], 严岳鸿[1], 陈建兵[1*]

[1]深圳市兰科植物保护研究中心, 兰科植物保护与利用国家林业和草原局重点实验室, 深圳 518114
[2]西南林业大学生物多样性保护学院, 昆明 650224; [3]中国科学院分子植物科学卓越创新中心, 上海 200032
[4]中国科学院植物研究所, 系统与进化植物学国家重点实验室, 北京 100093

摘要 多倍化或全基因组复制(WGD)是物种多样性发生的重要驱动力。目前, 在蕨类、菊科以及豆科等类群丰富的植物中已多次报道全基因组复制事件, 而兰科(Orchidaceae)全基因组复制事件报道极少, 与其丰富的物种多样性存在矛盾, 推测与前期样本量小但类群跨度大的研究策略有关。选取染色体数目变异丰富且多样性较高的兜兰属(*Paphiopedilum*)为兰科植物代表类群, 基于共享数据库中4种兜兰的转录组数据, 采用同义替换率(K_S)、系统发生基因组学以及相对定年的方法分析兜兰属植物是否发生过全基因组复制事件。结果表明, 在4种兜兰中均检测到3次全基因组复制事件, 分别发生在110.17–119.77 Mya (WGD1)、60.95–74.19 Mya (WGD2)和38.19–45.85 Mya (WGD3)。其中, WGD3为新检测到的全基因组复制事件, 推测其发生在杓兰亚科(Cypripedioideae)与姐妹类群分化后, 兜兰属与姐妹类群分化之前。此外, 3次全基因组复制事件发生后优先保留的基因拷贝在功能上多与当时的环境胁迫响应相关, 推测全基因组复制提高了兜兰属植物祖先对当时极端环境变化的适应性。

关键词 多倍化, 兰科, 杓兰亚科, 适应性演化, 同义替换率

王蒙, 王婷, 夏增强, 李廷章, 金效华, 严岳鸿, 陈建兵 (2021). 基于转录组数据揭示4种兜兰的全基因组复制历史. 植物学报 **56**, 699–714.

多倍化(polyploid)或全基因组复制(whole-genome duplication, WGD)是物种多样性发生的重要驱动力(De Bodt et al., 2005; Van de Peer et al., 2017; Mandáková and Lysak, 2018), 在植物演化历史中普遍存在, 尤其是维管束植物中多样性最高的类群被子植物和第二大类群蕨类反复发生过多轮全基因组复制(One Thousand Plant Transcriptomes Initiative, 2019; 汪浩等, 2019; Huang et al., 2020; 王婷等, 2021)。基于现有证据, 在蕨类植物、被子植物第一大科菊科(Asteraceae)、第三大科豆科(Fabaceae)中分别检测到19、41、28次全基因组复制事件(Huang et al., 2020; Zhang et al., 2021a; Zhao et al., 2021), 推测多倍化与蕨类植物和被子植物物种多样性较高类群的物种形成和多样化有关(De Bodt et al., 2005; Van de Peer et al., 2017; Mandáková and Lysak, 2018; Ren et al., 2018)。

全基因组复制使得染色体和基因组内全部基因均发生加倍, 为新性状演化和物种多样化提供了遗传材料(De Bodt et al., 2005; Wu et al., 2020)。而全基因组复制后的基因丢失、沉默、亚功能化和新功能化等基因水平的变异, 以及染色体重组等染色体水平变异促进了表型和物种的多样化(Wendel, 2000; Adams and Wendel, 2005; Mandáková and Lysak, 2018)。此外, 全基因组复制及后续变异导致一些类群染色体数目变异(Mandáková and Lysak, 2018)。以单子叶植物禾本科为例, 禾本科祖先的染色体基数为7条, 而在经历了全基因组复制事件后(Paterson et al., 2004; Salse et al., 2008), 水稻(*Oryza sativa*)、高粱(*Sorghum bicolor*)和谷子(*Setaria italica*)的染色体基数并未达到加倍后的14条, 而是表现为染色体数目不同程度地减少, 分别为12、10和9条(Murat et al., 2017; 王振怡和王希胤, 2020)。因此, 染色体数目的变化是

收稿日期: 2021-06-22; 接受日期: 2021-11-24
基金项目: 中央林业改革发展资金(粤财资环[2019]5号)
* 通讯作者。E-mail: cjb@cnocc.cn

全基因组复制发生及后续演化进程的重要特征之一。

兰科(Orchidaceae)含700余属、约26 000种，为被子植物第二大科，单子叶植物第一大科，是陆生植物中极具多样性的类群之一(Chase et al., 2015; Li et al., 2016)，同时表现出染色体数目变化较大(染色体基数从*x*=6到*x*=120)的特点(Da Conceição et al., 2006; 王筠竹等, 2019)，表明兰科植物的演化过程可能存在多次全基因组复制事件。然而，目前在兰科植物中已见报道的全基因组复制事件非常有限。基于兰科植物基因组证据(Cai et al., 2015; Zhang et al., 2016, 2017; Yuan et al., 2018; Hasing et al., 2020)以及千种植物转录组项目等转录组分析(One Thousand Plant Transcriptomes Initiative, 2019)，目前仅检测到1次兰科植物特异发生的全基因组复制事件，与蕨类(Huang et al., 2020)、菊科(Zhang et al., 2021a)和豆科(Zhao et al., 2021)等物种多样性丰富的类群多倍化研究结果不符。

分析上述情况的原因，我们推测可能与兰科植物种类及类群众多、前期研究样本量小但种类跨度大的研究策略有关。例如，千种植物转录组项目囊括了兰科7个样本，但却跨了香荚兰亚科(Vanilloideae)、兰亚科(Orchidoideae)和树兰亚科(Epidendroideae) 3个亚科7个属(One Thousand Plant Transcriptomes Initiative, 2019)；分析全基因组复制事件的5套全基因组数据同样覆盖了拟兰亚科(Apostasioideae)、香荚兰亚科、树兰亚科3个亚科5个属(Cai et al., 2015; Zhang et al., 2016, 2017; Yuan et al., 2018; Hasing et al., 2020)；关于杓兰亚科基因组进化的研究包括13个兰科植物转录组和基因组数据，覆盖了兰科所有亚科(拟兰亚科、香荚兰亚科、杓兰亚科(Cypripedioideae)、兰亚科和树兰亚科) 13个属(Unruh et al., 2018)。对于兰科这样包含26 000多种的特大类群，解析其全基因组复制历史需要借助更精细的尺度。

杓兰亚科具有囊状或倒盔状唇瓣、2个可育雄蕊和1个盾状退化雄蕊等特征，是兰科多样性的重要代表类群之一，包括杓兰属(*Cypripedium*)、南美杓兰属(*Selenipedium*)、美洲兜兰属(*Phragmipedium*)、镊萼兜兰属(*Mexipedium*)及兜兰属(*Paphiopedilum*) 5个属(Cox et al., 1997; Chen et al., 2009)。其中，兜兰属是杓兰亚科最大的属，约100多种，占杓兰亚科总物种数一半以上(Govaerts et al., 2021)。兜兰属植物的基因组普遍较大并存在一定程度的变异(16.5–35.9 pg/C)，且染色体数目变异丰富(2*n*=26–42) (Leitch et al., 2009)。因此，我们推测兜兰属可能存在全基因组复制事件，然而，在过去样本量小、跨大尺度的研究中并未检测到兜兰属特异发生的全基因组复制事件。因此，本研究基于NCBI共享数据，即杏黄兜兰(*Paphiopedilum armeniacum*)、同色兜兰(*P. concolor*)、带叶兜兰(*P. hirsutissimum*)以及麻栗坡兜兰(*P. malipoense*)的转录组数据，采用经典的同义替换率(K_s)、系统发生基因组学以及相对定年的方法对其进行全基因组复制事件检测，进而开展以下研究：(1) 过去未检测到全基因组复制事件历史的兜兰属植物是否发生了全基因组复制事件；(2) 若发生了全基因组复制事件，进一步分析其发生时间，以及是否为兜兰属内发生的全基因组复制事件；(3) 全基因组复制事件的发生对于兜兰属植物适应性演化的意义。

1 材料与方法

1.1 测序数据下载

从NCBI网站SRA数据库检索下载杏黄兜兰(*Paphiopedilum armeniacum* S.C.Chen & F.Y.Liu) (2*n*=26)、同色兜兰(*P. concolor* (Lindl. ex Bateman) Pfitzer) (2*n*=26)、带叶兜兰(*P. hirsutissimum* (Lindl. ex Hook.) Stein) (2*n*=26)以及麻栗坡兜兰(*P. malipoense* S.C. Chen & Z.H.Tsi) (2*n*=26)转录组测序的原始数据(raw data) (Cox et al., 1998; 杨志娟, 2006; Li et al., 2014; Zhang et al., 2017; Fang et al., 2020)，用于后续的组装与分析。同时，从NCBI网站Genome数据库下载深圳拟兰(*Apostasia shenzhenica* Z.J.Liu & L.J. Chen)基因组数据(GCA_002786265.1) (Zhang et al., 2017)用于物种间直系同源基因的K_s分析。将拟兰作为基于系统发生基因组学检测全基因组复制事件的外类群。

1.2 测序数据提取和质控

借助SRA Toolkit v2.10.8中的fastq-dump命令从原始数据中提取获得fastq文件，参数为--gzip --split-e (https://github.com/ncbi/sra-tools) (Leinonen et al., 2011)。利用Trimmomatic v0.39软件对fastq文件进行质控处理(参数设置：PE ILLUMINACLIP: TruSeq3-

PE.fa:2: 30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:15 MINLEN:50 TOPHRED33) (Bolger et al., 2014), 过滤去除接头序列及低质量碱基等, 获取高质量数据(clean data)用于后续组装。

1.3 转录组组装和质量评估

采用Trinity v2.11.0对高质量数据进行*de novo*组装(Haas et al., 2013), 参数设置为--seqType fq --min_kmer_cov 2 --normalize_reads --bflyCalculateCPU。随后, 利用cd-hit v4.8.1将相似性≥95%的转录本(transcript)聚为一组(参数: -c 0.95), 每一组聚类中输出最长序列, 得到非冗余的单基因簇(unigene) (Li and Godzik, 2006)。基于embryophyta_odb10数据库, 利用BUSCO v4.0.6软件对组装获得的转录本进行完整性评估(Simão et al., 2015)。

1.4 蛋白编码区及转录因子预测

在默认设置条件下, 利用TransDecoder v5.5.0对unigene序列进行蛋白编码区预测(https://github.com/TransDecoder/TransDecoder/releases/tag/TransDecoder-v5.5.0), 获得蛋白编码序列(protein coding sequence, CDS)和相应的蛋白序列。利用iTAK软件, 基于软件内置数据库进行植物转录因子(transcription factors, TFs)预测(Zheng et al., 2016)。

1.5 全基因组复制事件检测

根据文献报道的方法计算物种内旁系同源基因对的K_s值(Sollars et al., 2017), 并对其进行正态分布拟合, 以检测全基因组复制事件。首先, 分别对各物种的蛋白序列进行all against all序列相似性比对(BLASTP), 阈值设置为e^{-5}。然后, 应用脚本KSPlotter.py计算每个物种的K_s值(https://github.com/EndymionCooper/KSPlotting)。主要步骤为: 使用mcl-blastline pipeline构建基因家族(Enright et al., 2002), 借助MUSCLE对每个基因家族进行比对(Edgar, 2004), 最后利用CODEML软件(PAML包)计算每个物种的K_s值(Goldman and Yang, 1994; Yang, 2007)。为避免随机误差和同义替换饱和效应的影响(Blanc and Wolfe, 2004; Schlueter et al., 2004; Cui et al., 2006), 本研究仅保留0.1–5之间的K_s值用于后续分析。最后, 借助R包mclust中的高斯混合模型对保留的K_s值进行正态分布拟合(Scrucca et al., 2016), 以排除假阳性峰。

为分析全基因组复制事件与类群分化间的时间关系, 利用wgd软件, 采用wf2流程分别计算4种兜兰与深圳拟兰、3种兜兰与杏黄兜兰(位于兜兰基部类群)间直系同源基因的K_s值(Zwaenepoel and Van De Peer, 2019)。若物种内旁系同源基因的K_s峰值(代表全基因组复制事件)小于物种间直系同源基因的K_s峰值(代表类群分化事件), 则认为全基因组复制事件发生在类群分化事件后; 反之, 则认为全基因组复制事件发生在类群分化事件前(One Thousand Plant Transcriptomes Initiative, 2019)。

为验证K_s法检测结果的准确性, 应用tree2gd软件, 基于系统发生基因组学的方法再次检测全基因组复制事件(Zhang et al., 2020; Zhao et al., 2021)。(1) 以4种兜兰和深圳拟兰的蛋白序列为输入文件, 利用OrthoFinder v2.5.2筛选单拷贝直系同源基因(Emms and Kelly, 2019)。(2) 利用单拷贝直系同源基因构建物种树。首先, 采用MUSCLE v3.8.31对筛选得到的302个单拷贝直系同源基因进行多序列比对(Edgar, 2004); 随后, 基于比对结果使用Gblocks v0.91b筛选保守区域(Castresana, 2000; Talavera and Castresana, 2007), 并将筛选获得的保守区域串联形成多基因矩阵; 最后, 以ProtTest v3.4.2确定的PROTGAMMAJTTF为最优替代模型(Darriba et al., 2011), 利用RAxML v8.2.12软件, 采用最大似然法、基于保守序列矩阵、以深圳拟兰为外类群、在自举检验1 000次的设置下构建系统发生树(Stamatakis, 2014)。(3) 以第(2)步构建的系统发生树为物种树, 利用tree2gd v1.0.39软件, 基于默认参数检测全基因组复制事件(https://github.com/Dee-chen/Tree2gd) (Zhang et al., 2020)。

1.6 全基因组复制事件相对定年

基于同义替换速度恒定的假定前提, 根据物种内旁系同源基因K_s分布的峰值和公式$T=K_s/2r$, 采用深圳拟兰的绝对定年时间, 推算4种兜兰全基因组复制事件的发生时间(Badouin et al., 2017; Zhang et al., 2017)。先依据深圳拟兰的绝对定年信息(K_s=1, T=74 Mya) (Zhang et al., 2017)和公式$T=K_s/2r$, 推算出深圳拟兰的$r=6.76\times10^{-9}$ (同义替换/位点/年); 然后根据

正态分布拟合得到的K_s峰值，采用深圳拟兰的r值，推算4种兜兰全基因组复制事件的发生时间。

1.7 转录组功能注释和复制基因功能富集分析

首先，分别对每种兜兰的所有蛋白序列进行功能注释。使用eggNOG-mapper v2.0.6软件，基于eggNOG v5.0.1数据库对预测获得的蛋白序列进行功能注释(Huerta-Cepas et al., 2017, 2019)，注释结果用于分析保留复制基因的功能富集。然后，根据高斯混合模型拟合显著存在的峰值，分别提取4种兜兰各峰值95%置信区间的基因作为全基因组复制事件中保留的复制基因，对其进行GO功能富集分析。借助R包AnnotationForge，基于4种兜兰的功能注释结果，为每个物种分别构建数据库(https://bioconductor.org/packages/AnnotationForge/)；利用clusterProfiler分别对每种兜兰各全基因组复制事件中保留的复制基因进行GO功能富集分析(P<0.05) (Yu et al., 2012)。GO功能富集结果采用R包ggplot2 (https://github.com/tidyverse/ggplot2)和pheatmap (https://github.com/raivokolde/pheatmap)进行可视化。

2 结果与讨论

2.1 原始数据下载、组装和质量评估

从NCBI网站下载麻栗坡兜兰、同色兜兰、带叶兜兰以及杏黄兜兰的转录组原始数据，用于测序的组织分别为茎、叶或种子，原始数据量为3–14.1 Gb，总数据量为28.7 Gb。对获得的原始数据进行提取、质控及组装，分别组装得到76 006 (带叶兜兰) –239 105 (麻栗坡兜兰)个转录本，去冗余后对应获得62 565–201 606个unigenes。具体信息见表1。

为评估组装的完整性，基于包含1 614个单拷贝基因的embryophyta_odb10数据库进行BUSCO评估，完整覆盖基因的比例(complete BUSCOs)分别为麻栗坡兜兰(94.0%)>杏黄兜兰(92.5%)>同色兜兰(88.5%)>带叶兜兰(86.9%) (图1)。BUSCO评估结果显示，组装完整性较高，可用于后续分析。

2.2 蛋白编码区和转录因子预测

对4种兜兰转录组的unigenes序列进行蛋白编码区预测，共预测到228 075个CDS，其中带叶兜兰获得CDS的数量最少(33 207个)，平均长度最长(994.9 bp)；麻栗坡兜兰获得CDS的数量最多(79 854个)，平均长度最短(829.1 bp) (表2)。对4种兜兰的CDS进行转录因子预测，分别预测到1 181(带叶兜兰)–2 586 (麻栗坡兜兰)个转录因子，共预测到7 731个转录因子，归属于70个转录因子家族(附表1；表2)。

2.3 全基因组复制事件检测

基于预测的蛋白序列，分别计算4种兜兰种内旁系同源基因的K_s值，K_s值密度分布如图2所示。后续采用高斯混合模型对K_s值进行正态分布拟合(图2；表3)。结果显示，麻栗坡兜兰、同色兜兰以及杏黄兜兰均拟合到9个K_s峰，带叶兜兰拟合到7个K_s峰；同时，4种兜兰在K_s为1.5、0.8以及0.5左右均以较大比例拟合到3个K_s峰(分别为WGD1、WGD2和WGD3)，推测4种兜兰均至少经历了3次全基因组复制事件。

除计算物种内旁系同源基因的K_s值外，我们还分

表1 转录组原始数据和*de novo*组装结果

Table 1 The statistics of raw data and *de novo* assembly

	Paphiopedilum concolor	*P. hirsutissimum*	*P. malipoense*	*P. armeniacum*
Accession number	SRR1405683	SRR1405685	SRR5722160	SRR9842184
Tissues	Leaf	Leaf	Stem	Seed
Bases (Gb)	3.6	3	14.1	8
Number of transcripts	156581	76006	239105	164515
Average length of transcript (bp)	907.1	1162.3	884.5	993.2
N50 of transcript (bp)	1486	1971	1627	1856
Number of unigenes	116919	62565	201606	139203
Average length of unigene (bp)	863.3	1071.1	815.6	906.4
N50 of unigene (bp)	1438	1829	1480	1704
Source of raw data	Li et al., 2014	Li et al., 2014	Zhang et al., 2017	Fang et al., 2020

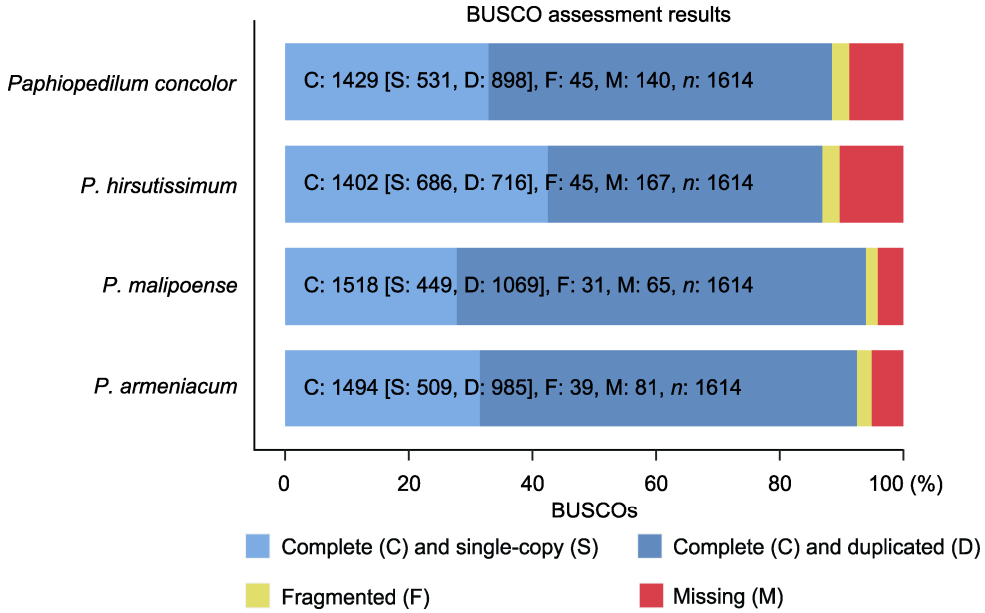

图1 BUSCO评估结果

C: 完整覆盖的基因数; S: 完整覆盖且为单拷贝的基因数; D: 完整覆盖且为多拷贝的基因数; F: 未完整覆盖的基因数; M: 缺失基因数

Figure 1 BUSCO assessment results

C: Complete BUSCOs; S: Complete and single-copy BUSCOs; D: Complete and duplicated BUSCOs; F: Fragmented BUSCOs; M: Missing BUSCOs

表2 蛋白编码区和转录因子预测结果

Table 2 The summary of protein coding sequence and transcription factor prediction

	Paphiopedilum concolor	*P. hirsutissimum*	*P. malipoense*	*P. armeniacum*
Number of protein coding sequences	56439	33207	79854	58575
Average length of protein coding sequence (bp)	936.1	994.9	829.1	914.7
N50 of protein coding sequence (bp)	1209	1308	1089	1215
Number of CDS identified as transcription factor	1950	1181	2586	2014
Number of transcription factor families	66	67	67	68

别计算了4种兜兰与兰科基部代表物种深圳拟兰、3种兜兰与兜兰基部代表物种杏黄兜兰间直系同源基因的K_s值(图2), 结果发现4种兜兰与深圳拟兰间直系同源基因的K_s峰值均小于全基因组复制事件WGD1和WGD2的K_s峰值, 大于全基因组复制事件WGD3的K_s峰值, 推测全基因组复制事件WGD1和WGD2发生在兜兰属与深圳拟兰分化事件之前, 而WGD3则发生在之后。杏黄兜兰与其它3种兜兰间直系同源基因的K_s峰值均小于全基因组复制事件WGD1、WGD2和WGD3, 推测3次全基因组复制事件均发生在兜兰属分化之前。

应用tree2gd v1.0.39软件, 基于系统发生基因组学方法再次进行全基因组复制事件检测, 判定标准参照Zhang等(2020)所述方法。满足以下任一条件则认为发生了全基因组复制事件: (1) 复制基因(gene duplication, GD) >500个, 其中(AB)(AB)类型的复制基因>250个; (2) 复制基因>1 500个, 其中(AB)(AB)类型的复制基因>100个, 且(AB)(AB)类型的复制基因与(AB)A类型或(AB)B类型的复制基因之和>1 000个。tree2gd分析结果(图3)表明, 4种兜兰的祖先(即结点2)保留了556个复制基因, 其中(AB)(AB)类型的复制基因为274个, 满足全基因组复制事件的判定条件, 推测在兜兰属与深圳拟兰分化之后、兜兰属分化之前(即结点3与结点2之间)发生了1次全基因组复制事件,

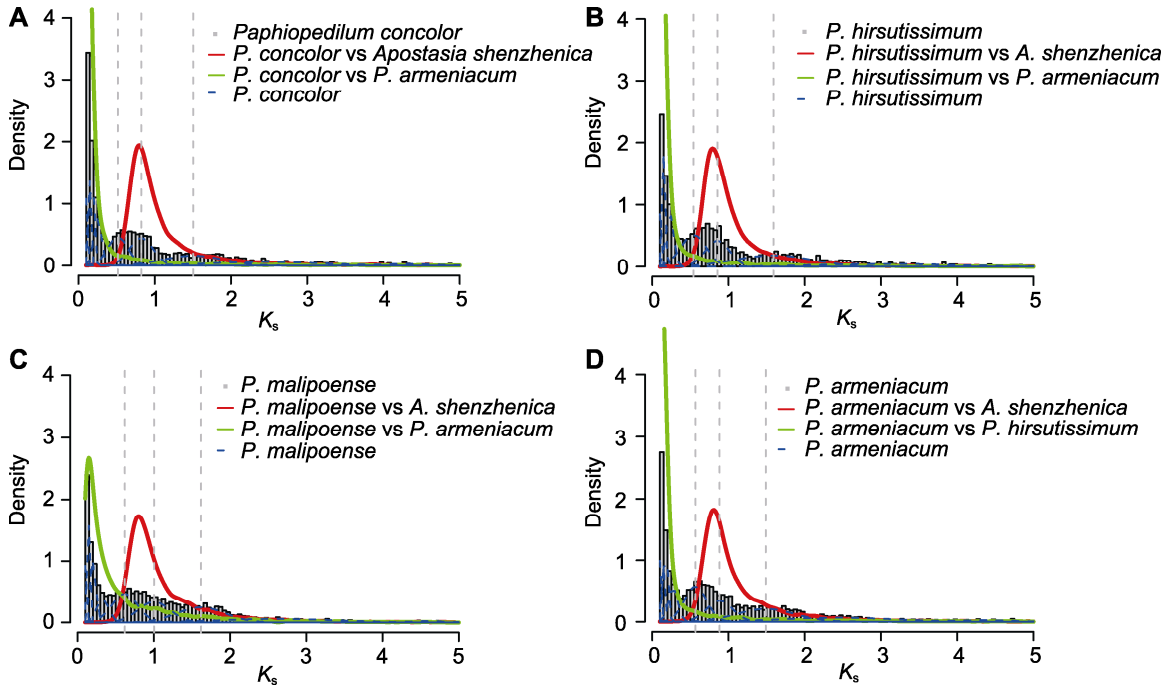

图2 4种兜兰的K_s密度分布

灰色直方图: 物种内旁系同源基因的K_s密度分布; 红色曲线: 4种兜兰与深圳拟兰间直系同源基因的K_s密度分布; 绿色曲线: 杏黄兜兰与其它3种兜兰间直系同源基因的K_s密度分布; 蓝色虚线: 物种内旁系同源基因K_s值的高斯混合模型拟合结果; 灰色虚线: 高斯混合模型显著拟合到的K_s峰值。

Figure 2 The density plot of K_s from four species of *Paphiopedilum*

Histograms filled in grey: The density distributions of intraspecies paralogue K_s values; Red solid curves: The K_s density plots of interspecies orthologues between four species of *Paphiopedilum* and *Apostasia shenzhenica*; Green solid curves: The K_s density plots of interspecies orthologues between *P. armeniacum* and other three species of *Paphiopedilum*; Blue dashed curves: The fitting results based on Gaussian mixture modeling of intraspecies paralogue K_s values; Grey dashed lines: The K_s values of significant peaks identified by Gaussian mixture modeling.

表3 基于高斯混合模型拟合的K_s结果

Table 3 The K_s value based on Gaussian mixture modeling

Species	No. of components	No. of duplicates	BIC	Variance	Mean (K_s)	Proportion
Paphiopedilum concolor	9	207	–4673.731	0.0000	0.1077	0.0527
	9	385	–4673.731	0.0002	0.1314	0.1048
	9	416	–4673.731	0.0007	0.1740	0.1194
	9	311	–4673.731	0.0026	0.2517	0.0964
	9	522	–4673.731	0.0175	0.5161	0.1544
	9	642	–4673.731	0.0241	0.8236	0.1741
	9	567	–4673.731	0.1557	1.5043	0.1594
	9	300	–4673.731	0.5765	2.4529	0.1148
	9	90	–4673.731	0.1277	4.3036	0.0240
P. hirsutissimum	7	196	–4604.688	0.0001	0.1146	0.0632
	7	277	–4604.688	0.0005	0.1504	0.0991
	7	290	–4604.688	0.0026	0.2292	0.1109
	7	558	–4604.688	0.0282	0.5407	0.2162
	7	508	–4604.688	0.0260	0.8544	0.1693
	7	585	–4604.688	0.2594	1.5894	0.2351
	7	236	–4604.688	0.8550	3.0527	0.1062

表3 (续)
Table 3 (continued)

Species	No. of components	No. of duplicates	BIC	Variance	Mean (K_s)	Proportion
P. malipoense	9	377	–14027.68	0.0000	0.1081	0.0399
	9	751	–14027.68	0.0003	0.1362	0.0890
	9	820	–14027.68	0.0016	0.2006	0.1005
	9	579	–14027.68	0.0067	0.3290	0.0768
	9	1611	–14027.68	0.0287	0.6196	0.1983
	9	1464	–14027.68	0.0447	1.0026	0.1778
	9	1634	–14027.68	0.1002	1.6186	0.1952
	9	683	–14027.68	0.5901	2.6205	0.1105
	9	105	–14027.68	0.0568	4.5773	0.0121
P. armeniacum	9	206	–8261.607	0.0000	0.1063	0.0359
	9	472	–8261.607	0.0002	0.1296	0.0870
	9	478	–8261.607	0.0008	0.1744	0.0950
	9	400	–8261.607	0.0039	0.2729	0.0849
	9	1089	–8261.607	0.0213	0.5664	0.2105
	9	778	–8261.607	0.0287	0.8825	0.1457
	9	999	–8261.607	0.1443	1.4888	0.1980
	9	455	–8261.607	0.5375	2.4393	0.1195
	9	120	–8261.607	0.1481	4.3256	0.0234

BIC: 贝叶斯信息标准 BIC: Bayesian information criterion

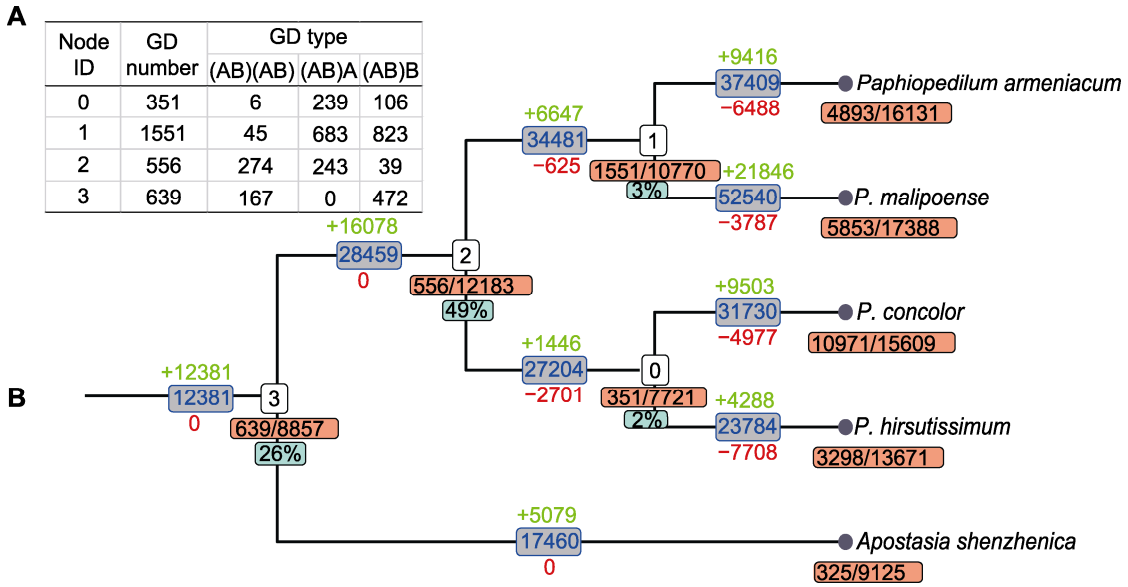

图3 基于系统发生基因组学的全基因组复制检测结果

(A) 各结点的复制基因家族情况统计，其中Node ID对应**(B)**中相应结点; GD number为各结点复制基因家族数量; GD type为复制基因家族每种类型的数量; **(B)** 结点下方黄色方框内数字为复制基因家族的数量/基因家族的数量，绿色方框内数字为(AB)(AB)类型复制基因家族占复制基因家族的比例; 分支上方绿色数字、下方红色数字分别表示基因家族的扩张和收缩。

Figure 3 The detection of whole-genome duplication based on phylogenomics

(A) The statistics of duplicated gene families, Node ID corresponds to the node number in **(B)**; GD number is the number of duplicated gene families at each node; GD type is the number of each type of duplicated gene families; **(B)** The numbers in yellow box below nodes is the number of duplicated gene families/gene families, the corresponding green box is the percentage of (AB)(AB) types; numbers above (green) and below (red) branches indicate the expansion and contraction of gene families, respectively.

与采用K_s检测到的WGD3相一致。由于tree2gd软件主要基于系统发生基因组学方法进行检测，因此受样本限制(仅4种兜兰和深圳拟兰)无法检测到兜兰属以外的全基因组复制事件。

2.4 全基因组复制事件的相对定年

根据高斯混合模型拟合到的K_s峰，采用深圳拟兰$r=6.76\times10^{-9}$ (同义替换/位点/年)，计算4种兜兰发生3次全基因组复制事件的时间，结果分别为110.17–119.77、60.95–74.19和38.19–45.85 Mya (表4)。

2.5 转录组功能注释和复制基因的功能富集分析

对4种兜兰的CDS序列进行功能注释，结果显示，共有102 214个CDS注释到GO数据库(包括同色兜兰27 528个，带叶兜兰15 721个，麻栗坡兜兰33 522个，杏黄兜兰25 443个)，占CDS总数的44.82%。对4种兜兰的功能注释结果进行二级分类统计(图4)，结果共注释到26个生物学过程(biological process, BP)、4种细胞组分(cellular component, CC)及17类分子功能(molecular function, MF)。在生物学过程中，参与细胞进程(cellular process)和代谢过程(metabolic process)的基因数量最多；在细胞组分中，表达细胞解剖实体(cellular anatomical entity)和含蛋白复合体组分(protein-containing complex)的基因数量最多；在分子功能中，参与催化活性(catalytic activity)和结合功能(binding)的基因数量最多。

分别提取4种兜兰各全基因组复制事件保留的复制基因进行GO功能富集分析，发现有1 694个GO条目得到显著富集($P<0.05$)。基于GO条目间的关系，在二级节点对其进行分类统计(图4)，共富集到20个生物学过程、2种细胞组分和12类分子功能；其中，70%生物学过程、100%细胞组分和50%分子功能在3次全基因组复制事件中均得到富集。对3次全基因组复制事件进行比较分析，发现WGD1无特异富集的生物学过程；WGD2特异富集了节律过程(rhythmic process)和移动(locomotion)的生物学过程；WGD3特异富集了色素沉积(pigmentation)、种内个体间相互作用(intraspecies interaction between organisms)和反应(behavior)的生物学过程。

为分析各全基因组复制事件的功能保留模式，我们提取4种兜兰中3种及以上物种共有的功能进行比较分析(图5)，发现在WGD1共同富集了36个GO条目，包括脂类代谢、单糖代谢、苯丙烷类的合成与代谢、叶片发育和衰老调控、活性氧代谢过程调控等生物学过程，以及氧化还原酶活性等分子功能。相较其它2次全基因组复制事件，WGD1特异富集了次生代谢过程、单糖代谢过程、叶片发育和衰老调控以及受伤的应激反应等生物学过程。在WGD2共同富集了3个GO条目，仅包括脱落酸激活的信号通路1个生物学过程。在WGD3共同富集了40个GO条目，包括酶联受体蛋白信号通路、根表皮细胞分化，以及特异富集了行为、植物表皮和保卫细胞的形态发生、毛状体分

表4 基于K_s峰值对全基因组复制(WGD)事件进行定年

Table 4 Dating the whole-genome duplication (WGD) event using K_s distribution peaks

Species	Name of WGD	Mean (K_s)	Age of WGD calculated by K_s mean value (Mya)	Age of WGD with 95% confidence interval (Mya)
Paphiopedilum concolor	WGD3	0.5161	38.19	37.35–39.03
	WGD2	0.8236	60.95	60.06–61.83
	WGD1	1.5043	111.32	108.91–113.72
P. hirsutissimum	WGD3	0.5407	40.01	38.98–41.04
	WGD2	0.8544	63.22	62.19–64.26
	WGD1	1.5894	117.61	114.56–120.67
P. malipoense	WGD3	0.6196	45.85	45.24–46.46
	WGD2	1.0026	74.19	73.39–74.99
	WGD1	1.6186	119.77	118.64–120.91
P. armeniacum	WGD3	0.5664	41.92	41.28–42.56
	WGD2	0.8825	65.31	64.43–66.19
	WGD1	1.4888	110.17	108.42–111.91

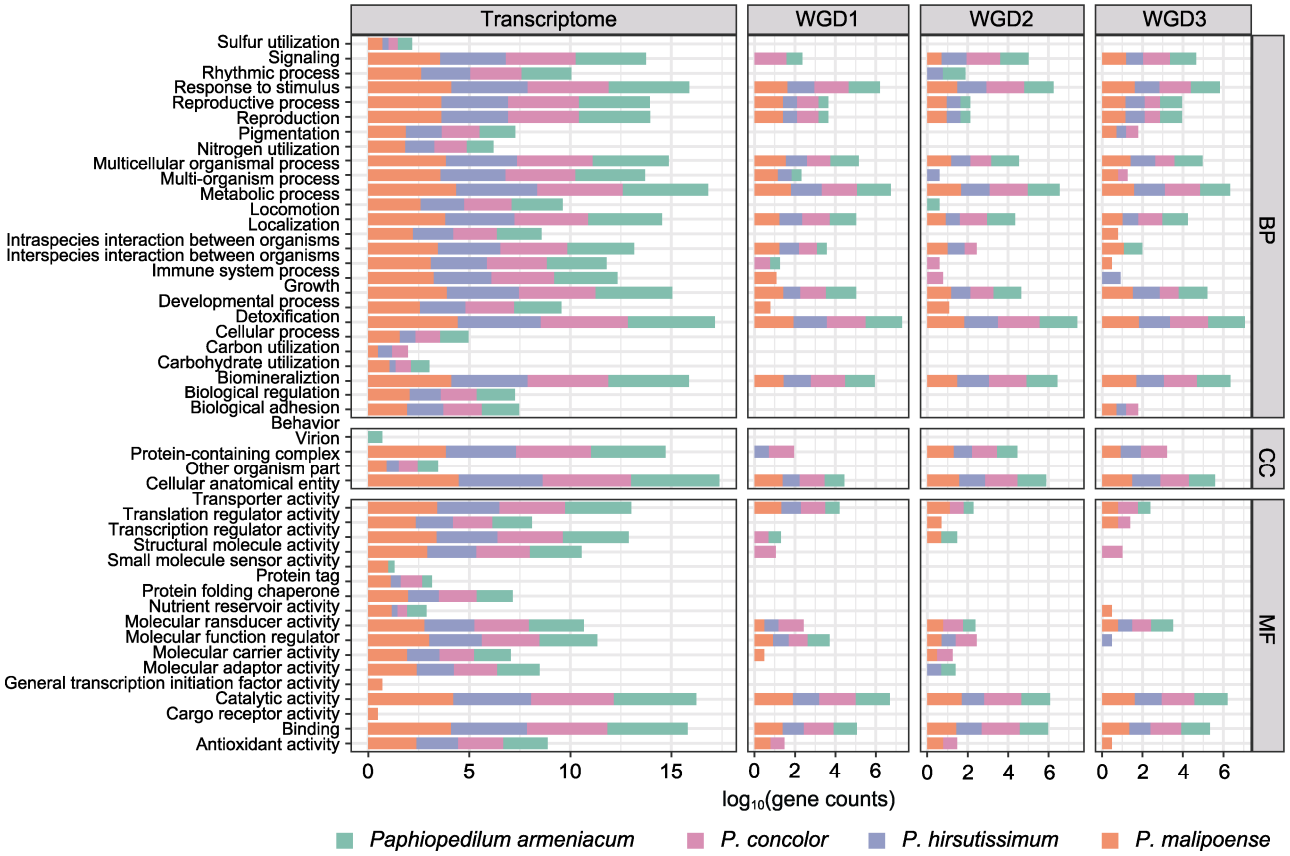

图4 转录组GO功能注释和复制基因功能富集的二级分类统计

Transcriptome: 转录组功能注释结果; WGD1: 全基因组复制事件WGD1中复制基因的功能富集结果(P<0.05); WGD2: 全基因组复制事件WGD2中复制基因的功能富集结果(P<0.05); WGD3: 全基因组复制事件WGD3中复制基因的功能富集结果(P<0.05); BP: 生物学过程; CC: 细胞组分; MF: 分子功能

Figure 4 The level 2 GO categories of transcriptome functional annotation and duplicated gene functional enrichment

Transcriptome: Results of transcriptome functional annotation; WGD1: Functional enrichment of duplicated gene from whole-genome duplication (WGD) event WGD1 (P<0.05); WGD2: Functional enrichment of duplicated gene from WGD2 (P<0.05); WGD3: Functional enrichment of duplicated gene from WGD3 (P<0.05); BP: Biological process; CC: Cellular component; MF: Molecular function

化、细胞壁β-葡聚糖和纤维素代谢过程、甘油酯和甘油磷脂代谢过程、细胞防御反应、色素沉着和细胞发育的正向调控等生物学过程。

2.6 讨论

2.6.1 兰科植物的全基因组复制历史

目前, 在兰科植物中仅检测到2次全基因组复制事件, 一次为大多数单子叶植物共享(110–135 Mya), 另一次为现存兰科植物共享(72–78 Mya)(Cai et al., 2015; Ming et al., 2015; Zhang et al., 2016, 2017; Yuan et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019; Hasing et al., 2020)。兜兰属是兰科多样性的重要代表类群, 本研究基于4种兜兰的转录组数据, 检测到3次全基因组复制事件, 分别发生在110.17–119.77 Mya (WGD1)、60.95–74.19 Mya (WGD2)和38.19–45.85 Mya (WGD3)。其中, WGD1和WGD2发生时间与前期研究得出的2次全基因组复制事件相近, 且物种间K_s分析表明, 二者均发生在兜兰属与深圳拟兰分化事件之前(图2), 因此推测WGD1为大多数单子叶植物共享、WGD2为现存兰科植物共享的全基因组复制事件。而本研究中检测到的全基因组复制事件WGD3 (38.19–45.85 Mya), 在蓝莓(blueberry)、茶树(*Camellia sinensis* var. *sinensis*)和胡萝卜(*Daucus carota*)中同一时期也检

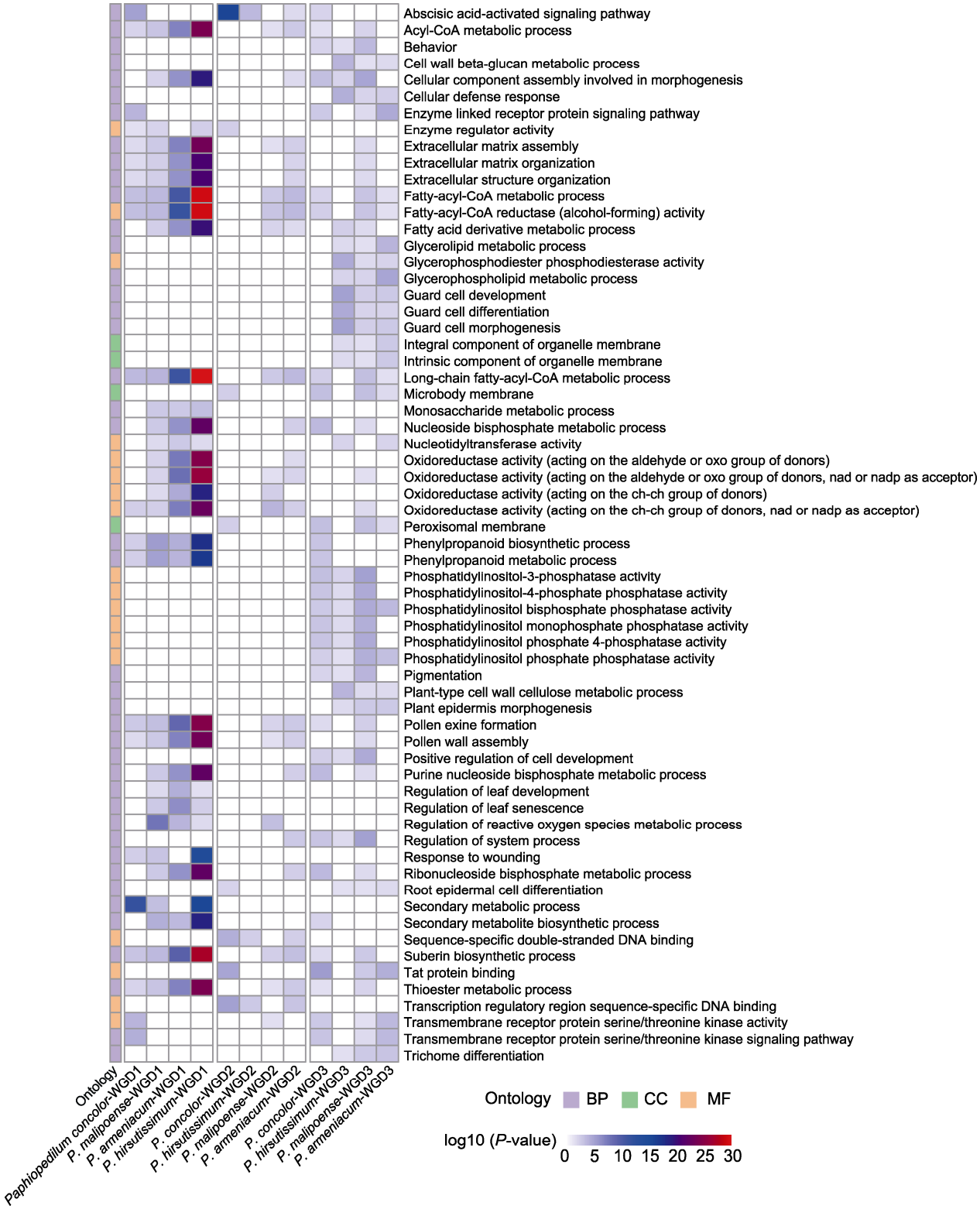

图5 基于GO功能注释结果的3种及以上兜兰物种共同富集的GO条目
BP、CC和MF同图4。

Figure 5 The shared GO terms of 3 or more *Paphiopedilum* species based on the results of GO functional enrichment BP, CC, and MF see Figure 4.

测到了全基因组复制事件(Iorizzo et al., 2016; Wei et al., 2018; Wang et al., 2020), 豆科中更是在该段时间检测到大量全基因组复制事件(17次, 23–55 Mya) (Zhao et al., 2021), 但在兰科植物中尚未见报道。

综合类群分化的时间信息、物种间K_s检测结果以及tree2gd检测结果, 进一步分析WGD3在兰科中的系统发生位置。兰科5个亚科的亲缘关系为(拟兰亚科(香荚兰亚科(杓兰亚科(兰亚科, 树兰亚科)))), 其中杓兰亚科与姐妹类群的分化时间约为64.97 Mya (48.54–84.93 Mya) (Kim et al., 2020), 冠群时间为33 Mya (19–50 Mya) (Gustafsson et al., 2010), 而WGD3的发生时间为38.19–45.85 Mya, 初步推测WGD3为杓兰亚科特异发生的全基因组复制事件。杓兰亚科包含5个属, 其亲缘关系为(杓兰属(南美杓兰属(兜兰属(美洲兜兰属, 镊萼兜兰属)))), 兜兰属与姐妹类群的分化时间为29.9 Mya (14.6–39.1 Mya) (http://www.timetree.org/), 冠群时间为7.09 Mya (5.88–8.41 Mya) (Tsai et al., 2020), 且物种间K_s分析结果(图2)和tree2gd检测结果(图3)均提示WGD3发生在兜兰物种间分化之前, 推测WGD3可能发生在兜兰属与美洲兜兰属、镊萼兜兰属分化之前。综上, 初步推测WGD3发生在杓兰亚科与姐妹类群分化之后, 兜兰属与美洲兜兰属、镊萼兜兰属分化之前。

2.6.2 全基因组复制事件对兜兰属植物适应性演化的意义

多倍化或全基因组复制, 特别是在稳定环境下, 常被认为是进化的终点(Comai, 2005; Oberlander et al., 2016)。然而, 在植物的演化过程中, 全基因组复制并非随机发生, 而是与全球气候变化、地质变化或者大规模灭绝等密切相关, 发生全基因组复制的个体在胁迫或极端环境条件下具有较二倍体祖先更强的适应性(Van de Peer et al., 2017, 2021; Ren et al., 2018; Wu et al., 2020)。与上述研究结果相似, 本研究检测到的3次全基因组复制事件发生时期出现了全球气候变化或大规模灭绝事件, 推测全基因组复制事件提高了兜兰属植物祖先应对极端环境变化的适应性。例如, WGD1 (110.17–119.77 Mya)发生在白垩纪(Cretaceous)阿普特阶(Aptian)至阿尔布阶(Albian), 随后出现了超级温室期(83.6–93.9 Mya) (Klages et al., 2020); WGD2 (60.95–74.19 Mya)发生在白垩纪与古近纪(Paleogene)交界, 出现了白垩纪-古近纪灭绝事件(K-Pg灭绝事件) (Vellekoop et al., 2016); WGD3 (38.19–45.85 Mya)发生在古近纪始新世(Eocene), 发生了古新世-始新世极热事件(56 Mya)和始新世-渐新世(Oligocene)全球变冷(Zachos et al., 2001; McInerney and Wing, 2011)。

全基因组复制事件保留了部分复制基因, 对保留的复制基因进行功能分析可为阐明全基因组复制事件对植物适应性演化的促进作用提供遗传证据。本研究分别对4种兜兰3次全基因组复制后的保留复制基因进行了GO功能富集分析, 发现3次全基因组复制事件富集到的功能存在差异(图4, 图5)。WGD1富集了脂类代谢、软木脂的生物合成、苯丙烷类的合成与代谢, 以及氧化还原酶活性和活性氧代谢过程的调控等功能(图5), 这可能与兜兰属植物应对超级温室期的干旱环境以及抵御干旱引起的活性氧失衡有关(Upchurch, 2008; Das and Roychoudhury, 2014; Brunner et al., 2015; Zhang et al., 2021b)。在K-Pg灭绝时期, 大气中充满了灰尘、硫酸盐气溶胶及碳黑颗粒, 黑暗和低温成为主要的胁迫因子(Vellekoop et al., 2016)。推测WGD2富集的脱落酸激活的信号通路以及昼夜节律等功能提高了兜兰属植物祖先对当时剧变环境的适应性(图4, 图5) (杨有新等, 2014; Vishwakarma et al., 2017)。WGD3之后, 兜兰属植物祖先经历了全球温度骤降, 推测富集的磷脂代谢、酶联受体蛋白信号通路、色素沉着, 以及保卫细胞分化与发育、根表皮细胞分化与毛状体分化等功能, 可能与应对低温引起的植物萎蔫、叶绿素含量减少以及细胞膜发生相变有关(王芳等, 2019)。综上, 推测保留的复制基因在功能上与当时特定的胁迫因子相关。

上述分析结果与前人有关被子植物主要分支的研究结论一致(Wu et al., 2020)。Wu等(2020)对包括被子植物主要分支在内的25个物种(双子叶植物10种, 单子叶植物12种, 基部被子植物、石松类植物和苔藓各1种)在3个历史时期(约120 Mya、约66 Mya及<20 Mya)发生全基因组复制后的保留复制基因进行了功能富集分析, 发现不同时期多倍化后的保留复制基因在功能上与当时的环境压力一致。(1) 在约120 Mya发生全基因组复制事件后的复制基因主要在响应缺水和盐胁迫的功能上显著富集, 当时地球正处于干旱环境; (2) 在K-Pg灭绝时期(约66 Mya)发生了全

球变冷、黑暗、酸雨和野火等, 该时期富集到了与冷、热、渗透、盐和水等胁迫相关的功能, 以及脱落酸激活的信号通路等与胁迫响应相关的其它生物学过程; (3) 在约20 Mya发生全基因组复制事件后的复制基因主要在响应盐胁迫、缺水和机械伤害的功能上显著富集, 与当时CO_2浓度低和相对低温有关。

参考文献

王芳, 王淇, 赵曦阳 (2019). 低温胁迫下植物的表型及生理响应机制研究进展. 分子植物育种 **17**, 5144–5153.

汪浩, 张锐, 张娇, 沈慧, 戴锡玲, 严岳鸿 (2019). 转录组测序揭示翼盖蕨(*Didymochlaena trancatula*)的全基因组复制历史. 生物多样性 **27**, 1221–1227.

王婷, 夏增强, 舒江平, 张娇, 王美娜, 陈建兵, 王慷林, 向建英, 严岳鸿 (2021). 全基因组复制事件的绝对定年揭示莲座蕨属植物的迟滞演化. 生物多样性 **29**, 722–734.

王笃竹, 陈跃, 秦德辉, 陈丽萍, 孙崇波 (2019). 兰科植物染色体研究现状及前景. 分子植物育种 **17**, 3717–3725.

王振怡, 王希胤 (2020). 染色体数目减少及B染色体产生的进化基因组学模型. 中国科学: 生命科学 **50**, 524–537.

杨有新, 王峰, 蔡加星, 喻景权, 周艳虹 (2014). 光质和光敏色素在植物逆境响应中的作用研究进展. 园艺学报 **41**, 1861–1872.

杨志娟 (2006). 兜兰属(*Paphiopedilum*)植物细胞学及其亲缘关系的研究. 硕士论文. 杨凌: 西北农林科技大学. pp. 12–36.

Adams KL, Wendel JF (2005). Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* **8**, 135–141.

Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Brière C, Owens GL, Carrère S, Mayjonade B, Legrand L, Gill N, Kane NC, Bowers JE, Hubner S, Bellec A, Bérard A, Bergès H, Blanchet N, Boniface MC, Brunel D, Catrice O, Chaidir N, Claudel C, Donnadieu C, Faraut T, Fievet G, Helmstetter N, King M, Knapp SJ, Lai Z, Le Paslier MC, Lippi Y, Lorenzon L, Mandel JR, Marage G, Marchand G, Marquand E, Bret-Mestries E, Morien E, Nambeesan S, Nguyen T, Pegot-Espagnet P, Pouilly N, Raftis F, Sallet E, Schiex T, Thomas J, Vandecasteele C, Varès D, Vear F, Vautrin S, Crespi M, Mangin B, Burke JM, Salse J, Muños S, Vincourt P, Rieseberg LH, Langlade NB (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152.

Blanc G, Wolfe KH (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678.

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.

Brunner I, Herzog C, Dawes MA, Arend M, Sperisen C (2015). How tree roots respond to drought. *Front Plant Sci* **6**, 547.

Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, Chen LJ, He Y, Xu Q, Bian C, Zheng ZJ, Sun FM, Liu WQ, Hsiao YY, Pan ZJ, Hsu CC, Yang YP, Hsu YC, Chuang YC, Dievart A, Dufayard JF, Xu X, Wang JY, Wang J, Xiao XJ, Zhao XM, Du R, Zhang GQ, Wang MN, Su YY, Xie GC, Liu GH, Li LQ, Huang LQ, Luo YB, Chen HH, Van de Peer Y, Liu ZJ (2015). The genome sequence of the orchid *Phalaenopsis equestris*. *Nat Genet* **47**, 65–72.

Castresana J (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540–552.

Chase MW, Cameron KM, Freudenstein JV, Pridgeon AM, Salazar G, van den Berg C, Schuiteman A (2015). An updated classification of Orchidaceae. *Bot J Linn Soc* **177**, 151–174.

Chen SC, Liu ZJ, Zhu GH, Lang KY, Tsi ZH, Luo YB, Jin XH, Cribb PJ, Wood JJ, Gale SW, Ormerod P, Vermeulen JJ, Wood HP, Clayton D, Bell A (2009). Orchidaceae. In: Wu ZY, Raven PH, Hong DY, eds. Flora of China, Vol. 25. Beijing: Science Press. pp. 381–382.

Comai L (2005). The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**, 836–846.

Cox AV, Abdelnour GJ, Bennett MD, Leitch IJ (1998). Genome size and karyotype evolution in the slipper orchids (Cypripedioideae: Orchidaceae). *Am J Bot* **85**, 681–687.

Cox AV, Pridgeon AM, Albert VA, Chase MW (1997). Phylogenetics of the slipper orchids (Cypripedioideae, Orchidaceae): nuclear rDNA ITS sequences. *Plant Syst Evol* **208**, 197–223.

Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16**, 738–749.

da Conceição LP, de Oliveira ALPC, Barbosa LV (2006). Characterization of the species *Epidendrum cinnabarium* salzm. (Epidendroideae: Orchidaceae) occurring in dunas

do abaeté-salvador, ba-brasil. *Cytologia* **71**, 125–129.

Darriba D, Taboada GL, Doallo R, Posada D (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165.

Das K, Roychoudhury A (2014). Reactive oxygen species (ROS) and response of antioxidants as ROS-scavengers during environmental stress in plants. *Front Environ Sci* **2**, 53.

De Bodt S, Maere S, Van de Peer Y (2005). Genome duplication and the origin of angiosperms. *Trends Ecol Evol* **20**, 591–597.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797.

Emms DM, Kelly S (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238.

Enright AJ, Van Dongen S, Ouzounis CA (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575–1584.

Fang L, Xu X, Li J, Zheng F, Li MZ, Yan JW, Li Y, Zhang XH, Li L, Ma GH, Zhang AY, Lv FB, Wu KL, Zeng SJ (2020). Transcriptome analysis provides insights into the non-methylated lignin synthesis in *Paphiopedilum armeniacum* seed. *BMC Genomics* **21**, 524.

Goldman N, Yang Z (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725–736.

Govaerts R, Bernet P, Kratochvil K, Gerlach G, Carr G, Alrich P, Pridgeon AM, Pfahl J, Campacci MA, Baptista DH, Tigges H, Shaw J, Cribb P, George A, Kreuz K, Wood J (2021). World checklist of Orchidaceae. https://wcsp.science.kew.org/. 2021-05-08.

Gustafsson ALS, Verola CF, Antonelli A (2010). Reassessing the temporal evolution of orchids with new fossils and a Bayesian relaxed clock, with implications for the diversification of the rare South American genus *Hoffmannseggella* (Orchidaceae: Epidendroideae). *BMC Evol Biol* **10**, 177.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494–1512.

Hasing T, Tang HB, Brym M, Khazi F, Huang TF, Chambers AH (2020). A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. *Nat Food* **1**, 811–819.

Huang CH, Qi XP, Chen DY, Qi J, Ma H (2020). Recurrent genome duplication events likely contributed to both the ancient and recent rise of ferns. *J Integr Plant Biol* **62**, 433–455.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* **34**, 2115–2122.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314.

Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang JY, Bowman M, Iovene M, Sanseverino W, Cavagnaro P, Yildiz M, Macko-Podgórni A, Moranska E, Grzebelus E, Grzebelus D, Ashrafi H, Zheng ZJ, Cheng SF, Spooner D, Van Deynze A, Simon P (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat Genet* **48**, 657–666.

Kim YK, Jo S, Cheon SH, Joo MJ, Hong JR, Kwak M, Kim KJ (2020). Plastome evolution and phylogeny of Orchidaceae, with 24 new sequences. *Front Plant Sci* **11**, 22.

Klages JP, Salzmann U, Bickert T, Hillenbrand CD, Gohl K, Kuhn G, Bohaty SM, Titschack J, Müller J, Frederichs T, Bauersachs T, Ehrmann W, van de Flierdt T, Pereira PS, Larter RD, Lohmann G, Niezgodzki I, Uenzelmann-Neben G, Zundel M, Spiegel C, Mark C, Chew D, Francis JE, Nehrke G, Schwarz F, Smith JA, Freudenthal T, Esper O, Pälike H, Ronge TA, Dziadek R (2020). Temperate rainforests near the South Pole during peak Cretaceous warmth. *Nature* **580**, 81–86.

Leinonen R, Sugawara H, Shumway M (2011). The sequence read archive. *Nucleic Acids Res* **39**, D19–D21.

Leitch IJ, Kahandawala I, Suda J, Hanson L, Ingrouille MJ, Chase MW, Fay MF (2009). Genome size diversity in orchids: consequences and evolution. *Ann Bot* **104**, 469–481.

Li D, Yin H, Zhao C, Zhu G, Lǚ F (2014). Transcriptome analysis of tessellated and green leaves in *Paphiopedilum* orchids using Illumina paired-end sequencing and discovery simple sequence repeat markers. *J Plant Biochem Physiol* **2**, 1000136.

Li MH, Zhang GQ, Lan SR, Liu ZJ, China Phylogeny Consortium (2016). A molecular phylogeny of Chinese orchids. *J Syst Evol* **54**, 349–362.

Li WZ, Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659.

Mandáková T, Lysak MA (2018). Post-polyploid diploidization and diversification through dysploid changes. *Curr Opin Plant Biol* **42**, 55–65.

McInerney FA, Wing SL (2011). The paleocene-eocene thermal maximum: a perturbation of carbon cycle, climate, and biosphere with implications for the future. *Annu Rev Earth Planet Sci* **39**, 489–516.

Ming R, VanBuren R, Wai CM, Tang HB, Schatz MC, Bowers JE, Lyons E, Wang ML, Chen J, Biggers E, Zhang JS, Huang LX, Zhang LM, Miao WJ, Zhang J, Ye ZY, Miao CY, Lin ZC, Wang H, Zhou HY, Yim WC, Priest HD, Zheng CF, Woodhouse M, Edger PP, Guyot R, Guo HB, Guo H, Zheng GY, Singh R, Sharma A, Min XJ, Zheng Y, Lee H, Gurtowski J, Sedlazeck FJ, Harkess A, McKain MR, Liao ZY, Fang JP, Liu J, Zhang XD, Zhang Q, Hu WC, Qin Y, Wang K, Chen LY, Shirley N, Lin YR, Liu LY, Hernandez AG, Wright CL, Bulone V, Tuskan GA, Heath K, Zee F, Moore PH, Sunkar R, Leebens-Mack JH, Mockler T, Bennetzen JL, Freeling M, Sankoff D, Paterson AH, Zhu XG, Yang XH, Smith JAC, Cushman JC, Paull RE, Yu QY (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* **47**, 1435–1442.

Murat F, Armero A, Pont C, Klopp C, Salse J (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat Genet* **49**, 490–496.

Oberlander KC, Dreyer LL, Goldblatt P, Suda J, Linder HP (2016). Species-rich and polyploid-poor: insights into the evolutionary role of whole-genome duplication from the Cape flora biodiversity hotspot. *Am J Bot* **103**, 1336–1347.

One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685.

Paterson AH, Bowers JE, Chapman BA (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* **101**, 9903–9908.

Ren R, Wang HF, Guo CC, Zhang N, Zeng LP, Chen YM, Ma H, Qi J (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol Plant* **11**, 414–428.

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11–24.

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004). Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**, 868–876.

Scrucca L, Fop M, Murphy TB, Raftery AE (2016). Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J* **8**, 289–317.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212.

Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, Cooper ED, Uauy C, Havlickova L, Worswick G, Studholme DJ, Zohren J, Salmon DL, Clavijo BJ, Li Y, He ZS, Fellgett A, McKinney LV, Nielsen LR, Douglas GC, Kjær ED, Downie JA, Boshier D, Lee S, Clark J, Grant M, Bancroft I, Caccamo M, Buggs RJA (2017). Genome sequence and genetic diversity of European ash trees. *Nature* **541**, 212–216.

Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Talavera G, Castresana J (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564–577.

Tsai CC, Liao PC, Ko YZ, Chen CH, Chiang YC (2020). Phylogeny and historical biogeography of *Paphiopedilum* pfitzer (Orchidaceae) based on nuclear and plastid DNA. *Front Plant Sci* **11**, 126.

Unruh SA, McKain MR, Lee YI, Yukawa T, McCormick MK, Shefferson RP, Smithson A, Leebens-Mack JH, Pires JC (2018). Phylotranscriptomic analysis and genome evolution of the Cypripedioideae (Orchidaceae). *Am J Bot* **105**, 631–640.

Upchurch RG (2008). Fatty acid unsaturation, mobilization, and regulation in the response of plants to stress. *Biotechnol Lett* **30**, 967–977.

Van de Peer Y, Ashman TL, Soltis PS, Soltis DE (2021). Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* **33**, 11–26.

Van de Peer Y, Mizrachi E, Marchal K (2017). The

evolutionary significance of polyploidy. *Nat Rev Genet* **18**, 411–424.

Vellekoop J, Esmeray-Senlet S, Miller KG, Browning JV, Sluijs A, van de Schootbrugge B, Damsté JSS, Brinkhuis H (2016). Evidence for Cretaceous-Paleogene boundary bolide 'impact winter' conditions from New Jersey, USA. *Geology* **44**, 619–622.

Vishwakarma K, Upadhyay N, Kumar N, Yadav G, Singh J, Mishra RK, Kumar V, Verma R, Upadhyay RG, Pandey M, Sharma S (2017). Abscisic acid signaling and abiotic stress tolerance in plants: a review on current knowledge and future prospects. *Front Plant Sci* **8**, 161.

Wang YS, Nie F, Shahid MQ, Baloch FS (2020). Molecular footprints of selection effects and whole genome duplication (WGD) events in three blueberry species: detected by transcriptome dataset. *BMC Plant Biol* **20**, 250.

Wei CL, Yang H, Wang SB, Zhao J, Liu C, Gao LP, Xia EH, Lu Y, Tai YL, She GB, Sun J, Cao HS, Tong W, Gao Q, Li YY, Deng WW, Jiang XL, Wang WZ, Chen Q, Zhang SH, Li HJ, Wu JL, Wang P, Li PH, Shi CY, Zheng FY, Jian JB, Huang B, Shan D, Shi MM, Fang CB, Yue Y, Li FD, Li DX, Wei S, Han B, Jiang CJ, Yin Y, Xia T, Zhang ZZ, Bennetzen JL, Zhao SC, Wan XC (2018). Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *Proc Natl Acad Sci USA* **115**, E4151–E4158.

Wendel JF (2000). Genome evolution in polyploids. *Plant Mol Biol* **42**, 225–249.

Wu SD, Han BC, Jiao YN (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol Plant* **13**, 59–71.

Yang ZH (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586–1591.

Yu GC, Wang LG, Han YY, He QY (2012). ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*: *J Integr Biol* **16**, 284–287.

Yuan Y, Jin XH, Liu J, Zhao X, Zhou JH, Wang X, Wang DY, Lai CS, Xu W, Huang JW, Zha LP, Liu DH, Ma X, Wang L, Zhou MY, Jiang Z, Meng HB, Peng HS, Liang YT, Li RQ, Jiang C, Zhao YY, Nan TG, Jin Y, Zhan ZL, Yang J, Jiang WK, Huang LQ (2018). The *Gastrodia elata* genome provides insights into plant adaptation to heterotrophy. *Nat Commun* **9**, 1615.

Zachos J, Pagani H, Sloan L, Thomas E, Billups K (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693.

Zhang CF, Huang CH, Liu M, Hu Y, Panero JL, Luebert F, Gao TG, Ma H (2021a). Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. *J Integr Plant Biol* **63**, 1273–1293.

Zhang CF, Zhang TK, Luebert F, Xiang YZ, Huang CH, Hu Y, Rees M, Frohlich MW, Qi J, Weigend M, Ma H (2020). Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol Biol Evol* **37**, 3188–3210.

Zhang CL, Chen JH, Huang WX, Song XQ, Niu J (2021b). Transcriptomics and metabolomics reveal purine and phenylpropanoid metabolism response to drought stress in *Dendrobium sinense*, an endemic orchid species in Hainan island. *Front Genet* **12**, 692702.

Zhang GQ, Liu KW, Li Z, Lohaus R, Hsiao YY, Niu SC, Wang JY, Lin YC, Xu Q, Chen LJ, Yoshida K, Fujiwara S, Wang ZW, Zhang YQ, Mitsuda N, Wang MN, Liu GH, Pecoraro L, Huang HX, Xiao XJ, Lin M, Wu XY, Wu WL, Chen YY, Chang SB, Sakamoto S, Ohme-Takagi M, Yagi M, Zeng SJ, Shen CY, Yeh CM, Luo YB, Tsai WC, Van de Peer Y, Liu ZJ (2017). The *Apostasia* genome and the evolution of orchids. *Nature* **549**, 379–383.

Zhang GQ, Xu Q, Bian C, Tsai WC, Yeh CM, Liu KW, Yoshida K, Zhang LS, Chang SB, Chen F, Shi Y, Su YY, Zhang YQ, Chen LJ, Yin YY, Lin M, Huang HX, Deng H, Wang ZW, Zhu SL, Zhao X, Deng C, Niu SC, Huang J, Wang MN, Liu GH, Yang HJ, Xiao XJ, Hsiao YY, Wu WL, Chen YY, Mitsuda N, Ohme-Takagi M, Luo YB, Van de Peer Y, Liu ZJ (2016). The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci Rep* **6**, 19029.

Zhao YY, Zhang R, Jiang KW, Qi J, Hu Y, Guo J, Zhu RB, Zhang TK, Egan AN, Yi TS, Huang CH, Ma H (2021). Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Mol Plant* **14**, 748–773.

Zheng Y, Jiao C, Sun HH, Rosli HG, Pombo MA, Zhang PF, Banf M, Dai XB, Martin GB, Giovannoni JJ, Zhao PX, Rhee SY, Fei ZJ (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol Plant* **9**, 1667–1670.

Zwaenepoel A, Van de Peer Y (2019). Wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153–2155.

Revealing the New Whole-genome Duplication Event of Four *Paphiopedilum* Species Based on Transcriptome Data

Meng Wang[1], Ting Wang[1, 2], Zengqiang Xia [1, 3], Tingzhang Li [1]
Xiaohua Jin[4], Yuehong Yan[1], Jianbing Chen[1*]

[1]Key Laboratory of National Forestry and Grassland Administration for Orchid Conservation and Utilization, The Orchid Conservation & Research Center of Shenzhen, Shenzhen 518114, China; [2]College of Biodiversity Conservation, South-west Forestry University, Kunming 650224, China; [3]Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai 200032, China; [4]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

Abstract Polyploidization or whole-genome duplication (WGD) is a major driving force in species diversification. Repeated WGD has been found in species-rich groups or families such as ferns, Asteraceae or Fabaceae. However, there is a paradox between the abundant species diversity of Orchidaceae and the rarely discovered WGD events. We hypothesized that it could be due to the early research strategy of a small sample size of species belonging to not closely related groups. *Paphiopedilum* has a rich variation on chromosome number and morphology. Here, we select it as the representative group of orchids, and use the synonymous substitution rate (K_s), phylogenomic and relative dating methods to detect whether WGD events have been occurred in *P. armeniacum*, *P. concolor*, *P. hirsutissimum* or *P. malipoense* based on the shared transcriptome data. The result shows that three WGDs are detected in all four species of *Paphiopedilum*, which occurred in 110.17–119.77 Mya (WGD1), 60.95–74.19 Mya (WGD2), and 38.19–45.85 Mya (WGD3), respectively. WGD3 is a newly detected whole-genome duplication event in orchids and speculated to occur after the differentiation of Cypripedioideae and its sister groups, but before the differentiation of *Paphiopedilum* and its sister groups. In addition, the retained duplicated genes of three WGDs are functionally related with the environmental stress response at that time. It is speculated that WGD has improved the adaptability of the ancestors of *Paphiopedilum* to extreme environmental changes.

Key words polyploidization, Ochidaceae, Cypripedioideae, adaptive evolution, synonymous substitution rate

Wang M, Wang T, Xia ZQ, Li TZ, Jin XH, Yan YH, Chen JB (2021). Revealing the new whole-genome duplication event of four *Paphiopedilum* species based on transcriptome data. *Chin Bull Bot* **56**, 699–714.

* Author for correspondence. E-mail: cjb@cnocc.cn

(责任编辑：白羽红)

附表 1 4 种兜兰转录因子家族
Appendix table 1 Transcription factor family in the four *Paphiopedilum* species

https://www.chinbullbotany.com/fileup/1674-3466/PDF/t21-100.pdf

附表 1 4 种兜兰转录因子家族
Appendix table 1 Transcription factor family in the four *Paphiopedilum* species

TF family	Member number			
	P. concolor	*P. hirsutissimum*	*P. malipoense*	*P. armeniacum*
AP2/ERF-AP2	24	13	23	24
AP2/ERF-ERF	106	61	131	105
AP2/ERF-RAV	3	1	2	1
Alfin-like	20	10	16	15
B3	47	24	78	74
B3-ARF	35	27	41	38
BBR-BPC	24	11	15	15
BES1	17	14	18	16
BSD	4	3	1	4
C2C2-CO-like	8	6	7	5
C2C2-Dof	25	11	40	30
C2C2-GATA	38	28	43	41
C2C2-LSD	16	20	15	13
C2C2-YABBY	4	3	5	5
C2H2	96	77	181	138
C3H	149	90	153	115
CAMTA	8	7	12	14
CPP	9	10	17	8
CSD	7	2	7	3
DBB	9	4	10	8
DBP	5	2	7	9
DDT	19	8	15	17
E2F-DP	13	8	22	22
EIL	4	2	3	2
FAR1	62	38	79	72
GARP-ARR-B	14	9	5	9
GARP-G2-like	55	30	88	67
GRAS	69	39	53	47
GRF	5	2	24	16
GeBP	20	11	15	17
HB-BELL	21	18	29	21
HB-HD-ZIP	55	23	53	43
HB-KNOX	5	4	14	10
HB-PHD	1	1	3	2
HB-WOX	2	2	3	12
HB-other	35	15	41	36
HRT	2	1	1	1
HSF	23	16	31	29
LFY	/	/	/	1
LIM	7	5	8	5
LOB	3	3	18	20
MADS-M-type	9	9	27	20
MADS-MIKC	18	10	39	46
MYB	56	32	105	77

TF family	Member number			
	P. concolor	*P. hirsutissimum*	*P. malipoense*	*P. armeniacum*
MYB-related	88	64	155	94
NAC	103	49	115	103
NF-X1	4	5	6	2
NF-YA	30	19	29	18
NF-YB	23	14	30	19
NF-YC	27	17	36	24
OFP	2	2	12	13
PLATZ	8	8	16	10
RWP-RK	6	3	5	7
SAP	/	/	/	1
S1Fa-like	2	1	/	/
SBP	33	23	48	26
SRS	/	1	3	5
STAT	4	2	2	3
TCP	34	27	32	29
TUB	25	15	28	26
Tify	32	9	55	28
Trihelix	38	32	46	38
ULT	3	2	3	6
VOZ	2	1	4	2
WRKY	83	23	106	56
Whirly	7	3	4	4
bHLH	121	72	168	111
bZIP	115	67	146	103
zf-HD	8	12	22	13
zn-clus	/	/	17	/