nature genetics
OPEN

The genome sequence of the orchid *Phalaenopsis equestris*

Jing Cai[1–3,15], Xin Liu[4,15], Kevin Vanneste[5,6,15], Sebastian Proost[5,6,15], Wen-Chieh Tsai[7,15], Ke-Wei Liu[1–3,15], Li-Jun Chen[1], Ying He[5,6], Qing Xu[8], Chao Bian[4], Zhijun Zheng[4], Fengming Sun[4], Weiqing Liu[4], Yu-Yun Hsiao[9], Zhao-Jun Pan[9], Chia-Chi Hsu[9], Ya-Ping Yang[9], Yi-Chin Hsu[9], Yu-Chen Chuang[9], Anne Dievart[10], Jean-Francois Dufayard[10], Xun Xu[4], Jun-Yi Wang[4], Jun Wang[4], Xin-Ju Xiao[1], Xue-Min Zhao[11], Rong Du[11], Guo-Qiang Zhang[1], Meina Wang[1], Yong-Yu Su[12], Gao-Chang Xie[1], Guo-Hui Liu[1], Li-Qiang Li[1], Lai-Qiang Huang[1–3,12], Yi-Bo Luo[8], Hong-Hwa Chen[9,13], Yves Van de Peer[5,6,14] & Zhong-Jian Liu[1,2,12]

Orchidaceae, renowned for its spectacular flowers and other reproductive and ecological adaptations, is one of the most diverse plant families. Here we present the genome sequence of the tropical epiphytic orchid *Phalaenopsis equestris*, a frequently used parent species for orchid breeding. *P. equestris* is the first plant with crassulacean acid metabolism (CAM) for which the genome has been sequenced. Our assembled genome contains 29,431 predicted protein-coding genes. We find that contigs likely to be underassembled, owing to heterozygosity, are enriched for genes that might be involved in self-incompatibility pathways. We find evidence for an orchid-specific paleopolyploidy event that preceded the radiation of most orchid clades, and our results suggest that gene duplication might have contributed to the evolution of CAM photosynthesis in *P. equestris*. Finally, we find expanded and diversified families of MADS-box C/D-class, B-class *AP3* and *AGL6*-class genes, which might contribute to the highly specialized morphology of orchid flowers.

Ever since Darwin published *Fertilization of Orchids* in 1862 (ref. 1), orchids have attracted great interest from evolutionary biologists and botanists. Orchidaceae is one of the largest plant families, with between 22,075 and 26,567 species in 880 genera (see URLs), and is known for its diversity in specialized reproductive and ecological strategies. The specific development of the labellum (the 'lip') and gynostemium (a fused structure of the stamens and pistils) to trick pollinators and facilitate pollination is well documented, and the coevolution of orchid flowers and pollinators is well known[2]. In addition to the highly sophisticated floral structure contributing to the diversification of orchids[3], CAM and epiphytism[4] might also be linked to the adaptive radiation of orchids.

Phalaenopsis species are popular ornamental plants worldwide because of their elegant appearance and extended longevity, and they are of great economic importance for the floral industry. *P. equestris* is an important breeding parent because of its many colorful flowers in a single inflorescence. It has a karyotype of $2N = 2X = 38$ with uniform small-size chromosomes of 1–2.5 μm in length[5]. Its genome size is estimated to be 1.6×10^9 bp per haploid genome, which is relatively small in comparison to the genomes of other species in the same genus[5] or even other genera[6]. Some transcriptome sequence data have previously been generated for this species, and a transcriptome database, OrchidBase[7,8], has been established. Here we present the first whole-genome sequence and analysis of *P. equestris*, providing fundamental knowledge for further research in orchid biology.

RESULTS

Genome sequencing and assembly

We adopted a whole-genome shotgun strategy to sequence and assemble the genome of *P. equestris* (**Supplementary Table 1**) and estimated the genome size to be 1.16 Gb (**Supplementary Fig. 1** and **Supplementary Note**), which is smaller than previous estimates by flow cytometry (~1.6 Gb)[5]. We assembled 236,185 scaffolds (**Supplementary Table 2**), with about 90% of the total assembled genome (~980 Mb) contained in the 6,359 longest scaffolds.

[1]Shenzhen Key Laboratory for Orchid Conservation and Utilization, National Orchid Conservation Center of China and Orchid Conservation and Research Center of Shenzhen, Shenzhen, China. [2]Center for Biotechnology and BioMedicine, Shenzhen Key Laboratory of Gene & Antibody Therapy, State Key Laboratory of Health Science & Technology (prep) and Division of Life & Health Sciences, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. [3]School of Life Science, Tsinghua University, Beijing, China. [4]BGI-Shenzhen, Shenzhen, China. [5]Department of Plant Systems Biology, VIB, Ghent, Belgium. [6]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [7]Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan, Taiwan. [8]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China. [9]Department of Life Sciences, National Cheng Kung University, Tainan, Taiwan. [10]Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR Amélioration Génétique et Adaptation des Plantes Méditerranéennes et Tropicales (AGAP), Montpellier, France. [11]State Forestry Administration, Beijing, China. [12]College of Forestry, South China Agriculture University, Guangzhou, China. [13]Orchid Research Center, National Cheng Kung University, Tainan, Taiwan. [14]Department of Genetics, Genomics Research Institute, Pretoria, South Africa. [15]These authors contributed equally to this work. Correspondence should be addressed to L.-Q.H. (huanglq@mail.tsinghua.edu.cn), Y.-B.L. (luoyb@ibcas.ac.cn), H.-H.C. (hhchen@mail.ncku.edu.tw), Y.V.d.P. (yves.vandepeer@psb.vib-ugent.be) or Z.-J.L. (liuzj@sinicaorchid.org).

Received 2 September; accepted 29 October; published online 24 November 2014; corrected after print 9 January 2015; corrected after print 6 February 2015; doi:10.1038/ng.3149

The total genome assembly amounted to 1.086 Gb (1.0 Gb without unknown (N) bases), which is ~93% (~86% without N bases) of the estimated total genome size. We assessed the coverage of the genome assembly using BACs (**Supplementary Tables 3** and **4**, and **Supplementary Note**). For contigs longer than 1 kb in length (979 in total) generated by sequencing and assembling BAC pools, 97% could be mapped to the assembled genome (**Supplementary Fig. 2**). Comparison of ten randomly sequenced BAC scaffolds indicated a low error rate (**Supplementary Table 4**). Of the 248 conserved core eukaryotic genes that were used to assess genome completeness[9], 234 (94%) were uncovered in our genome assembly (**Supplementary Table 5** and **Supplementary Note**), and a majority of transcription fragments could be mapped to the genome assembly (**Supplementary Tables 6**–**8**). Further detailed information is provided in the Online Methods.

Genome annotation and gene expression analysis

About 62% of the genome assembly was found to be composed of repetitive DNA, a higher proportion than the 29% in rice[10] and 41% in grape[11] but similar to the proportion in sorghum (61%)[12]. Interspersed repeats and transposable elements (TEs) occupied about 59% of the genome, and tandem repeats accounted for 3% (**Supplementary Table 9**). Among the TEs, LTRs (long terminal repeats) were the most abundant and occupied ~46% of the genome, followed in frequency by LINEs (long interspersed nuclear elements), which accounted for ~8% of the genome (**Supplementary Tables 10** and **11**).

We analyzed the distribution of the divergence times for the complete LTRs in the genome assembly and estimated that most LTRs (71%) in the orchid genome arose during a relatively recent insertion between 11.7 and 43 million years ago (**Supplementary Fig. 3**), long after the origin of the last common ancestor of orchids (74–86 million years ago)[13]. Separate divergence analysis of the *Copia* and *Gypsy* TEs suggested that these two types of LTRs experienced a recent burst (**Supplementary Fig. 3**).

Next, protein-coding gene models were constructed using a pipeline combining *de novo* prediction, homology-based prediction and RNA sequence–aided prediction. In total, we predicted 29,431 protein-coding gene models, excluding genes with similarity to known TEs (**Supplementary Table 12**). We also predicted alternatively spliced forms of these protein-coding genes (**Supplementary Table 13**) and found 9,021 splice variants for 6,389 genes (21.7% of the total protein-coding genes). Of the 29,431 predicted genes, 20,398 (69.3%) were supported by transcriptome data from at least 1 of 4 tissues examined (leaf, flower, stem and root), and 15,530 gene models (52.8%) were supported by all 3 prediction methods, namely, homology-based, *de novo* and transcriptome-based prediction, and are therefore considered to represent the high-confidence gene set (**Supplementary Fig. 4**). For gene model validation, we manually examined 500 randomly selected genes from 2,038 of the gene families shared among monocots that only had 1 copy in both *P. equestris* and rice (**Supplementary Table 14**). Most of these 500 genes were correctly predicted (**Supplementary Note**), although 20 contained potential annotation errors, overall reflecting an annotation accuracy of 96%.

Gene similarity clustering for the set of 29,431 predicted genes yielded 3,694 gene families containing at least 2 orchid genes. Furthermore, 4,171 orchid genes could not be grouped with any of the genes from the following species—*Populus trichocarpa*, *Arabidopsis thaliana*, *Vitis vinifera*, *Oryza sativa*, *Brachypodium distachyon*, *Sorghum bicolor*, *Zea mays*, *Physcomitrella patens*, *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus*—and are referred to as orphan genes. A 4-way comparison of orchid, rice, grape and *A. thaliana* (**Fig. 1a**) showed that there were 5,696 gene families shared by all 4 species, and 4,775 gene families were unique to *P. equestris*, more than in *A. thaliana* (2,647) and grape (3,634) but fewer than in rice (10,905). For all species, expanded and contracted gene families (in comparison to ancestors) were compared with those of *P. equestris* to identify gene families that were only expanded or contracted in *P. equestris*. In total, 2,497 gene families were expanded in *P. equestris*, whereas only 3 gene families were contracted (**Supplementary Tables 15** and **16**). For the gene families specifically expanded in *P. equestris*, we conducted Gene Ontology (GO) enrichment analysis and found enrichment for 'transition metal ion binding' and 'zinc ion binding', probably reflecting the importance of genes for the binding of metal ions (including iron and zinc) in orchid.

To obtain functional annotations for the coding genes, we first used InterProScan[14] to identify known protein domains. This approach uncovered 94,693 domains encoded by 17,931 genes (60.93% of all genes). On the basis of their encoding specific protein domains, 12,739 genes were linked to 129,064 GO annotations. Second, we used high-quality functional annotations from rice, excluding GO labels inferred through electronic annotation alone, in combination with a tree-based approach to transfer these labels to orchid[15]. In this way, 8,518 new or more specific GO labels were assigned to 3,090 genes. The combination of both approaches resulted in a total of 13,954 genes (47.41%) with GO terms assigned to them (**Supplementary Table 17**).

To quantify general gene expression levels, we mapped all RNA sequencing (RNA-seq) data to the annotated genes and calculated RPKM values (reads per kilobase per million mapped reads) for every gene in the four different tissues analyzed (**Supplementary Table 18**). Differentially expressed genes were detected using the method described by Chen *et al.*[16]. Using false discovery rate (FDR) $P < 0.05$ as the threshold for significance, we found 2,283, 1,499, 947 and 1,288 genes that were preferentially expressed in flower, leaf, stem and root, respectively. Among the genes preferentially expressed in flowers, there was high enrichment for GO terms related to 'cell wall', 'cell wall modification' and 'pectinesterase activity' ($P < 0.05$; **Supplementary Table 19**), confirming the correlation of modification and organization of cell walls with flower development and wilting[17]. In leaves, GO terms related to 'photosynthesis', 'electron transport chain' and 'photosystem' were significantly enriched ($P < 0.05$), consistent with photosynthesis being the major function of this tissue. The 'heme binding' and 'iron ion binding' functions were enriched in both root and stem, suggesting that both organs have an important role in metal ion metabolism in orchid. GO terms related to 'transition metal ion metabolism' were also uncovered in GO term enrichment analysis for gene families with *P. equestris*–specific expansion and might suggest that the metabolism of transition metals (including iron, zinc and copper) in both stems and roots has an important role in adaptation to *P. equestris*–specific epiphytic growth niches. Mineral ions in epiphytic growth niches are unevenly distributed, usually diluted and only sporadically available in comparison to their distribution in growth niches in soils[18]. Gene duplication in these metabolic pathways might have contributed to different regulation of mineral ion metabolism adapted to epiphytic niches. It would be very interesting to see whether the same gene families have also been expanded in other non-orchid epiphytes (such as bromeliads).

Heterozygosity

To estimate the level of heterozygosity in the genome, we carried out *k*-mer distribution approximation with simulated heterozygous genome sequences and found that the real *k*-mer distribution was fitted

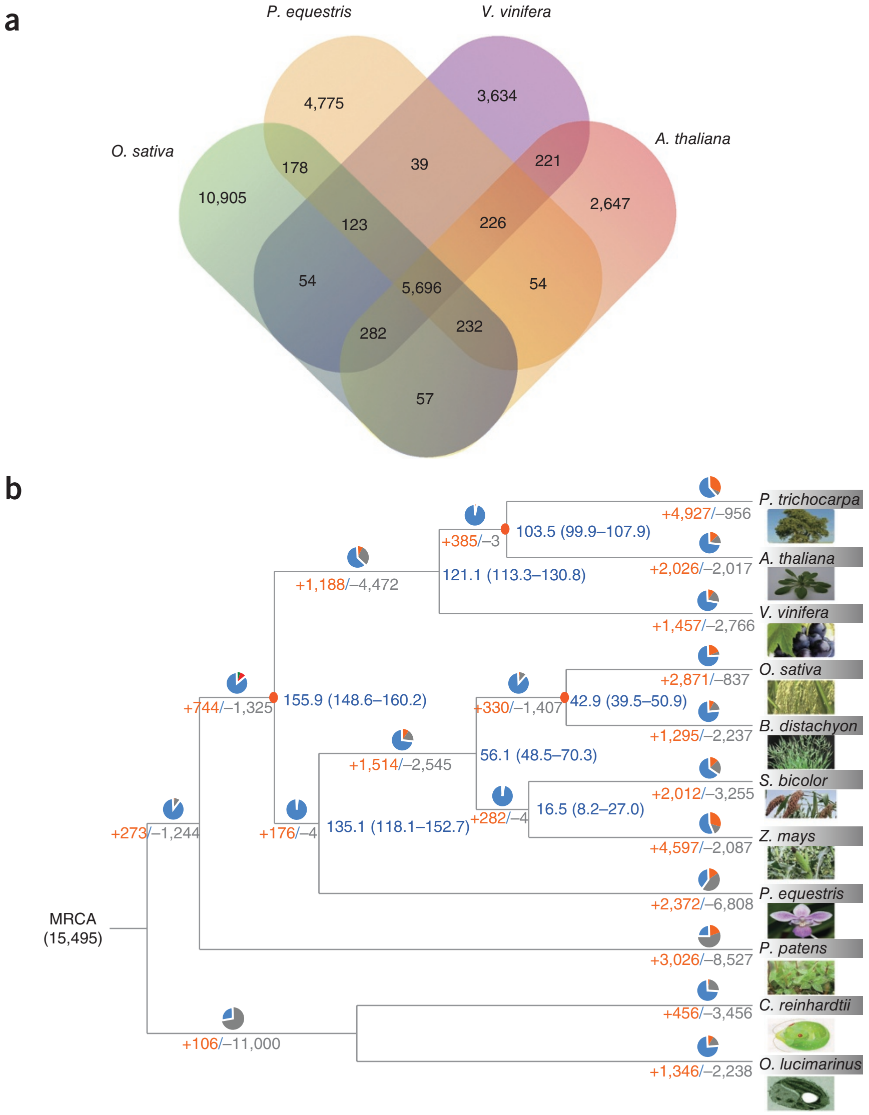

Figure 1 Evolution of *P. equestris*. (**a**) Comparison of the number of gene families in orchid (*P. equestris*), rice (*O. sativa*), grapevine (*V. vinifera*) and *A. thaliana*. (**b**) Phylogenetic tree and gene family expansion and contraction. The phylogenetic tree was constructed from a concatenated alignment of 72 single-copy gene families from 11 green plant species. Gene family expansions are indicated in orange, and gene family contractions are indicated in gray; the corresponding propotions among total changes are shown using the same colors in the pie charts. Inferred divergence dates (in millions of years) are denoted at each node in blue. MRCA, most recent common ancestor. Blue portions of the pie charts represent the conserved gene families.

best by a simulated *k*-mer (*k* represents the chosen length of substrings) distribution with 1.2% (between 1.1% and 1.3%) heterozygosity (**Supplementary Fig. 1**). We also investigated the level of heterozygosity in the assembled part of the genome by mapping all reads back to the assembly, finding that the heterozygosity was about 0.4%. Because the part of the genome showing the lowest heterozygosity was the best assembled part, 0.4% is probably an underestimation due to sampling bias.

With a heterozygous genome, the assembler might assemble the two alleles for a site separately, which would result in an excess of the assembly with half of the average sequencing depth. To plot the sequencing depth distribution, we mapped all the sequencing reads to the assembly; we found that, although the major peak of sequencing depth was at ~100×, there was indeed a minor peak at ~50× (**Supplementary Fig. 5**), indicating that there were genomic regions with half of the average sequencing depth due to underassembly of allelic regions with high heterozygosity. Using the depth distribution, we estimated the length of the genome assembly from these heterozygous regions to be 135 Mb (**Supplementary Fig. 6**). We then identified 58,241 contigs with half of the average sequencing depth (total length of 131.2 Mb, consistent with the estimation of heterozygous region length) (**Supplementary Fig. 7**); these contigs explained the 50× peak for sequencing depth, as the depth distribution was normal after excluding them (**Supplementary Fig. 5** and **Supplementary Note**). The 2,454 genes from these contigs (**Supplementary Table 20**) were significantly over-represented in the GO terms 'apoptosis', (*P* value 4.19×10^{-6}) 'programmed cell death' (*P* value 4.19×10^{-6}) and 'defense response' (*P* value 1.77×10^{-4}) and are possibly related to self-incompatibility[19,20] (**Supplementary Fig. 8** and **Supplementary Table 21**). The heterozygous contigs suggest that there is a block-wise distribution of heterozygosity, and we further identified heterozygous SNPs in the genome to characterize their distribution (**Supplementary Fig. 9**). We indeed found that the 1.7 million high-quality heterozygous SNPs identified were not distributed randomly in the genome.

TE insertions in introns

We compared the gene models in 13 plant species and found that the average intron length for *P. equestris* was substantially longer than for the other species (**Table 1**). In comparing the distributions of TE proportions in introns, we identified a distinctive major peak near 45% in orchid (**Supplementary Fig. 10** and **Supplementary Table 22**). In addition, the proportion of genes with long introns seemed substantially higher in *P. equestris* than in most other species, even after excluding TEs (**Supplementary Fig. 11**). Further comparison of intron length distributions showed that *P. equestris* had the highest proportion of introns with a length of ≥2,000 bp (27.7%) (**Supplementary Table 23**). To explore the functional consequences of intronic TE insertions, we compared the expression levels of genes with TE insertions to those of corresponding paralogous genes without intronic TE insertions. Overall, the expression levels of genes with TE insertions were lower than those of their paralogs ($P < 1 \times 10^{-11}$, Wilcoxon rank-sum paired test) in all four tissues examined (**Supplementary Fig. 12**). A decrease in the expression levels of these genes can probably be explained by negative selection due to an increase in the transcription cost for the longer transcript, which is consistent with previous findings suggesting

Table 1 Comparison of the gene models in different species

	Am	At	Bd	Gr	Si	Mu	Os	Pe	Pb	Sb	Sm	Vv	Zm
Genome size (Mb)	706.30	119.70	271.90	775.20	400.10	473.00	382.80	1086.20	509.00	738.50	212.80	486.20	2058.60
Gene (transcript) number	26,846	26,637	26,413	40,976	38,125	36,536	33,127	29,431	42,341	27,159	22,285	25,329	78,483
Gene length (Mb)	151.50	50.87	75.61	101.86	115.82	131.34	72.40	283.85	117.94	79.90	36.63	155.24	183.15
Coding sequence (Mb)	25.40	33.10	33.98	45.25	41.58	37.97	33.45	26.42	56.99	34.25	25.53	29.82	62.16
Intron number	82,937	112,745	106,747	144,200	124,169	161,132	95,376	86,301	180,965	104,612	100,572	129,570	188,877
All Intron length (Mb)	126.69	17.76	41.63	48.90	54.79	93.37	38.95	252.19	60.95	45.66	11.14	125.42	121.00
Average intron length (bp)	1,528	158	390	339	441	579	408	2,922	337	436	111	968	641
Length of TEs in introns (Mb)	46.74	0.46	2.85	0.87	11.04	2.55	1.37	125.58	6.16	8.80	2.81	47.22	40.02
Total TEs (Mb)	425.66	23.87	77.43	79.43	186.03	31.90	97.99	667.90	190.44	435.33	102.15	239.96	1524.76
Total TE proportion (%)	60.30	19.90	28.50	10.20	46.50	6.70	25.60	61.50	37.40	58.90	48.00	49.40	74.10
TE proportion in introns (%)	36.89	2.56	6.84	1.78	20.15	2.73	3.53	49.80	10.11	19.27	25.17	37.65	33.07

Am, *Amborella trichopoda*; At, *Arabidopsis thaliana*; Bd, *Brachypodium distachyon*; Gr, *Gossypium raimondii*; Si, *Setaria italica*; Mu, *Musa acuminata*; Os, *Oryza sativa*; Pe, *Phalaenopsis equestris*; Pb, *Pyrus bretschneideri*; Sb, *Sorghum bicolor*; Sm, *Selaginella moellendorffii*; Vv, *Vitis vinifera*; Zm, *Zea mays*. For all species, information on the data used can be found in the **Supplementary Note** and **Supplementary Table 22**. TEs were extracted from the original data set, and proportions were calculated accordingly.

that natural selection favors short introns in highly expressed genes to minimize the cost of transcription and other molecular processes such as splicing[21]. Previous studies on paralogs formed by whole-genome duplication (WGD) showed significant differences in expression levels, with genes with weaker expression having a tendency to accumulate more transposon insertions[22].

Genome evolution

We constructed a phylogenetic tree on the basis of a concatenated sequence alignment of the 72 single-copy genes shared by orchid and 10 other green plant species (**Fig. 1b**). In this phylogenetic tree, orchid, as expected, clustered with other monocots, although the evolutionary distance from orchid to cereals such as rice, *Brachypodium* species, sorghum and maize was relatively large. Although the 72 genes already provided enough phylogenetic signals for phylogeny construction in the 11 green plant species, accurate dating of the divergence times between orchid and other monocots requires a larger gene set while compromising the phylogenetic coverage. Applying the PAML MCMCTree program to the 342 single-copy genes shared among orchid and 7 other species (monocots and dicots), the divergence time between *P. equestris* and the other monocots was estimated at 135.1 ± 17 million years ago, which is in line with estimates from both angiosperm-wide[23] and Orchidaceae-specific[13] molecular sequence divergence estimation studies. We also studied gene family expansion and contraction in different evolutionary lineages (**Fig. 1b**).

Like many other plant genomes sequenced thus far[24–27], the orchid genome harbored the remnants of one or more large-scale duplication events. Although only a small fraction of the genome (3.51%) showed collinearity (conservation of gene order and content) with other regions in the genome, this proportion most likely constitutes a substantial underestimate. Indeed, in total, 12,000 orchid genes resided on scaffolds with fewer than 20 genes, which are of limited use for the intragenome detection of collinearity. Furthermore, about 6,500 genes were located on scaffolds with fewer than 5 genes. However, a considerable number of genes (5,492) were contained in syntenic regions that showed conservation of gene content, regardless of gene order. The notable difference between retained homeologs in collinear versus syntenic regions can probably be explained by a high degree of reshuffling of genes after duplication, fractionation (loss of either homeolog) and the low gene density of *P. equestris* (which has about the same number of genes as *A. thaliana* with a genome size that is about ten times larger). In the absence of a close outgroup and because other sequenced monocot species diverged more than 100 million years ago, the above factors render multi-level collinearity with other sequenced angiosperm species hard to detect. However, analysis of the number of synonymous substitutions per synonymous site (K_S) for the whole paranome (the set of all duplicated genes in the genome), identified a peak between 0.6 and 1.1 that corresponds to contemporarily created gene duplicates (**Supplementary Fig. 13a**) and most likely represents an ancient WGD event[28]. Furthermore, when only duplicates retained in collinear regions were considered, duplicates from small-scale duplications were excluded from the distribution and the WGD signature peak in the K_S distribution became even more pronounced (**Supplementary Fig. 13b**). Putative peaks at higher K_S values might represent more ancient WGD events in the monocot lineage that might have been shared by orchids or their ancestors[29,30] (**Supplementary Fig. 13b**).

We performed absolute dating through phylogenomic analysis to establish the age of this paleopolyploidy in relation to the monocot phylogeny[27]. The absolute dating of genes present under the signature WGD peak in the K_S distribution rendered an absolute age distribution with a clear peak at ~76 million years ago (**Fig. 2a**). More specifically, the WGD event was dated at 75.57 million years ago, with lower and upper 90% confidence interval limits of 71.50 and 80.73 million years ago, respectively (**Supplementary Note**). As the common ancestor of the crown group of orchids is supposed to have lived during the Late Cretaceous period sometime between 76 and 84 million years ago[13], this finding would suggest that the orchid-specific WGD event occurred in association with the origin of this clade, and polyploidy is indeed proposed as a frequent mechanism of speciation in angiosperms[31,32]. In contrast, many members of the Orchidaceae family underwent drastic rate shifts (transition and transversion) during their evolutionary history due to periods of accelerated molecular evolution caused by their short life cycles and altered life history strategies[33–35]. These rate shifts complicate absolute dating[27], especially considering the long distant relationship from orchid to the other monocot species for which the complete genome sequence is currently available, such that the current WGD age estimate could be an overestimate, with the actual age most likely closer to the lower confidence interval boundary. This WGD age estimate would suggest that paleopolyploidy enabled survival across the Cretaceous-Paleogene boundary, as witnessed in many other angiosperms[36] (**Fig. 2b**). Determining whether

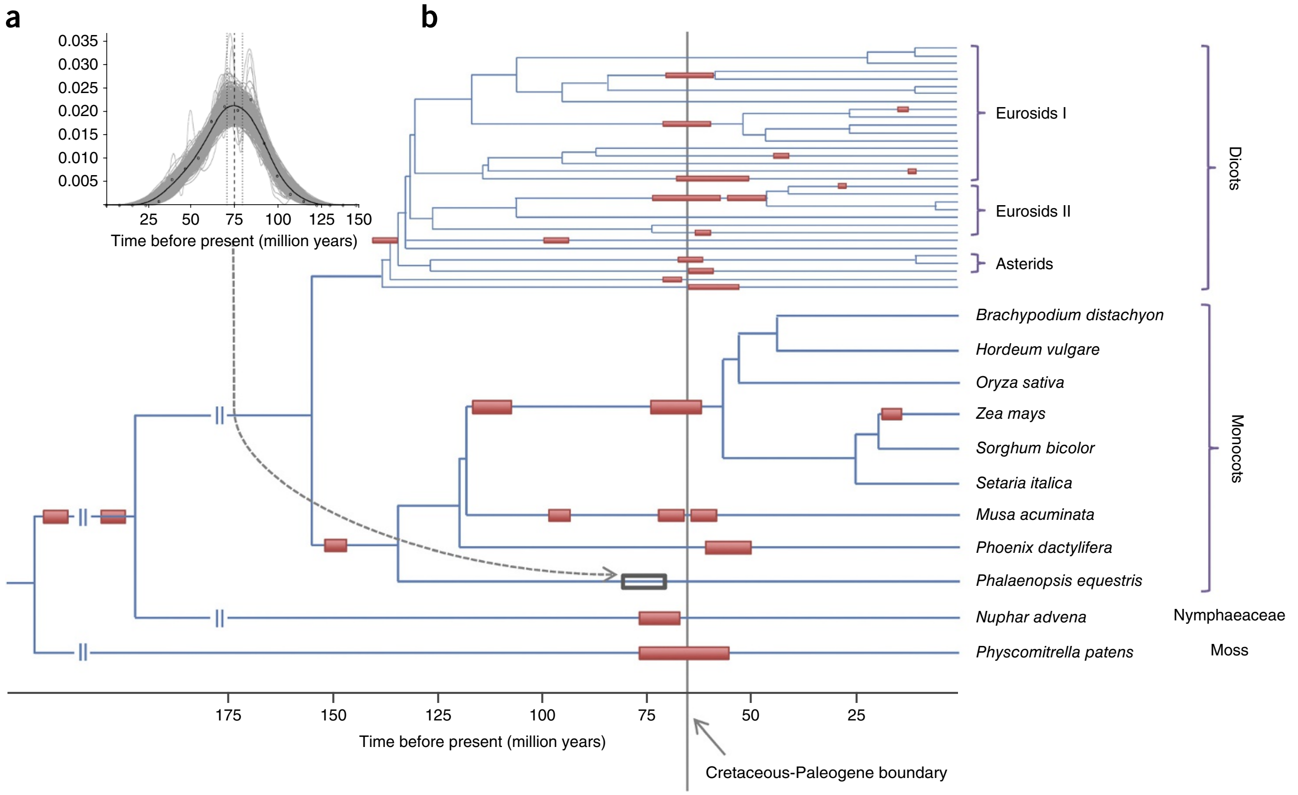

Figure 2 Dating the paleopolyploidy event in *P. equestris*. (**a**) Absolute age distribution obtained by phylogenomic dating of *P. equestris* homeologs. The solid black line represents the kernel density estimate of the dated homeologs, and the vertical dashed black line represents its peak, used as the WGD age estimate. Gray lines represent the density estimates for the 1,000 bootstrap replicates, and the vertical black dotted lines represent the corresponding 90% confidence intervals for the WGD age estimate. The original raw distribution of dated homeologs is also indicated by dots. The mode used as an estimate for the consensus WGD age is found at 75.57 million years ago with lower and upper 90% confidence interval boundaries at 71.50 and 80.73 million years ago, respectively *y* axis represents percentage of gene pairs. (**b**) Phylogenetic tree of the angiosperms. A wave of WGD events (indicated by colored bars) appears to be associated with the Cretaceous-Paleogene extinction event ~66 million years ago[27]. The orchid-specific WGD is indicated by the unfilled bar.

orchid-specific paleopolyploidy occurred in association with either its origin or the Cretaceous-Paleogene boundary will necessitate information on other non-cereal monocot genomes. Nevertheless, the orchid-specific paleopolyploidy identified in this study followed by the documented vast radiation of orchid not long after the Cretaceous-Paleogene boundary[13], which enabled Orchidaceae to become the second largest angiosperm plant family with a remarkable diversity in flower morphology[37], might suggest that the WGD event contributed to the success of orchids[38].

Evolutionary analysis of CAM genes

In contrast to all vascular plants, of which about 10% are estimated to be epiphytes[39], most orchid species (72%) are epiphytes, with the majority being restricted to tropical regions[40,41]. Many orchids use the CAM pathway for photosynthesis rather than the C_3 pathway, which is considered to be an adaptation to their epiphytic lifestyle that limits water supply. CAM is an important metabolic pathway that evolved convergently in many different plant lineages, and it has been estimated that CAM pathway components are encoded by at least 343 genera in 35 plant families, comprising ~6% of flowering plant species[42–44]. The CAM pathway bears resemblance to the C_4 pathway in that both act to concentrate CO_2 around RuBisCO, thereby increasing its efficiency. However, where C_4 plants concentrate CO_2 spatially, CAM plants concentrate CO_2 temporally, providing CO_2 during the day but not at night when respiration is the dominant reaction. *P. equestris* is the only CAM plant thus far for which the genome has been sequenced.

To gain further insight into the evolution of the CAM pathway from the ancestral C_3 pathway, we identified genes encoding the key enzymes of the CAM pathway (**Fig. 3** and **Supplementary Table 24**) and compared the *P. equestris* CAM genes with their homologs in Poaceae, including *O. sativa*, *S. bicolor* and *Z. mays*, using the dicot *A. thaliana* as an outgroup. We analyzed six key enzymes in CAM, namely, carbonic anhydrase (*CA*), malic enzyme (*ME*), malate dehydrogenase (*MDH*), pyruvate phosphate dikinase (*PPDK*), phosphoenolpyruvate carboxykinase (*PPCK*) and phosphoenolpyruvate carboxylase. Gene trees were constructed from alignments of the coding sequences for each gene family (**Supplementary Fig. 14** and **Supplementary Data Set**). We identified gene duplication and loss events along the lineage leading to *P. equestris* by manually inspecting each gene tree individually. In particular, we uncovered one gene loss and one gene duplication event in the *PEPC* gene family, two gene duplications in the *ME* gene family (in the *NADP-ME* subfamily), one gene loss in the *PPCK* gene family and at least six gene duplications in the *CA* gene family (**Fig. 3**). *CA* catalyzes the reaction converting carbon dioxide into carbonate, which is the first step in CO_2 fixation. There are two *CA* subfamilies (α and β) in *P. equestris*. The most obvious expansion for a gene family was found in the α *CA* gene family. However, the functional differentiation between the two gene families is still not clear. Gene family expansion might substantially increase

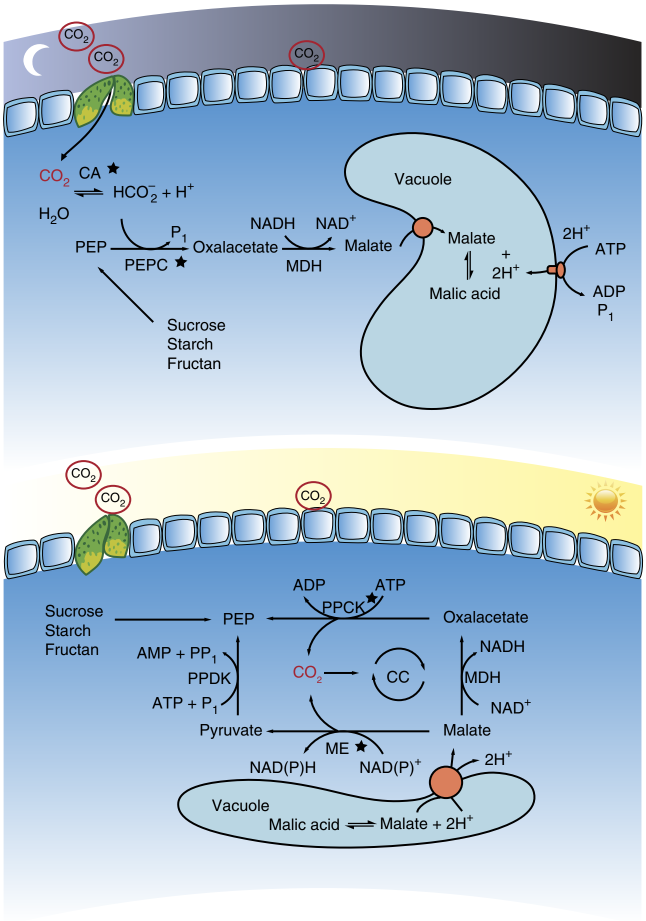

Figure 3 Overview of crassulacean acid metabolism (CAM) pathway evolution. Overview of the CAM pathway with components for which the respective gene family underwent gene duplication or loss as indicated. A star indicates components whose gene families underwent gene loss or gain. CA, carbonic anhydrase; CC, Calvin cycle; PEP, phosphoenolpyruvic acid; PEPC, phosphoenolpyruvate carboxylase; PPCK, phosphoenolpyruvate carboxykinase; MDH, malate dehydrogenase; ME, malic enzyme; PPDK, pyruvate phosphate dikinase.

the efficiency of the reaction through dosage effects and might also provide the possibility of adaptive evolution of the duplicated copies. We did not find any CAM genes among the retained duplicates in the WGD-derived homeologous segments and therefore did not find any obvious evidence that the paleopolyploidy event has been of crucial importance for CAM evolution in *P. equestris*.

MADS-box gene family analysis

MADS-box genes are known to be involved in many important processes during plant development but are especially known for their roles in flower development. Because orchids are famous for their flower morphology, we focused on identifying and characterizing the MADS-box genes in more detail. In total, 51 putative functional MADS-box genes and 9 pseudogenes were identified (**Table 2**). Perhaps surprisingly, these numbers are smaller than what is documented for most other sequenced angiosperms. *P. equestris* has 29 type II MADS-box genes, much fewer than the number found in rice (48) or other cereals. Phylogenetic analysis (**Supplementary Fig. 15**) showed that most of the genes in the type II MADS-box clades had been duplicated, except those in the B-PI, Bs, SVP and MIKC* clades. Among the duplicated type II clades, the E-class (six members), C/D-class (five members, three in C class and two in D class), B-class *AP3* (four members) and *AGL6* clades (three members) contained more genes than *A. thaliana* and rice (**Supplementary Table 25**). However, genes from the *FLC*, *AGL12* and *AGL15* clades could not be found in the *P. equestris* genome. Recently, *FLC* genes have been found in cereals, but they are difficult to identify because they are highly divergent and relatively short[45]. However, genes in both the *AGL12* and *AGL15* clades are present in the genomes of rice and *A. thaliana*; therefore, orthologs of *FLC*, *AGL12* and *AGL15* might have been specifically lost in orchids. Although *AGL12*-like genes (*XAL1* in *A. thaliana*) are necessary for root development and flowering[46], it seems that a different mechanism has evolved in *P. equestris* for the same function. Genes in the B-class *AP3*, C/D-class and E-class clades are well known for their roles in the specification of floral organ identity and have been well studied in *P. equestris*. These expanded clades including members with differential expression patterns in orchid floral organs, as well as divergent encoded protein domains, support the unique evolutionary routes of these floral organ identity genes associated with the unique labellum and gynostemium innovation in orchids[47–50]. Notably, one of the three gene copies in the expanded *AGL6* clade had an expression pattern similar to that of the *AP3*-like *PeMADS4* gene, which is specifically expressed in the labellum[51]. The *AGL6*-like gene *OsMADS6* could specify floral organ identities and meristem fate by interacting with the floral homeotic genes *SUPERWOMAN1*, *MADS3*, *MADS58*, *MADS13* and *DROOPING LEAF* in rice[52]. The *OsMADS6* gene product has also been shown to act as an integrator by forming complexes with B-class and C/D-class MADS-box proteins[53,54]. Combinatorial protein interaction networks among the members of the expanded B-class, C/D-class, E-class and *AGL6* clades during orchid floral development might have led to the evolutionary novelties of orchid flowers directly involved in speciation[50].

Table 2 MADS-box genes in the *P. equestris*, poplar, *A. thaliana* and rice genomes

	P. equestris		Poplar[a]		*A. thaliana*[a]		Rice[a]	
Category	Functional	Pseudo	Functional	Pseudo	Functional	Pseudo	Functional	Pseudo
Type II (total)	29	1	64	3	47	5	48	1
vMIKC[b]	28	1	55	2	43	4	47	1
vMIKC*	1	0	2	0	2	0	1	0
vMδ	0	0	7	1	4	1	0	0
Type I (total)	22	8	41	9	62	36	23	6
vMα	10	6	23	4	20	23	15	2
vMβ	0	0	12	5	17	5	9	1
vMγ	12	2	6	0	21	8	8	3
Total	51	9	105	12	107	41	80	7

Genes encoding a stop codon in the MADS-box domain were categorized as pseudogenes.
[a] Data from Leseberg *et al.*[57].

Only 22 putative functional type I genes and 8 pseudogenes were found (**Table 2**), suggesting that the *P. equestris* type I MADS-box genes have experienced a lower birth rate or, alternatively, a higher death rate than type II MADS-box genes. Tandem gene duplications seem to have contributed to the increase in the number of type I genes in the α group (type I Mα) and suggest that the type I MADS-box genes have mainly been duplicated by smaller-scale and more recent duplications[55]. Interestingly, the *P. equestris* genome does not contain the β group of type I MADS-box genes (type I Mβ), although these do exist in *A. thaliana*, poplar and rice. Interactions among type I MADS-box genes are important for the initiation of endosperm development[56].

We also determined the expression levels of all orchid MADS-box genes (**Supplementary Table 26**) and found that 20 of these genes were preferentially expressed in flower tissue. In particular, five genes (*PEQU_41930* (D), *PEQU_16438* (D), *PEQU_12328* (Bs), *PEQU_17261* (Mα) and *PEQU_09539* (MIKC*)) were exclusively expressed in flower, suggestive of a distinct role in orchid floral morphogenesis.

DISCUSSION

Here we have presented a high-quality draft sequence of *P. equestris*, the first orchid for which the genome sequence has been determined. All around the world, orchids are highly endangered species because of illegal collection and loss of habitat. The complete genome sequence of *P. equestris* will provide an important resource to start exploring orchid diversity and evolution at the genome level, which will be important for ecological and conservation purposes. The genome sequence will also be a key resource for the development of new concepts and techniques in genetic engineering, such as molecular marker–assisted breeding and the production of transgenic plants, which are necessary to increase the efficiency of orchid breeding and aid orchid horticulture research.

URLs. OrchidBase, http://orchidbase.itps.ncku.edu.tw/; Angiosperm Phylogeny Website, http://www.mobot.org/MOBOT/research/APweb/; World Checklist of Orchidaceae, http://apps.kew.org/wcsp/; RepeatProteinMask, http://www.repeatmasker.org/cgi-bin/RepeatProteinMaskRequest; EMBOSS, http://emboss.sourceforge.net/; BESTORF, http://linux1.softberry.com/berry.phtml?topic=bestorf&group=help&subgroup=gfind.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing data, annotations and analyses results have all been uploaded to the FTP site ftp://ftp.genomics.org.cn/from_BGISZ/20130120/ for evaluation. The data have also been submitted to the NCBI database under BioProject PRJNA192198.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J.C. Pires for helpful discussions. The authors acknowledge support from the 948 Program of the State Forestry Administration (China) (2011-4-53) to Z.-J.L. and X.-M.Z., the Shenzhen Municipal Science & Technology Programs for Building State and Shenzhen Key Laboratories (no. 2006464, 200712 and CXB201005260070A) to L.-Q.H., the Youth Talent Fund of the Shenzhen municipal government (JC201005310690A) to J.C. and the Ghent University Multidisciplinary Research Partnership 'Bioinformatics: From Nucleotides to Networks' to Y.V.d.P. Y.V.d.P. also acknowledges support from the European Union Seventh Framework Programme (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739–DOUBLE-UP. Part of this work was carried out using the Stevin Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Hercules Foundation and the Flemish Government Department of Economy, Science and Innovation.

AUTHOR CONTRIBUTIONS

J.C., Z.-J.L., L.-Q.H., J.W., H.-H.C., Y.V.d.P., X.L., S.P., K.V., W.-C.T., Y.-B.L., K.-W.L., X.-M.Z. and R.D. planned and coordinated the project and wrote the manuscript. W.-C.T., Y.-Y.S., Z.-J.P., C.-C.H., Y.-P.Y., Y.-C.H., Y.-C.C., L.-J.C., X.-J.X., G.-Q.Z., M.W., G.-C.X., G.-H.L. and L.-Q.L. collected and grew the plant material. J.C., W.-C.T., K.-W.L., L.-J.C. and Q.X. prepared samples. X.L., C.B., Z.Z., W.L., F.S. and K.-W.L. sequenced and processed the raw data. X.L., C.B., Z.Z., X.X., J.W. and F.S. annotated the genome. C.B., X.L., J.C., Y.H. and S.P. analyzed the gene families. C.B., X.L., S.P., K.V., Y.H. and Z.Z. conducted genome evolution analysis. J.C. conducted CAM analysis. J.C., X.L. and F.-M.S. conducted TE insertion analysis. Z.-J.L. and J.-Y.W. conducted protein kinase analysis. W.-C.T., Y.-Y.H., Z.-J.P., C.-C.H., Y.-P.Y., Y.-C.H., Y.-C.C., A.D. and J.-F.D. conducted the MADS-box gene analysis. X.L., J.C. and K.-W.L. conducted transcriptome sequencing and analysis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Darwin, C. *On the Various Contrivances by Which British and Foreign Orchids are Fertilised by Insects* (Cambridge University Press, 2011).
2. Schiestl, F.P. *et al.* The chemistry of sexual deception in an orchid-wasp pollination system. *Science* **302**, 437–438 (2003).
3. Cozzolino, S. & Widmer, A. Orchid diversity: an evolutionary consequence of deception? *Trends Ecol. Evol.* **20**, 487–494 (2005).
4. Silvera, K., Santiago, L.S., Cushman, J.C. & Winter, K. Crassulacean acid metabolism and epiphytism linked to adaptive radiations in the Orchidaceae. *Plant Physiol.* **149**, 1838–1847 (2009).
5. Lin, S. *et al.* Nuclear DNA contents of *Phalaenopsis* sp. and *Doritis pulcherrima*. *J. Am. Soc. Hortic. Sci.* **126**, 195–199 (2001).
6. Leitch, I.J. *et al.* Genome size diversity in orchids: consequences and evolution. *Ann. Bot.* **104**, 469–481 (2009).
7. Tsai, W.C. *et al.* OrchidBase 2.0: comprehensive collection of Orchidaceae floral transcriptomes. *Plant Cell Physiol.* **54**, e7 (2013).
8. Fu, C.H. *et al.* OrchidBase: a collection of sequences of the transcriptome derived from orchids. *Plant Cell Physiol.* **52**, 238–243 (2011).
9. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
10. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* **102**, 5454–5459 (2005).
11. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
12. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
13. Ramírez, S.R., Gravendeel, B., Singer, R.B., Marshall, C.R. & Pierce, N.E. Dating the origin of the Orchidaceae from a fossil orchid with its pollinator. *Nature* **448**, 1042–1045 (2007).
14. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
15. Proost, S. *et al.* PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* **21**, 3718–3731 (2009).
16. Chen, S. *et al.* *De novo* analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS ONE* **5**, e15633 (2010).
17. O'Donoghue, E.M., Somerfield, S.D. & Heyes, J.A. Organization of cell walls in *Sandersonia aurantiaca* floral tissue. *J. Exp. Bot.* **53**, 513–523 (2002).
18. Lüttge, U. *Vascular Plants as Epiphytes: Evolution and Ecophysiology* (Springer-Verlag, 1989).
19. Bosch, M., Poulter, N.S., Vatovec, S. & Franklin-Tong, V.E. Initiation of programmed cell death in self-incompatibility: role for cytoskeleton modifications and several caspase-like activities. *Mol. Plant* **1**, 879–887 (2008).
20. Dixit, R. & Nasrallah, J.B. Recognizing self in the self-incompatibility response. *Plant Physiol.* **125**, 105–108 (2001).
21. Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V. & Kondrashov, F.A. Selection for short introns in highly expressed genes. *Nat. Genet.* **31**, 415–418 (2002).
22. Schnable, J.C., Springer, N.M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. USA* **108**, 4069–4074 (2011).
23. Bell, C.D., Soltis, D.E. & Soltis, P.S. The age and diversification of the angiosperms re-revisited. *Am. J. Bot.* **97**, 1296–1303 (2010).
24. Cui, L. *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).
25. Proost, S., Pattyn, P., Gerats, T. & Van de Peer, Y. Journey through the past: 150 million years of plant genome evolution. *Plant J.* **66**, 58–65 (2011).
26. Soltis, D.E. *et al.* Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336–348 (2009).

27. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
28. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* **30**, 177–190 (2013).
29. Kim, C. *et al.* Comparative analysis of Miscanthus and Saccharum reveals a shared whole-genome duplication but different evolutionary fates. *Plant Cell* **26**, 2420–2429 (2014).
30. Jiao, Y., Li, J., Tang, H. & Paterson, A.H. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* **26**, 2792–2802 (2014).
31. Wood, T.E. *et al.* The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. USA* **106**, 13875–13879 (2009).
32. Soltis, D.E., Visger, C.J. & Soltis, P.S. The polyploidy revolution then…and now: Stebbins revisited. *Am. J. Bot.* **101**, 1057–1078 (2014).
33. Whitten, W.M. *et al.* Molecular phylogenetics of *Maxillaria* and related genera (Orchidaceae: Cymbidieae) based on combined molecular data sets. *Am. J. Bot.* **94**, 1860–1889 (2007).
34. Whitten, W.M., Williams, N.H. & Chase, M.W. Subtribal and generic relationships of Maxillarieae (Orchidaceae) with emphasis on Stanhopeinae: combined molecular evidence. *Am. J. Bot.* **87**, 1842–1856 (2000).
35. Douzery, E.J. *et al.* Molecular phylogenetics of diseae (Orchidaceae): a contribution from nuclear ribosomal ITS sequences. *Am. J. Bot.* **86**, 887–899 (1999).
36. Vanneste, K., Maere, S. & Van de Peer, Y. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Phil. Trans. R. Soc. Lond. B* **369**, 20130353 (2014).
37. Campbell, C.S., Judd, W.S. & Kellogg, E.A. *Plant Systematics: A Phylogenetic Approach* (Sinauer Associates, 1999).
38. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009).
39. Lüttge, U. Ecophysiology of crassulacean acid metabolism (CAM). *Ann. Bot.* **93**, 629–652 (2004).
40. Benzing, D.H. Vascular epiphytism: taxonomic participation and adaptive diversity. *Ann. Mo. Bot. Gard.* **74**, 183–204 (1987).
41. Gravendeel, B., Smithson, A., Slik, F.J. & Schuiteman, A. Epiphytism and pollinator specialization: drivers for orchid diversity? *Phil. Trans. R. Soc. Lond. B* **359**, 1523–1535 (2004).
42. Pospišilová, J. Vascular plants as epiphytes. Evolution and ecophysiology. *Biol. Plant.* **33**, 500 (1991).
43. Holtum, J.A., Winter, K., Weeks, M.A. & Sexton, T.R. Crassulacean acid metabolism in the ZZ plant, *Zamioculcas zamiifolia* (Araceae). *Am. J. Bot.* **94**, 1670–1676 (2007).
44. Silvera, K. *et al.* Evolution along the crassulacean acid metabolism continuum. *Funct. Plant Biol.* **37**, 995–1010 (2010).
45. Ruelens, P. *et al.* FLOWERING LOCUS C in monocots and the tandem origin of angiosperm-specific MADS-box genes. *Nat. Commun.* **4**, 2280 (2013).
46. Tapia-López, R. *et al.* An *AGAMOUS*-related MADS-box gene, *XAL1* (*AGL12*), regulates root meristem cell proliferation and flowering transition in *Arabidopsis*. *Plant Physiol.* **146**, 1182–1192 (2008).
47. Pan, Z.J. *et al.* The duplicated B-class MADS-box genes display dualistic characters in orchid floral organ identity and growth. *Plant Cell Physiol.* **52**, 1515–1531 (2011).
48. Tsai, W.C. *et al.* Interactions of B-class complex proteins involved in tepal development in *Phalaenopsis* orchid. *Plant Cell Physiol.* **49**, 814–824 (2008).
49. Tsai, W.C., Kuoh, C.S., Chuang, M.H., Chen, W.H. & Chen, H.H. Four DEF-like MADS box genes displayed distinct floral morphogenetic roles in *Phalaenopsis* orchid. *Plant Cell Physiol.* **45**, 831–844 (2004).
50. Pan, Z.J. *et al.* Flower development of *Phalaenopsis* orchid involves functionally divergent *SEPALLATA*-like genes. *New Phytol.* **202**, 1024–1042 (2014).
51. Hsiao, Y.Y. *et al.* Transcriptomic analysis of floral organs from *Phalaenopsis* orchid by using oligonucleotide microarray. *Gene* **518**, 91–100 (2013).
52. Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
53. Seok, H.Y. *et al.* Rice ternary MADS protein complexes containing class B MADS heterodimer. *Biochem. Biophys. Res. Commun.* **401**, 598–604 (2010).
54. Favaro, R. *et al.* Ovule-specific MADS-box proteins have conserved protein-protein interactions in monocot and dicot plants. *Mol. Genet. Genomics* **268**, 152–159 (2002).
55. Parenicová, L. *et al.* Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. *Plant Cell* **15**, 1538–1551 (2003).
56. Masiero, S., Colombo, L., Grini, P.E., Schnittger, A. & Kater, M.M. The emerging importance of type I MADS box transcription factors for plant reproduction. *Plant Cell* **23**, 865–872 (2011).
57. Leseberg, C.H., Li, A., Kang, H., Duvall, M. & Mao, L. Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. *Gene* **378**, 84–94 (2006).

ONLINE METHODS

Sample preparation and sequencing. For genome sequencing, we collected leaves and flowers from several individuals of an inbred line of *P. equestris* and extracted genomic DNA using the modified CTAB protocol[58]. Sequencing libraries with insert sizes ranging from 160 bp to 20 kb were then constructed using a library construction kit (Illumina). Libraries were sequenced on the Illumina HiSeq 2000 platform. A library with an insert size of 40 kb was constructed using a modified fosmid library construction pipeline as described previously[59]. The raw reads generated were filtered according to sequencing quality and with regard to adaptor contamination and duplicated reads. Thus, only high-quality reads were remained and used in the genome assembly.

Genome assembly and assessment. We adopted a whole-genome shotgun strategy to sequence and assemble the genome of *P. equestris* and obtained 119.4 Gb of data from seven DNA libraries (**Supplementary Table 1**). Using *k*-mer frequency distribution analysis, we estimated the genome size to be 1.16 Gb (**Supplementary Fig. 1** and **Supplementary Note**), which is smaller than previous estimates by flow cytometry (~1.6 Gb; ref. 5). In the *k*-mer distribution, we also observed a secondary peak indicating considerable heterozygosity, which posed serious challenges for the genome assembly algorithm. In general, genome assembly was carried out using SOAPdenovo[60]. Contig construction, scaffolding and gap filling processes were performed using the corresponding methods provided by SOAPdenovo. The parameters used in genome assembly were as follows: pregraph -s Pha.lib -a 200 -p 12 -K 35 -d 2 -o Pha; contig -g Pha_1213 -M 3; map -s Pha.lib -g Pha; scaff -g Pha. Using these data, we assembled 236,185 scaffolds, with a total length of ~1.1 Gb and an N50 length of 359.1 kb (**Supplementary Table 2**). About 90% of the total assembled genome (~980 Mb) was contained in the 6,359 longest scaffolds.

The total genome assembly amounted to 1.086 Gb (1.0 Gb without N bases), which is ~93% (~86% without N bases) of the estimated total genome size. We assessed the coverage of the genome assembly using BACs (**Supplementary Note**). First, 18,486 BAC-end sequences (>100 bp in length) from a previous study[61] and our data (W.-C. Tsai and H.-H. Chen, unpublished data) were downloaded and mapped back to the scaffolds using BLAT[62]. Of the BAC-end sequences, 92.9% could be mapped to the assembled genome. We then used a pooling strategy to sequence the BAC clones from a BAC library of *P. equestris*. We made mixtures of 10, 20 and 30 randomly selected BAC clones, each with an independent replication. Each of the six pooled BAC clones was amplified with liquid medium and then used to extract BAC DNA. The pooled BAC DNA was sheared to generate sequencing libraries of short insert size, and these libraries were sequenced. For each BAC pool, we obtained more than 5 Gb of data on average. Taking these data, we used SOAPdenovo to assemble each BAC pool. Although the overall assembly results of these BACs had some gaps (**Supplementary Table 27**), we were able to generate some long contigs and use these to assess the quality of the whole-genome assembly. We mapped these long contigs back to the genome assembly with BLASTZ. The mapped length was then calculated, and the mapping details were displayed. For contigs longer than 1 kb in length (979 in total), 97% could be mapped to the assembled genome (**Supplementary Fig. 2**). Finally, we sequenced and assembled ten randomly chosen BAC clones using 454 sequencing technology. Comparison of these assembled BAC scaffolds with our assembly also indicated a low error rate (**Supplementary Note**).

We also assessed the completeness and accuracy of the assembly using conserved genes and RNA-seq data (**Supplementary Note**). Of the 248 conserved core eukaryotic genes that were used to assess genome completeness[6], 234 (94%) were uncovered in our genome assembly (**Supplementary Table 5**). Using the ~9 Gb of RNA-seq data from four different *P. equestris* tissues (leaf, flower, stem and root), we assembled transcription fragments, and 93% (root) and ~97% (leaf) of the assembled sequences could be mapped to the genome assembly with 90% identity and 90% mapping coverage (**Supplementary Table 6**). Thus, the coverage of the assembly in gene-rich regions was estimated to be > 93%, which was higher than the estimated genome coverage overall.

Repeat annotation. Tandem repeats and TEs in the genome were identified separately. The repeat annotation process was similar to that applied and described in a previous study[59]. Tandem repeats were identified using TRF[63] and RepeatMasker[64] (version 3.2.7). To identify TEs, we first used RepeatMasker with the Repbase[65] database of known repeat sequences to search the TEs in the orchid genome. We then used LTR_FINDER[66] (version 1.0.3), PILER[67] and RepeatScout[68] (version 1.05) to construct a repeat sequence database for orchid. Further, applying this *de novo* repeat sequence database, we used RepeatMasker to search for repeats in the genome. We also used RepeatProteinMask (version 3.2.2) implemented in RepeatMasker to identify repeat proteins. All the repeat sequences identified by the different methods were combined into the final repeat annotation. The repeat elements were categorized in a hierarchical way, as described previously[59].

To study the divergence of LTRs, we identified LTRs with complete structure using LTR_STRUC[69] with default parameters. Divergence was then estimated as described previously[59]. The LTRs with complete structure were aligned using MUSCLE[70], and divergence was estimated using the Kimura two-parameter model in distmat implemented in the EMBOSS package.

Gene annotation. Gene annotation was performed as described previously[59]. We carried out gene annotation using the following methods: (i) *de novo* gene prediction, (ii) homolog prediction, (iii) RNA-seq annotation and (iv) integration of the gene set. First, two *de novo* gene prediction programs, AUGUSTUS[71] (version 2.03) and SNAP[72], were applied to predict genes on the masked genome sequences. In homolog prediction, we utilized the protein sequences from four angiosperm genomes (*A. thaliana*, *O. sativa*, *S. bicolor* and *Z. mays*) to align against the unmasked genome using TBLASTN, with an *E*-value cutoff of 1 × 10^{-5}, and then used Genewise[73] (version 2.2.0) to predict gene structures. In RNA-seq annotation, the RNA-seq reads from four tissues were aligned against the reference genome using TopHat[74] (version 1.0.14). After alignment, the transcripts were assembled using Cufflinks[75] (version 0.8.2). BESTORF was then used to predict ORFs with parameters trained on monocot genes without filtering out UTRs. Finally, we generated an integrated gene set using GLEAN[76]. Details on software parameters are provided in the **Supplementary Note**.

To detect alternative splicing, we first used an in-house script to train fifth-order Markov parameters with our annotated gene set. Using this training set, we predicted an ORF for each transcript generated by Cufflinks. Transcripts without predicted ORFs were discarded. The transcripts predicted were compared to our gene models. Redundant transcripts, transcripts encoding short proteins (<50 amino acids in length) and transcripts for which the protein product was shorter than 30% of the proteins encoded in the gene set were filtered out.

For gene model validation, we manually examined 500 randomly selected genes from 2,038 of the gene families shared among monocots that only had one copy in both *P. equestris* and rice. Most of these 500 genes were correctly predicted, although 20 contained dubious annotation errors, reflecting an annotation accuracy of 96%. We further manually checked the well-studied MADS-box genes in *P. equestris* and found that the annotation of those genes was consistent with previous gene models determined by the sequencing of full-length cDNAs (W.-C.T., Y.-Y.H., K.-W.L., Z.-J.L., G.-Q.Z. *et al.*, unpublished data). Of 40 genes, only 3 had incomplete annotations, again reflecting the high quality of the gene annotation.

Building gene families. To build gene families, the PLAZA pipeline[15] was used. Along with the orchid genome, we included the genomes of four monocots (*O. sativa*, *B. distachyon*, *S. bicolor* and *Z. mays*), three eudicots (*A. thaliana*, *P. trichocarpa* and *V. vinifera*) and three outgroup species (*P. patens*, *C. reinhardtii* and *O. lucimarinus*). First, all pairwise similarities between the 364,344 coding genes in the data set were calculated using all-against-all BLASTP[77], retaining the top 1,250 hits for each gene with an *E*-value cutoff of 1 × 10^{-5}. Using tribeMCL[78], these similarities were clustered into homologous gene families (in mclblastline, using $I = 2$ and scheme = 4; other parameters were left at default values).

Genes included in a family with a similarity score (BLASTP) of less than 25% of the median similarity score for gene pairs within that family were flagged as outliers and were not included in the alignment and phylogenetic tree. For each family, the amino acid sequences encoded by all genes were obtained and aligned using MUSCLE. Ambiguously aligned positions (sites with gaps in the majority of the sequences and misalignments) were automatically removed from the alignments. Note that each singleton, for example, a gene with no homologs, was considered to represent a separate gene family (as shown in **Fig. 1a**).

Functional annotation. Two methods were used for functional annotation. First, InterProScan[14] (using default settings) was run to map known protein domains to all genes. Using InterPro to GO mapping, GO labels were obtained for the InterPro domains. Second, on the basis of phylogenetic trees, reliable rice orthologs were identified for orchid genes. Functional annotations from the orthologous rice genes, excluding those with an evidence tag of 'inferred from electronic annotation', were transferred to the orchid genes.

Detection of genomic homology. Genomic homology was detected using i-ADHoRe 3.0 (ref. 79), included in the PLAZA pipeline, using the following settings: alignment method gg2, gap size 30, tandem gap 30, cluster gap 35, q value 0.85, prob cutoff 0.01, anchor points 5 and multiple hypothesis correction FDR). The output was processed by the pipeline and included in a relational database to which visualization programs can connect and on which additional statistical analysis can be performed. For synteny detection, the cloud mode was enabled (cluster_type = cloud) and appropriate settings were selected: cloud_gap_size 20, cloud_cluster_gap 20, cloud_filter_method binomial, prob cutoff 0.01, anchor points 5, multiple hypothesis correction FDR and level_2_only true.

Relative dating using synonymous substitutions. K_S values for homologous gene pairs were calculated by first aligning the coding sequences with ClustalW[80] using the protein sequences as a guide. Positions aligned with low confidence (regions near gaps in the alignment) were stripped. Codeml (PAML package[81]) was used to determine the actual K_S value of each pair. To build the orchid paranome (all duplicated genes) K_S age distribution, a correction was performed as described in Maere *et al.*[10].

Gene family evolution and phylogenomic dating. We used 72 single-copy families shared by all 11 species to construct a phylogenetic species tree. We applied the Café program[82] to identify gene families that had undergone expansions or contractions.

To date the divergence times of orchid and other monocots, we used 342 single-copy gene families shared by orchid, the other 4 monocots (*O. sativa*, *B. distachyon*, *S. bicolor* and *Z. mays*) and 3 eudicots (*A. thaliana*, *P. trichocarpa* and *V. vinifera*). The PAML MCMCTree program was used to estimate species divergence times, with the options 'correlated molecular clock' and 'JC69' model. The 'alpha' parameter was estimated by PhyML using the same set of sequence data. MCMC analysis was run for 20,000 generations, using a burn-in of 1,000 iterations. Other parameters were left at default settings. Phase 1 (nondegenerate) sequences for all single-copy gene families that were identified by the PLAZA pipeline were used as the input file for PhyML and MCMCTree. We used the *O. sativa* and *B. distachyon* divergence time (40–54 million years ago[83]), the *P. trichocarpa* and *A. thaliana* divergence time (100–120 million years ago[84]) and the monocot and eudicot divergence time (130–240 million years ago[11]) as calibrators to predict the divergence time of other nodes and obtained a predicted divergence time of 135.1 million years ago (95% credibility interval of 118.1–152.7 million years ago) for *P. equestris* and cereals.

Evolutionary analysis of CAM genes. We identified putative CAM genes by searching the InterProScan result of all predicted *P. equestris* proteins. We identified orthologs of the *P. equestris* CAM genes in the genomes of other species, including *A. thaliana*, *Z. mays*, *O. sativa* and *S. bicolor*, using a reciprocal best hit (RBH) strategy implemented with NCBI BLAST and custom scripts. We then aligned the coding sequences of each gene family using MUSCLE implemented in MEGA5 (ref. 85). Before constructing the phylogeny, we removed dubious short reading frames and obviously unrelated genes resulting from the relaxed annotation of InterProScan. A gene tree was then constructed with MEGA5 using maximum likelihood for each gene family.

MADS-box gene analysis. MADS-box genes were identified by searching the InterProScan result of all predicted *P. equestris* proteins. MADS-box domains comprising 60 amino acids, identified by SMART[86] for all the MADS-box genes, were then aligned using ClustalW. An unrooted neighbor-joining phylogenetic tree was constructed in MEGA5 with default parameters. Bootstrap analysis was performed using 1,000 iterations.

58. Murray, M.G. & Thompson, W.F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325 (1980).
59. Zhang, G. *et al.* Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* **30**, 549–554 (2012).
60. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
61. Hsu, C.-C. *et al.* An overview of the *Phalaenopsis* orchid genome through BAC end sequence analysis. *BMC Plant Biol.* **11**, 3 (2011).
62. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
63. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
64. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* Chapter 4, Unit 4.10 (2009).
65. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
66. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
67. Edgar, R.C. & Myers, E.W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (suppl. 1), i152–i158 (2005).
68. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–i358 (2005).
69. McCarthy, E.M. & McDonald, J.F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362–367 (2003).
70. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
71. Stanke, M. *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
72. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
73. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
74. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
75. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
76. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
77. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
78. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
79. Proost, S. *et al.* i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, e11 (2012).
80. Thompson, J.D., Gibson, T.J. & Higgins, D.G. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* Chapter 2, Unit 2.3 (2002).
81. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
82. De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
83. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
84. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
85. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
86. Letunic, I. *et al.* Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* **30**, 242–244 (2002).

doi:10.1038/ng.3149

Corrigendum: The genome sequence of the orchid *Phalaenopsis equestris*

Jing Cai, Xin Liu, Kevin Vanneste, Sebastian Proost, Wen-Chieh Tsai, Ke-Wei Liu, Li-Jun Chen, Ying He, Qing Xu, Chao Bian, Zhijun Zheng, Fengming Sun, Weiqing Liu, Yu-Yun Hsiao, Zhao-Jun Pan, Chia-Chi Hsu, Ya-Ping Yang, Yi-Chin Hsu, Yu-Chen Chuang, Anne Dievart, Jean-Francois Dufayard, Xun Xu, Jun-Yi Wang, Jun Wang, Xin-Ju Xiao, Xue-Min Zhao, Rong Du, Guo-Qiang Zhang, Meina Wang, Yong-Yu Su, Gao-Chang Xie, Guo-Hui Liu, Li-Qiang Li, Lai-Qiang Huang, Yi-Bo Luo, Hong-Hwa Chen, Yves Van de Peer & Zhong-Jian Liu

***Nat. Genet.* 47, 65–72 (2015); published online 24 November 2014; corrected after print 9 January 2015**

In the version of this article initially published, the divergence time estimates in Figure 1b were incorrect, and the genome size estimation for *Phalaenopsis equestris* was incorrectly stated as 1.6 × 10^6 instead of the correct 1.6 × 10^9. Finally, Figure 3 incorrectly showed maleic acid in the vacuole reaction, which should have been malic acid. The errors have been corrected in the HTML and PDF versions of the article.

Corrigendum: The genome sequence of the orchid *Phalaenopsis equestris*

Jing Cai, Xin Liu, Kevin Vanneste, Sebastian Proost, Wen-Chieh Tsai, Ke-Wei Liu, Li-Jun Chen, Ying He, Qing Xu, Chao Bian, Zhijun Zheng, Fengming Sun, Weiqing Liu, Yu-Yun Hsiao, Zhao-Jun Pan, Chia-Chi Hsu, Ya-Ping Yang, Yi-Chin Hsu, Yu-Chen Chuang, Anne Dievart, Jean-Francois Dufayard, Xun Xu, Jun-Yi Wang, Jun Wang, Xin-Ju Xiao, Xue-Min Zhao, Rong Du, Guo-Qiang Zhang, Meina Wang, Yong-Yu Su, Gao-Chang Xie, Guo-Hui Liu, Li-Qiang Li, Lai-Qiang Huang, Yi-Bo Luo, Hong-Hwa Chen, Yves Van de Peer & Zhong-Jian Liu

Nat. Genet. **47, 65–72 (2015); published online 24 November 2014; corrected after print 9 January 2015; corrected after print 6 February 2015**

In the version of this article initially published, the legend for Figure 1b referred to red arrows indicating the inferred divergence dates. No arrows are depicted in the figure, so this sentence has been removed from the figure legend in the HTML and PDF versions of the article.