The Apostasia genome and the evolution of orchids

Guo-Qiang Zhang1, Ke-Wei Liu1, Zhen Li2,3, Rolf Lohaus2,3, Yu-Yun Hsiao4,5, Shan-Te Ni1,6, Jie-Yu Wang1,7, Yao-Cheng Lin8,9, Qing Xu1, Li-Jun Chen1, Kouki Yoshida8, Sumire Fujibara9, Zhi-Wei Wang10, Yong-Qiang Zhang11, Nobutaka Mitsuda12, Meina Wang4, Guo-Hui Liu1, Lorenzo Pecoraro1, Hui-Xia Huang1, Xin-Ju Xiao1, Min Lin1, Xin-Yi Wu1, Wan-Lin Wu1,4, You-Yi Chen1,3, Song-Bin Chang1,2, Shingo Sakamoto9, Masaru Ohme-Takagi11,12, Masafumi Yagi12, Si-Jin Zeng3,7, Ching-Yu Shen13, Chuan-Ming Yeh1, Y-Chi Lo4,13, Wen-Chieh Tsai4,5,13, Yves Van de Peer2,3,14 & Zhong-Jian Liu1,7,15,16

Constituting approximately 10% of flowering plant species, orchids (Orchidaceae) display unique flower morphologies, possess an extraordinary diversity in lifestyle, and have successfully colonized almost every habitat on Earth1,2. Here we report the draft genome sequence of Apostasia shenzhenica3, a representative of one of two genera that form a sister lineage to the rest of the Orchidaceae, providing a reference for inferring the genome content and structure of the most recent common ancestor of all extant orchids and improving our understanding of their origins and evolution. In addition, we present transcriptome data for representatives of Vanilloideae, Cypripedioideae and Orchidoideae, and novel third-generation genome data for two species of Epidendroideae, covering all five orchid subfamilies. A. shenzhenica shows clear evidence of a whole-genome duplication, which is shared by all orchids and occurred shortly before their divergence. Comparisons between A. shenzhenica and other orchids and angiosperms also permitted the reconstruction of an ancestral orchid gene toolkit. We identify new gene families, gene family expansions and contractions, and changes within MADS-box gene classes, which control a diverse suite of developmental processes, during orchid evolution. This study sheds new light on the genetic mechanisms underpinning key orchid innovations, including the development of the labellum and gynostemium, pollinia, and seeds without endosperm, as well as the evolution of epiphytism; reveals relationships between the Orchidaceae subfamilies; and helps clarify the evolutionary history of orchids within the angiosperms.

The Apostasiioideae are a small subfamily of orchids that includes only two genera (Apostasia and Neuwiedia2-5), consisting of terrestrial species confined to the humid areas of southeast Asia, Japan, and northern Australia6. Although Apostasiioideae share some synapomorphies with other orchids (for example, small seeds with a reduced embryo and a myco-heterotrophic protocorm stage), they possess several unique traits, the most conspicuous of which is their floral morphology7. Apostasia has a non-resupinate, solanum-type flower with anthers closely encircling the stigma (including post-genital fusion), a long ovary, and an actinomorphic perianth with an undifferentiated labellum. Three stamens (two of which are fertile) are basally fused to the style, forming a relatively simple gynostemium, and the anthers contain powdery pollen (grains not unified into pollinia). These characteristics (Extended Data Fig. 1a) differ from those of other Orchidaceae subfamilies, which have three sepals, three petals (of which one has specialized to form the labellum), and stamens and pistil fused into a more complex gynostemium (Extended Data Fig. 1b), but are similar to those of some species of Hypoxidaceae (a sister family to Orchidaceae, in the order Asparagales).

We sequenced the A. shenzhenica genome using a combination of different approaches; the total length of the final assembly was 349 Mb (see Methods and Supplementary Tables 1–4). We confidently annotated 21,841 protein-coding genes, of which 20,202 (92.50%) were supported by transcriptome data (Supplementary Fig. 1 and Supplementary Table 5). Using single-copy orthologues, we performed a BUSCO assessment that indicated that the completeness of the genome was 93.62%, suggesting that the A. shenzhenica genome assembly is of high quality (Supplementary Table 6). For comparative analyses, we also improved the quality of the previously published genome assemblies of the orchids Phalaenopsis equestris8 and Dendrobium catenatum9 (see Methods and Supplementary Tables 6 and 7).

We constructed a high-confidence phylogenetic tree and estimated the divergence times of 15 plant species using genes extracted from a total of 439 single-copy families (Fig. 1 and Extended Data Fig. 2). We undertook a computational analysis of gene family sizes (CAFÉ 2.211) to study gene family expansion and contraction during the evolution of orchids and related species (Fig. 1 and Supplementary Note 1.1). By comparing 12 plant species, we found 474 gene families (Extended Data Fig. 3) that appeared unique to orchids (Supplementary Note 1.2). Gene Ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis found these gene families to be specifically enriched in the terms ‘O-methyltransferase activity’, ‘cysteine-type peptidase activity’, ‘lilavone and flavonol biosynthesis’ and ‘stilbenoid, diarylheptanoid and gingerol biosynthesis’ (Supplementary Note 1.2).

Distributions of synonymous substitutions per synonymous site (Ks) for paralogous A. shenzhenica genes showed a clear peak at Ks ~ 1 (Extended Data Fig. 4). Similar peaks at Ks values of 0.7 to 1.1 were identified in 11 other orchids, covering all 5 orchid subfamilies (Supplementary Fig. 2). These peaks might reflect multiple independent whole-genome duplication (WGD) events across orchid sublineages or, more parsimoniously, a single WGD event shared by all (extant) orchids. Comparisons of orchid paralogue Ks distributions with Ks distributions of orthologues between orchid species, and
Figure 1 | Phylogenetic tree showing divergence times and the evolution of gene family sizes. The phylogenetic tree shows the topology and divergence times for 15 plant species. As expected, as a member of the Apostasioideae, *A. shenzhenica* is sister to all other orchids. In general, the estimated orchid divergence times are in good agreement with recent broad scale orchid phylogenies. Divergence times are indicated by light blue bars at the internodes; the range of these bars indicates the 95% confidence interval of the divergence time. Numbers at branches indicate the expansion and contraction of gene families (see Methods and Extended Data Fig. 2). MRCA, most recent common ancestor. The number in parentheses is the number of gene families in the MRCA as estimated by CAFE.

Table 1 | MADS-box genes in the *A. shenzhenica*, *P. equestris*, *D. catenatum*, *P. trichocarpa*, *A. thaliana* and *O. sativa* genomes

| Category | *A. shenzhenica* | *P. equestris* | *D. catenatum* | *P. trichocarpa* | *A. thaliana* | *O. sativa* |
|----------|-----------------|----------------|----------------|-----------------|--------------|-------------|
|          | Functional Pseudo | Functional Pseudo | Functional Pseudo | Functional Pseudo | Functional Pseudo | Functional Pseudo |
| Type II (Total) | 27 | 4 | 29 | 1 | 35 | 11 | 64 | 3 | 47 | 5 | 48 | 1 |
| MIKC³ | 25 | 3 | 28 | 1 | 32 | 9 | 55 | 2 | 43 | 4 | 47 | 1 |
| MIKC³ | 2 | 1 | 1 | 0 | 3 | 2 | 2 | 0 | 2 | 0 | 1 | 0 |
| Mh | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 1 | 4 | 1 | 0 | 0 |
| Type I (Total) | 9 | 0 | 22 | 8 | 28 | 1 | 41 | 9 | 62 | 36 | 32 | 6 |
| Mi | 5 | 0 | 10 | 6 | 15 | 1 | 23 | 4 | 20 | 23 | 15 | 2 |
| Mj | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 5 | 17 | 5 | 9 | 1 |
| Ml | 4 | 0 | 12 | 2 | 13 | 0 | 6 | 0 | 21 | 8 | 8 | 3 |
| Total | 36 | 4 | 51 | 9 | 63 | 12 | 105 | 12 | 107 | 41 | 80 | 7 |

*Genes with stop codon in MADS-box domain were categorized as pseudogenes²⁹.
†Nine MADS-box genes belonging to the M₁ subgroup were identified³⁵.
orchids (Fig. 4b). and E-class members resulted in the derived labellum found in other retention and subsequent sub- and neo-functionalization of B-AP3 the ancestral state. The Apostasia subsequent loss of paralogous genes in non- apostasioid orchids or, alternatively, in the common ancestor of B-AP3 and E classes may have expanded independently only in the Apostasia Fig. 7). These similarities suggest that the lower gene numbers in MADS-box B-AP3 and E classes in Apostasia represent an ancestral state, responsible for producing the plesiomorphic flower with an undeveloped labellum and partially fused gynostemium. The B-AP3 and E classes may have expanded independently only in the non-apostasioid orchids or, alternatively, in the common ancestor of all extant orchids, possibly as a result of the shared orchid WGD, with subsequent loss of paralogous genes in Apostasia causing reversion to the ancestral state. The B-AP3 gene tree topology and some evidence from co-linearity analysis of orchid B-AP3 genes (Supplementary Fig. 10) suggest the latter. We hypothesize that differential paralogue retention and subsequent sub- and neo-functionalization of B-AP3 and E-class members resulted in the derived labellum found in other orchids (Fig. 4b).

The packaging of pollen grains into a compact unit known as the pollinium, specialized for transfer as a unit by pollinating vectors, was a key innovation in the evolutionary history of Orchidaceae and may have played a role in promoting the tremendous radiation of the group. In seed plants, the P- and S-subclades of MIKC*-type genes are major regulators of male gametophytic development. The P-subclade, however, is absent in all orchids except A. shenzhenica (Extended Data Fig. 8). Gene expression analysis showed that, in orchids and M. capitulata, MIKC*-type genes are expressed in the pollinia or pollen, suggesting they play roles in its development (Extended Data Fig. 9). Although most orchids have a pollinium, Apostasia has scattered pollen, similar to M. capitulata, Oryza sativa (rice), and Arabidopsis thaliana. Therefore, we propose that the loss of the P-subclade members of MIKC*-type genes is related to the evolution of the pollinium (Fig. 4a, c and Supplementary Note 3).

Figure 2 | Ks and co-linearity analysis of the A. shenzhenica WGD. a, Distribution of Ks for the one-to-one P. equestris–D. catenatum, A. shenzhenica–D. catenatum, A. shenzhenica–P. equestris and A. shenzhenica–A. officinalis orthologues (filled grey curves and left-hand y-axis). Distribution of Ks for duplicated anchors found in co-linear regions of A. shenzhenica (green lines), D. catenatum (red lines) and P. equestris (blue lines). The filled grey curves and dashed coloured lines are actual data points from the distribution; the solid coloured lines are kernel density estimates (KDE) of the anchor-pair (duplicated genes found in co-linear regions) data scaled to match the corresponding dashed lines. All anchor-pair data are scaled up >15 (right-hand y-axis) compared to the orthologue data. b, Syntenic dot plot of the self-comparison of A. shenzhenica. Only co-linear segments with at least 15 anchor pairs are shown. The sections on each scaffold with co-linear segments are shown in grey. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each position; see Methods). The co-linear regions in green indicate the four co-linear segments that have a common orthologous co-linear segment in A. trichopoda as shown in (c). Co-linear alignment of A. shenzhenica and A. trichopoda. The colours of genes in the alignment indicate gene orientation, with blue for forward strands and green for reverse strands. The grey links connect orthologues between A. shenzhenica and A. trichopoda. Scf86, scaffold00086 of the A. trichopoda genome (v1.0).

Figure 3 | Phylogenomic analysis of orchid WGD events. The numbers on the branches of the species tree indicate the number of gene families with one or more anchor pairs from at least one of the three orchids with genomes that coalesced on the respective branch (top), as well as the individual contributions of anchor pairs from the three orchids (bottom: A. A. shenzhenica; D. D. catenatum; P. P. equestris). The two WGD events identified are depicted by stars. Species with published genomes are in bold. All the duplication events have bootstrap values over 80% (see Methods; for results for bootstrap values over 50% see Supplementary Fig. 15).
orchids23–25. The
we did not find similar genes in epiphytic orchids, suggesting that the
differentiation26. of terrestrial orchids such as
Arabidopsis
and Supplementary Fig. 14). Notably,
in
Arabidopsis
contains one AGL12 clade gene, as do
A. shenzhenica
and rice. In addition, we found transcripts similar to AGL12 in
M. capitulata.
Apostasioideae
(Orchidaceae) from China.
Apostasia shenzhenica:
A new species of Apostasioideae
Orchid historical biogeography, diversification, Antarctica
Proc. R. Soc. B
55 (2016).
Givnish, T. J. Orchid phylogenomics and multiple drivers of their
Published online 13 September 2017.
Received 6 June 2016; accepted 7 August 2017.

Figure 4 | MADS-box genes involved in orchid morphological evolution. a, Phylogenetic analysis of MADS-box genes among
A. shenzhenica, P. equestris, O. sativa and Arabidopsis. The B-AP3 and E-class, MIKC*, M3, and AGL12 and ANR1 subclades are marked by
purple, orange, green and blue shading, respectively. b, A. shenzhenica,
with fewer B-AP3 class and E class MADS-box genes, keeps an
undifferentiated labellum and partially fused gynostemium, while
P. equestris, with more B-AP3 class and E class MADS-box genes, develops
the specialized labellum and column (in red). c, Loss of the P-subclade
genes of MIKC* in P. equestris is likely to be related to the evolution of
pollinia. d, The failed development of endosperm in orchids might be
related to the missing type I M3 MADS-box genes (Extended Data Fig. 9).
e, A. shenzhenica, containing the AGL12 gene and expanded ANR1 genes,
is a terrestrial orchid, while epiphytic orchids, such as P. equestris, have lost
the AGL12 gene and some ANR1 genes.

The reduction of seed volume and content to an absolute minimum is a pivotal aspect of Orchidaceae evolution: in all orchid species,
endosperm is absent from the seed. Type I MADS-box genes are impor-
tant for the initiation of endosperm development22, and transcripts of
type I Mox and Mγ MADS-box genes were found in developing seeds of
A. shenzhenica, P. equestris, and M. capitulata (Extended Data Fig. 10
and Supplementary Fig. 11). Notably, the three orchid genomes do not
contain any type I M3 MADS-box genes (Fig. 4a and Supplementary
Fig. 12), which are found in Arabidopsis, Populus trichocarpa (poplar),
O. sativa (Table 1), and in M. capitulata (Supplementary Fig. 13).
The lack of endosperm in orchids might therefore be related to the missing
type I M3 MADS-box genes (Fig. 4d).
Orchids are one of very few flowering plant lineages that have been
able to successfully colonize epiphytic or lithophytic niches, clinging
to trees or rocks and growing in dry conditions using crassulacean acid
metabolism23–25. The roots of epiphytic orchids, such as
Phalaenopsis
dendrobium, are extremely specialized and differ from the roots
of terrestrial orchids such as
Apostasia.
These aerial roots develop the
velamen radicum, a spongy epidermis that traps the nutrient-rich flush
during rainfall, representing an important adaptation of epiphytic
orches.23–25 The Arabidopsis AGL12 gene is involved in root cell
differentiation26. A. shenzhenica contains one AGL12 clade gene, as do
Arabidopsis and rice. In addition, we found transcripts similar to AGL12
in M. capitulata. In both A. shenzhenica and M. capitulata, these genes
are highly expressed in root tissue (Supplementary Fig. 14). Notably,
we did not find similar genes in epiphytic orchids, suggesting that the
loss of these gene(s) may be involved in losing the ability to develop
true roots for terrestrial growth (Fig. 4e).

Online Content Methods, along with any additional Extended Data display items and
Source Data, are available in the online version of the paper; references unique to
these sections appear only in the online paper.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

1. Roberts, D. L & Dixon, K. W. Orchids. Curr. Biol. 18, R325–R329 (2008).
2. Givnish, T. J. et al. Orchid phylogenomics and multiple drivers of their
extraordinary diversification. Proc. R. Soc. B 282, 1553 (2015).
3. Givnish, T. J. et al. Orchid historical biogeography: diversification, Antarctica
and the paradox of orchid dispersal. J. Biogeogr. 43, 1905–1916 (2016).
4. Chen, L. J. & Liu, Z. J. Apostasia shenzhenica: a new species of Apostasioideae
(Orchidaceae) from China. Plant Science Journal 29, 38–41 (2011).
23. Zotz, G. & Winkler, U. Aerial roots of epiphytic orchids: the velamen radicum

21. Kwantes, M., Liebsch, D. & Verelst, W. How MIKC* or

22. Masiero, S., Colombo, L., Grini, P. E., Schnittger, A. & Kater, M. M. The emerging

7. Kocyan, A. & Endress, P. K. Floral structure and development of Apostasia and

10. Zhang, G.-Q.

6. Dressler, R. L. Phylogeny and Classification of the Orchid Family (Disocordiae, 1993).

5. Kocyan, A., Qiu, Y.-L., Endress, P. K. & Conti, E. A phylogenetic analysis of Apostasiaeae (Orchidaceae) based on ITS, trnL-F and matK sequences. Plant Syst. Evol. 247, 203–213 (2004).

8. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).

9. Cai, J. et al. The genome sequence of the orchid Phalaenopsis equestris. Nat. Genet. 47, 65–72 (2015).

12. 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-Palaeogene boundary. Genome Res. 24, 1334–1347 (2014).

13. Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated syntenic and phylogenetic analyses reveal an ancient genome duplication in monocots. Plant Cell 26, 2792–2802 (2014).

15. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of UV-B damage, and a large dated phylogeny implies multiple gains and losses of this function during the Cenozoic. New Phytol. 205, 1330–1341 (2015).

24. Gravendeel, B., Smithson, A., Silk, J. F. W. & Schutteman, A. Epiphism and pollinator specialization: drivers for orchid diversity? Phil. Trans. R. Soc. Lond. B 359, 1523–1535 (2004).

27. Ibarra-Laclette, E. et al. Architecture and evolution of a minute plant genome. Nature 498, 94–98 (2013).

28. Zhang, H. & Ford, B. G. An Arabidopsis MADS box gene that controls nutrient-induced changes in root architecture. Science 279, 407–409 (1998).

29. 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).

30. Árora, R. et al. MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. BMC Genomics 8, 242 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge support from The Funds for Environmental Project of Shenzhen, China (no. 2012-02); The 948 programme of the State Forestry Administration P.R. China (no. 2011-4-53); The Funds for Forestry Science and Technology Innovation Project of Guangdong, China (no. 2016JKJCX025; no. 2013JCX040-05); Fundamental Research Project of Shenzhen, China (no. JCYJ20170307170746099; no. JCYJ20150430150235943); and the teamwork projects funded by Guangdong Natural Science Foundation (no. 2017A030312004) awarded to Z.-J.L. Y.v.d.P. acknowledges the Multidisciplinary Research Partnership ‘Bioinformatics: from nucleotides to networks’ Project (no. 01MR0310W) of Ghent University and the European Union Seventh Framework Programme (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739-DOUBLEUP.

Author Contributions Z.-J.L. managed the project; Z.-J.L., G.-Q.Z., Y.-Q.Z., K.-W.L., S.-C.N, J.-Y.W., Q.X., S.S., Y.-B.L. (luoyb@ibcas.ac.cn) and C.-M.Y. (cmyeh@mail.saitama-u.ac.jp). Requests for materials should be addressed to Z.-J.L. (liuzj@sinicaorchid.org), G.-Q.Z., Y.-Q.Z., K.-W.L., S.-C.N, J.-Y.W., Q.X., S.S., Y.-B.L. (luoyb@ibcas.ac.cn) and C.-M.Y. (cmyeh@mail.saitama-u.ac.jp). Claims in published maps and institutional affiliations. Correspondence and interests. Readers are welcome to comment on the online version of the paper. www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Z.-J.L. (liuzj@sinicaorchid.org), Y.v.d.P. (yves.vandepeer@psb.vib-ugent.be), W.-C.T. (tsaiwc@mail.ncku.edu.tw), Y.-B.L. (luoyb@ibcas.ac.cn) and C.-M.Y. (cmyeh@mail.saitama-u.ac.jp).

Reviewer Information Nature thanks V. Albert, J. Leebens-Mack, J. C. Pires and S. Ramirez for their contribution to the peer review of this work.
METHODS

No statistical methods were used to predetermine sample size. Sample preparation and sequencing. For genome sequencing, we collected leaves, stems, and flowers from wild A. shenzhenica, a self-pollinating species found in southeast China4 that has a karyotype of 2N = 2X = 68 with uniform small chromosomes (Supplementary Fig. 16). We extracted genomic DNA using a modified cetymethylammonium bromide (CTAB) protocol. Sequencing libraries with insert sizes ranging from 180 bp to 2 kb (Supplementary Table 1) were constructed using a library prep kit (Illumina). These libraries were then sequenced using an Illumina HiSeq 2000 platform. The 80.02-Gb raw reads generated were filtered according to sequencing quality, the presence of adaptor contamination, and duplication. Only high-quality reads were used for genome assembly.

Total RNA was extracted from this study's samples using the RNAprep Pure Plant Kit and genomic DNA contamination was removed using RNase-Free DNase I (both from TaKaRa). The integrity of RNA was evaluated on a 1.0% agarose gel stained with ethidium bromide (EB), and its quality and quantity were assessed using a NanoPhotometer spectrophotometer (IMPLEN) and an Agilent 2100 Bioanalyzer (Agilent Technologies). As the RNA integrity number (RIN) was greater than 7.0 for all samples, they were used in cDNA library construction and Illumina sequencing, which was completed by Beijing Novogene Bioinformatics Technology Co., Ltd. The cDNA library was constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB) and 3 μg RNA per sample, following the manufacturer’s recommendations. The PCR products obtained were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system. Library preparations were sequenced on an Illumina Hiseq 2000 platform, generating 100-bp paired-end reads.

Gene and non-coding RNA prediction. MAKER41 was used to generate a consensus gene set based on de novo predictions from AUGUSTUS42 and GlimmerHMM43, homology annotation with the universal single-copy genes from CEGMA44 and the genes from Arabidopsis (TAIR10) and another four sequenced monocots (O. sativa, P. equestris, S. bicolor and Zea mays) using exonere45, and RNA-seq prediction by Cufflinks46 and TopHat47. These results were integrated into a final set of protein-coding genes for annotation (Supplementary Table 5).

The total length of the final assembly for A. shenzhenica was 349 Mb with a scaffold N50 size of 3.029 Mb and corresponding contig N50 size of 80.1 Kb. (Supplementary Table 4). For the two previously published orchid genomes of P. equestris and D. catenatum, the scaffold N50 size as well as the completeness (see below) improved considerably: for P. equestris, the scaffold N50 size increased from 359.12 Kb to 1.217 Mb and the corresponding contig N50 size from 20.56 Kb to 45.79 Kb, while for D. catenatum the scaffold N50 size increased from 391.46 Kb to 1,055 Mb, and the corresponding contig N50 size from 33.1 Kb up to 51.7 Kb.

Repeat prediction. A total of 146.65 Mb of repetitive elements occupying more than 42.05% of the A. shenzhenica genome were annotated using a combination of structural information and homology prediction48. Retrotransposable elements, known to be the dominant form of repeats in angiosperm genomes, constituted a large part of the A. shenzhenica genome and included the most abundant subtypes, such as LTR/Copia (9.47%), LTR/Gypsy (11.84%), LINE/L1 (2.78%) and LINE/RT-BovB (9.32%), among others. In addition, the percentage of de novo predicted repeats was notably larger than that obtained for homologous repeats based on Repbase49, indicating that A. shenzhenica has multiple unique repeats compared with other sequenced plant species (Supplementary Table 9).

Sample preparation and sequencing. No statistical methods were used to predetermine sample size. Sample preparation and sequencing. For genome sequencing, we collected leaves, stems, and flowers from wild A. shenzhenica, a self-pollinating species found in southeast China4 that has a karyotype of 2N = 2X = 68 with uniform small chromosomes (Supplementary Fig. 16). We extracted genomic DNA using a modified cetymethylammonium bromide (CTAB) protocol. Sequencing libraries with insert sizes ranging from 180 bp to 2 kb (Supplementary Table 1) were constructed using a library prep kit (Illumina). These libraries were then sequenced using an Illumina HiSeq 2000 platform. The 80.02-Gb raw reads generated were filtered according to sequencing quality, the presence of adaptor contamination, and duplication. Only high-quality reads were used for genome assembly.

Total RNA was extracted from this study's samples using the RNAprep Pure Plant Kit and genomic DNA contamination was removed using RNase-Free DNase I (both from TaKaRa). The integrity of RNA was evaluated on a 1.0% agarose gel stained with ethidium bromide (EB), and its quality and quantity were assessed using a NanoPhotometer spectrophotometer (IMPLEN) and an Agilent 2100 Bioanalyzer (Agilent Technologies). As the RNA integrity number (RIN) was greater than 7.0 for all samples, they were used in cDNA library construction and Illumina sequencing, which was completed by Beijing Novogene Bioinformatics Technology Co., Ltd. The cDNA library was constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB) and 3 μg RNA per sample, following the manufacturer’s recommendations. The PCR products obtained were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system. Library preparations were sequenced on an Illumina Hiseq 2000 platform, generating 100-bp paired-end reads.

Repeat prediction. A total of 146.65 Mb of repetitive elements occupying more than 42.05% of the A. shenzhenica genome were annotated using a combination of structural information and homology prediction48. Retrotransposable elements, known to be the dominant form of repeats in angiosperm genomes, constituted a large part of the A. shenzhenica genome and included the most abundant subtypes, such as LTR/Copia (9.47%), LTR/Gypsy (11.84%), LINE/L1 (2.78%) and LINE/RT-BovB (9.32%), among others. In addition, the percentage of de novo predicted repeats was notably larger than that obtained for homologous repeats based on Repbase49, indicating that A. shenzhenica has multiple unique repeats compared with other sequenced plant species (Supplementary Table 9).

Gene and non-coding RNA prediction. MAKER41 was used to generate a consensus gene set based on de novo predictions from AUGUSTUS42 and GlimmerHMM43, homology annotation with the universal single-copy genes from CEGMA44 and the genes from Arabidopsis (TAIR10) and another four sequenced monocots (O. sativa, P. equestris, S. bicolor and Zea mays) using exonere45, and RNA-seq prediction by Cufflinks46 and TopHat47. These results were integrated into a final set of protein-coding genes for annotation (Supplementary Table 5).

Using the same annotation pipeline as for A. shenzhenica, 29.545 and 29.257 protein-coding genes were predicted for P. equestris and D. catenatum, respectively (Supplementary Table 7). A. shenzhenica was found to have a greater average gene length (here we considered the start and stop codons as the two boundaries for a gene) than most other sequenced plants, but this length was similar to that of P. equestris and D. catenatum (Supplementary Fig. 18 and Supplementary Table 10), in both of which this is due to a long average intron length50,51.

We then generated functional assignments of the A. shenzhenica genes with BLAST (version 2.2.28+) by aligning their protein-coding regions to sequences in public protein databases, including KEGG (59.3)46, SwissProt (release 2013_06)46, TrEMBL (release 2013_06)46 and NCBI non-redundant protein database (20150617), and InterProScan (v5.11-51.0)51 was also used to provide functional annotation (Supplementary Table 11). We were able to generate functional assignments for 84.2% of the A. shenzhenica genes from at least one of the public protein databases (Supplementary Table 11).

The RNA genes were searched by tRNAscan-SE52. For RNA identification, we downloaded the Arabidopsis RNA sequences from NCBI and aligned them with the A. shenzhenica genome to identify possible RNAs. Additionally, other types of non-coding RNAs, including miRNA and snRNA, were identified by using INFERNAL16 to search from the Rfam database. In the end, we identified 43 microRNAs, 203 transfer RNAs, 452 ribosomal RNAs and 93 small nuclear RNAs in the A. shenzhenica genome (Supplementary Table 12).

Transcriptome assembly. Before assembly, we got high-quality reads by removing adaptor sequences and filtered low-quality reads by using TRIMMOMATIC53 from raw reads with parameters: ILLUMINACLIP: path:adapter:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36. The resulting high-quality reads were de novo assembled and annotated with the TRINITY program54. The commands and parameters used for running TRINITY were as follows: Trinity --seqType fq --jM200G --leftSample 1_lq --rightSample 2_lq --normalize_by_read --CPU 32 --output sample --min_kmer_cov 2. Protein sequences and coding sequences of transcripts were predicted using TransDecoder (http://transdecoder.sourceforge.net) by aligning their protein-coding regions to sequences in the Pfam database and comparing the translated coding sequences with the PFAM domain database55. For genes with more than one transcript, the longest one was used to calculate transcript abundance and coverage. Transcript abundance level was normalized using the fragments per kilobase per million mapped reads (FPKM) method, and FPKM values were used as proposed by Mortazavi et al.56.

Transcriptomes of Agave deserti and Allium cepa were downloaded from Dryad (h368B) and NCBI (PRJNA175466), respectively. We removed the redundant unigenes in A. cepa by CD-HIT-EST with 99% identity and used TransDecoder to predict proteins with default parameters.

We carried out BLASTP (http://blast.ncbi.nlm.nih.gov/blast) with e-value < 1 × 10−5 to search the best hits for the proteins predicted in the transcriptomes against a customized database, built with proteins from the genomes of A. shenzhenica, P. equestris, D. catenatum56, and A. officinalis (GenBank accession number GCF_001876935.1) as well as public databases, such as NCBI Plant RefSeq (release 80), Ensembl (release 77), Ensembl Metazoa (release 24), Ensembl Fungi (release 24), and Ensembl Protists (release 24),
Only plant-homologous proteins were retained in the transcriptomes to eliminate the effects of genes derived from commensal organisms, laboratory contaminants, and artefacts resulting from incorrect assembly (Supplementary Table 13).

**Gene family identification.** We downloaded genome and annotation data of *A. trichopoda* (http://amborella.hack.psu.edu, version 1.0), *A. comosus* (GenBank accession number GCF_001540865.1), *A. thaliana* (TAIR 10), *A. officinalis* (GenBank accession number GCF_001876935.1), *B. distachyon* (purple false brome; Phytozome v9.0), *M. acuminata* (http://ensemblgenomes.org, release-21), *O. sativa* (Nipponbare, IRGSP-1.0), *P. dactylifera* (http://qatar-weiil.cornell.edu/research/datepalmGenome), *P. trichocarpa* (http://ensemblgenomes.org, release-21), *S. bicolor* (sorghum; Phytozome v9.0), *S. polyrhiza* (common duckweed; http://www.spirodelagenome.org), and *V. vinifera* (Phytozome v9.0). We chose the longest transcript to represent each gene and removed gene models with open reading frames shorter than 150 bp. Gene family clustering was performed using OrthoMCL [5] based on the set of 21,841 predicted genes of *A. shenzhenica* and the protein sets of the above ten other monocots, three dicots and the outgroup *A. trichopoda*. This analysis yielded 11,995 gene families in *A. shenzhenica* containing 18,268 predicted genes (83.6% of the total genes identified; orthologous genes in the 15 sequenced plant species are shown in Supplementary Fig. 19 and Supplementary Table 14) (see also Supplementary Note 1).

**Phylogenetic tree construction and phylogenetic dating.** We constructed a phylogenetic tree based on a concatenated sequence alignment of 439 single-copy gene families and their orthologues in the other plant species described [3]. In brief, the paraphyletic tree based on a concatenated sequence alignment of them using PhyML [5] with GTR + Γ model. A run without data was performed to ensure proper placement of the marginal calibration prior distributions [6]. The MCCM for each orthogroup was run for 10 million generations with sampling every 1,000 generations, resulting in a sample size of 10,000. The resulting trace files of all orthogroups were evaluated manually using *Tracer* v1.5 [7] with a burn-in of 1,000 samples to ensure proper convergence (minimum ESS for all statistics was at least 200). In total, 303 orthogroups were accepted, and all age estimates for the node uniting the WGD paralogous pairs were then grouped into one absolute age distribution (Extended Data Fig. 5; too few anchor pairs were available to evaluate them separately from the peak-based duplicates), for which K estimates were obtained using BayesTraits [8] to estimate the peak consensus WGD age and its 90% confidence interval boundaries, respectively. More detailed methods are available in Vanneste et al. [9].

To compare the relative timing of speciations and WGD event(s) in orchids based on *K* troublesome, we first identified 839 anchors from *D. catenatum* and 355 anchors from *P. equestris* using i-ADHoRe 3.0 and calculated their *K* distribution as described above. Identification of orthologues between *A. shenzhenica* and *A. officinalis*, *A. shenzhenica* and *P. equestris*, *A. shenzhenica* and *D. catenatum*, and *P. equestris* and *D. catenatum* was performed first by reciprocal BLASTP with *E* value < 1 × 10^-10 for proteins from the three orchids and asparagus, followed by sorting BLAST hits by bit-scores and *E* values. Reciprocal best hits from the four comparisons were selected as orthologues. In this way, we identified 9,142 orthologues between *A. shenzhenica* and *A. officinalis*, 10,699 orthologues between *A. shenzhenica* and *P. equestris*, 11,386 orthologues between *A. shenzhenica* and *D. catenatum*, and 13,139 orthologues between *P. equestris* and *D. catenatum*. For each pair of orthologues, ClustalW [10] alignment was carried out to perform sequence alignment using the parameter for amino acids recommended by Hall [11]. PAL2NAL [12] was then used to back-translate aligned protein sequences into codon sequences and to remove any gaps in the alignment. Estimates of *K* values were obtained from CODEML in *PAML* using the Goldman-Yang model with codon frequencies estimated by the F3 × 4 model.

We performed pairwise co-linearity analysis within *A. shenzhenica* and between *A. shenzhenica* and *A. officinalis*, *A. comosus*, *V. vinifera*, and *A. trichopoda*. Homologous pairs of *A. shenzhenica* and the above species were identified by all-against-all BLASTP (*E* value < 1 × 10^-5), followed by the removal of weak matches by applying a c-score of 0.5 (indicating their BLASTP bit-scores were below 50% of the best-scores of the best matches) [13]. Then, i-ADHoRe 3.0 was used to identify co-linear segments with parameters as described above except using `level_2_only` = FALSE, enabling the functionality to detect highly degenerated gene families from more ancient large-scale duplications (this is achieved by recursively building phylogenetic trees based on relatively recent co-linear segments). All co-linear dot plots were drawn by selecting co-linear segments according to a specified required number of anchor pairs (given in the figure legend of each of the dot plots). For the comparisons between *A. shenzhenica* and the chromosome-level assembled genomes (*A. officinalis*, *A. comosus*, and *V. vinifera*) we retained co-linear segments with at least ten anchor pairs (Extended Data Fig. 6 and Supplementary Figs 5, 7, 8). For the comparisons with fragmented genomes, like *A. trichopoda*, and the self-comparison of *A. shenzhenica*, we keep pair with five anchor pairs (Fig. 2b and Supplementary Figs 3, 4). The start
and end boundaries of selected co-linear segments were used to define broader regions containing such segments on the chromosomes or scaffolds by further connecting co-linear segments if they overlapped with each other. Then, duplication depths, that is, the number of connected co-linear segments overlapping at each position of a broader region, were illustrated in the margins of the plots by mapping the connected co-linear segments over each other. The number of anchors required in the co-linear segments could affect the duplication depth in such a way that increasing the number of anchors required tends to reduce the co-linear segments originating from more complex WGDs due to increased gene loss.

To identify the duplication events that resulted in the 1,488 anchor pairs in *A. shenzenica*, the 839 anchor pairs in *D. catenatum*, and the 355 anchor pairs in *P. equestris*, we performed phylogenomic analyses employing protein-coding genes from 20 species, including 12 orchids across all five subfamilies of Orchidaceae (the three orchids with genomes (*A. shenzenica*, *D. catenatum* and *P. equestris*) plus nine orchid transcriptomes (Supplementary Table 13)), four non-orchid Asparagales (/*A. officinalis* (genome), *M. capitulata* (Supplementary Table 13), *A. deserti* and *A. cepa*), three commelinioid monocots (*Elaeis guineensis*, *P. dactylifera*, and *A. comosus*), and *A. trichopoda*. OrthoMCL (v2.0.9) was used with default parameters to identify gene families based on sequence similarities resulting from an all-against-all BLASTP with E value $< 1 \times 10^{-5}$. Then, 1,101 of the 2,582 anchor pairs with $K_i$ values greater than five were removed. If the remaining anchors fell into different gene families, indicating incorrect assignment of *gene* families by OrthoMCL, we merged the corresponding gene families. In this way, we obtained 32,217 multi-gene families. Next, phylogenetic trees were constructed for the subset of 777 gene families with no more than 300 genes that had at least one pair of anchors and one gene from *A. trichopoda*. Multiple sequence alignments were produced by MUSCLE (v3.8.31) using default parameters. These were trimmed by trimAl (v1.4) to remove low-quality regions based on a heuristic approach (-automated1) that depends on a distribution of residue similarities inferred from the alignments for each gene family. RAxML (v8.2.0) was then applied. The timing of the duplication event for each anchor pair relative to the lineage divergence events was then inferred using the following approach (Supplementary Fig. 20): we first mapped internodes from a gene tree to the species tree, the duplication was thus considered to have occurred on the species tree. If the two nodes were directly connected by a single branch on the species tree, or a 'dubious' node. A duplication node is a node that shares at least one parologue with the gene tree. The number of anchors in eukaryotic genomes. *Bioinformatics* 27, 757–763 (2011).

**Evolution and expression analysis of orchid MADS box genes.** We identified candidates of MADS-box genes by searching the InterProScan1 result of all the published protein sequences. The main group of MADS-box genes was further determined by SMART45, which identified MADS-box domains comprised by 60 amino acids. The protein-sequence set of the MADS-box gene candidates was BLAST against the assembled *A. shenzenica* transcriptomes with the TBLASTN program. The matched transcript sequences were then assembled with the candidates of MADS-box genes using Sequencher v5.1 (*Gene Codes Corp.*). The exon structure of the final MADS-box genes was manually edited (Supplementary Data 1).

The end boundaries of selected co-linear segments were used to define broader regions containing such segments on the chromosomes or scaffolds by further connecting co-linear segments if they overlapped with each other. Then, duplication depths, that is, the number of connected co-linear segments overlapping at each position of a broader region, were illustrated in the margins of the plots by mapping the connected co-linear segments over each other. The number of anchors required in the co-linear segments could affect the duplication depth in such a way that increasing the number of anchors required tends to reduce the co-linear segments originating from more complex WGDs due to increased gene loss.

To identify the duplication events that resulted in the 1,488 anchor pairs in *A. shenzenica*, the 839 anchor pairs in *D. catenatum*, and the 355 anchor pairs in *P. equestris*, we performed phylogenomic analyses employing protein-coding genes from 20 species, including 12 orchids across all five subfamilies of Orchidaceae (the three orchids with genomes (*A. shenzenica*, *D. catenatum* and *P. equestris*) plus nine orchid transcriptomes (Supplementary Table 13)), four non-orchid Asparagales (/*A. officinalis* (genome), *M. capitulata* (Supplementary Table 13), *A. deserti* and *A. cepa*), three commelinioid monocots (*Elaeis guineensis*, *P. dactylifera*, and *A. comosus*), and *A. trichopoda*. OrthoMCL (v2.0.9) was used with default parameters to identify gene families based on sequence similarities resulting from an all-against-all BLASTP with E value $< 1 \times 10^{-5}$. Then, 1,101 of the 2,582 anchor pairs with $K_i$ values greater than five were removed. If the remaining anchors fell into different gene families, indicating incorrect assignment of *gene* families by OrthoMCL, we merged the corresponding gene families. In this way, we obtained 32,217 multi-gene families. Next, phylogenetic trees were constructed for the subset of 777 gene families with no more than 300 genes that had at least one pair of anchors and one gene from *A. trichopoda*. Multiple sequence alignments were produced by MUSCLE (v3.8.31) using default parameters. These were trimmed by trimAl (v1.4) to remove low-quality regions based on a heuristic approach (-automated1) that depends on a distribution of residue similarities inferred from the alignments for each gene family. RAxML (v8.2.0) was then used with the GTR+I model to estimate a maximum likelihood tree starting with 200 rapid bootstrap runs followed by maximum likelihood optimizations on every fifth bootstrap tree. Gene trees were rooted based on genes from *A. trichopoda* if these formed a monophyletic group in the tree; otherwise, midpoint rooting was applied. The timing of the duplication event for each anchor pair relative to the lineage divergence events was then inferred using the following approach (Supplementary Fig. 20): we first mapped internodes from a gene tree to the species phylogeny according to the common ancestor of the genes in the gene tree. Each internode of the gene tree was then defined as either a duplication node, a speciation node, or a ‘dubious’ node. A duplication node is a node that shares at least one pair of paralogues, a speciation node is a node that has no paralogues and is consistent with divergence in the species phylogeny, and a ‘dubious’ node is a node that has no paralogues and is inconsistent with divergence in the species phylogeny. Then, if a pair of anchors coalesced to a duplication node, we traced back its parental node(s) until we reached a speciation node in the gene tree. In this way, we circumscribed the duplication event as between these two nodes with the duplication node as the lower bound and the speciation node as the upper bound on the species tree. If the two nodes were directly connected by a single branch on the species tree, the duplication was thus considered to have occurred on the branch. To reduce biased estimations, we used the bootstrap value on the branch leading to the common ancestral node of an anchor pair as support for a duplication event. In total, 628 anchor pairs in 493 gene families coalesced as duplication events on the species phylogeny, and duplication events from 318 anchor pairs in 262 gene families (or from 448 anchor pairs in 367 gene families) had bootstrap values greater than or equal to 80% (or 50%).

**Evolution and expression analysis of orchid MADS box genes.** We identified candidates of MADS-box genes by searching the InterProScan1 result of all the published protein sequences. The main group of MADS-box genes was further determined by SMART45, which identified MADS-box domains comprised by 60 amino acids. The protein-sequence set of the MADS-box gene candidates was BLAST against the assembled *A. shenzenica* transcriptomes with the TBLASTN program. The matched transcript sequences were then assembled with the candidates of MADS-box genes using Sequencher v5.1 (*Gene Codes Corp.*). The exon structure of the final MADS-box genes was manually edited (Supplementary Data 1).

In the end, we aligned all the identified MADS-box genes using the ClustalW program46. An unrooted neighbour-joining phylogenetic tree was constructed in MEGA7 with default parameters. Transcriptomic analysis of other orchids. In addition, 53 more transcriptomes derived from 9 more taxa and 8 tissues (*flower bud, anther, pollenium, shoot, stem, leaf, aerial root and root*) (Supplementary Table 13) were sampled to investigate the roles of the genes that may be important for the evolution of orchid traits. The gene expression levels were indicated by FPKM for the longest assembled transcript.
62. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

63. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).

64. Tuskan, G. A. et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).

65. Jallion, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).

66. Magallón, S., Hiu, K. W. & Quandt, D. Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am. J. Bot.* **100**, 556–573 (2013).

67. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of genome duplications in extremely large data sets. *Nucleic Acids Res.* **40**, e11 (2012).

68. 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).

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

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

71. Proost, S. et al. iADHoRe 3.0—a fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* **40**, W609–W612 (2012).

72. Foster, J. et al. A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* **27**, 749–756 (2011).

73. Ostlund, G. et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196–D203 (2010).

74. D’Hont, A. et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213–217 (2012).

75. Al-Dous, E. K. A new fossil flower from the Turonian of New Jersey and implications regarding the history of bee pollination. *Am. J. Bot.* **85**, 1122–1133 (1998).

76. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock model of angiosperm radiation Malpighiales. *Proc. Natl Acad. Sci. USA* **109**, 17519–17524 (2012).

77. Clarke, J. T., Warnock, R. C. & Donoghue, P. C. Establishing a time-scale for plant evolution. *New Phytol.* **192**, 256–301 (2011).

78. Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Bothwell, R. D. C. Inferences from a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **107**, 18724–18729 (2010).

79. Crepet, W. & Nixon, K. Fossil Clusiaceae from the late Cretaceous (Turonian) of New Jersey and implications regarding the history of bee pollination. *Am. J. Bot.* **85**, 1122–1133 (1998).

80. Putnam, N. H. *Phylogenomics and a posteriori datation resolving the Cretaceous angiosperm radiation Malpighiales*. *Proc. Natl Acad. Sci. USA* **109**, 17519–17524 (2012).

81. Xi, Z. et al. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc. Natl Acad. Sci. USA* **109**, 17519–17524 (2012).

82. 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).

83. Janssen, T. & Bremer, K. The age of major monocot groups inferred from 800 rbcL sequences. *Bot. J. Linn. Soc.* **146**, 385–398 (2004).

84. Oliver, T., Schmidt, B., Nathan, D., Clemens, R. & Maskell, D. Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics* **21**, 3431–3432 (2005).

85. Hall, B. G. *Phylogenetic Trees Made Easy* (Sinauer, 2004).

86. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).

87. Putnam, N. H. et al. The sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86–94 (2007).

88. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for post-analysis of large phylogenies. *Bioinformatics* **26**, 515–521 (2010).

89. Heled, J. & Drummond, A. J. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.* **61**, 138–149 (2012).

90. Oliver, T., Schmidt, B., Nathan, D., Clemens, R. & Maskell, D. Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics* **21**, 3431–3432 (2005).

91. Putnam, N. H. et al. The sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86–94 (2007).

92. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

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

94. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* **43**, D257–D260 (2015).

95. Tamura, K. et al. MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
Extended Data Figure 1 | The morphology of orchid flowers. a, Illustration of an *Apostasia* flower. b, Illustration of a *Phalaenopsis* flower.
Extended Data Figure 2 | Phylogenetic tree showing the topology and divergence times for 15 genomes (A. trichopoda, P. trichocarpa, A. thaliana, V. vinifera, Spirodela polyrhiza, O. sativa, Brachypodium distachyon, Sorghum bicolor, A. comosus, Musa acuminata, Phoenix dactylifera, A. officinalis, A. shenzhenica, P. equestris and D. catenatum) and 10 transcriptomes (Apostasia odorata, Cypripedium margaritaceum, Galeola faberi, Habenaria delavayi, Hemipilia forrestii, Lecanorchis nigricans, M. capitulata, Neuwiedia malipoensis, Paphiopedilum malipoense, Vanilla shenzhenica). The unigenes of the transcriptomes of the 10 'transcriptome' species were aligned to the 439 single-copy gene families of the 15 'genome' species. One hundred and thirty-two single-copy gene families for the 25 species could be identified, and were used to construct a phylogenetic tree based on the PhyML software with the GTR+Γ model, while divergence times (indicated by light blue bars at the internodes) were predicted by MCMCTREE. The range of the bars indicates the 95% confidence interval of the divergence times.
Extended Data Figure 3 | Venn diagram showing unique and shared gene families among members of Orchidaceae, dicots, and Poaceae, and *M. acuminata* and *P. dactylifera*. Numbers represent the number of gene families. Comparison of the 4 groups revealed 474 gene families unique to Orchidaceae and which exist in all 3 Orchidaceae species. If we consider lineage-specific gene families for each group (that is, gene families present in one or a few but not all species in a group), then there are 4,958 unique gene families for Orchidaceae, 7,503 for Poales, 4,494 for the dicots, and 1,560 for the group of *M. acuminata* and *P. dactylifera*. 
Extended Data Figure 4 | A. shenzhenica $K_s$-based age distributions.

a. Distribution of $K_s$ for the whole A. shenzhenica paragene.
b. Distribution of $K_s$ for duplicated anchors found in co-linear regions as identified by i-ADHoRe. A WGD event is identified in both distributions with its peak centred on a $K_s$ value of 1. The dashed lines indicate the $K_s$ boundaries used to extract duplicate pairs for absolute phylogenomic dating of the WGD event (see Methods and Extended Data Fig. 5).
Extended Data Figure 5 | Absolute age of the *A. shenzhenica* WGD event. Absolute age distribution obtained by phylogenomic dating of *A. shenzhenica* paralogues. The solid black line represents the KDE of the dated paralogues, and the vertical dashed black line represents its peak at 74 Ma, which was used as the consensus WGD age estimate. The grey lines represent density estimates from 2,500 bootstrap replicates and the vertical black dotted lines represent the corresponding 90% confidence interval for the WGD age estimate, 72–78 Ma (see Methods). The histogram shows the raw distribution of dated paralogues.
Extended Data Figure 6 | Co-linearity and synteny between *A. shenzhenica* and *A. comosus*. Only co-linear segments with at least 20 anchor pairs are shown. The sections on each scaffold with co-linear segments between *A. shenzhenica* and *A. comosus* are shown in grey. The red bars below the dot plot illustrate the duplication depths (the number of connected co-linear segments overlapping at each scaffold/chromosomal position; see Methods). Only connected co-linear segments with at least ten anchor pairs were used to calculate the duplication depths. The co-linear regions in green highlight the four co-linear segments in *A. shenzhenica* that correspond to a specific set of four co-linear segments in *A. comosus*, which originated from one of the seven ancestral pre-τ-WGD chromosomes in monocots (known as Anc6)14. The phylogenetic tree above the dot plot indicates how Anc6 evolved into (segments of) the current four chromosomes in *A. comosus* (the pair of paired LG18 and LG04, and LG13 and LG23; see Figure 2 in Ming et al.14) through two rounds of WGDs. Names of very small *A. shenzhenica* scaffolds are omitted for clarity. A part of the alignment of the co-linear segments between *A. shenzhenica* and *A. comosus* is shown below. The colours of genes in the alignment indicate anchor pairs with genes of the same colour being homologous. The grey links connect anchor pairs between the two closest segments.
Extended Data Figure 7 | Phylogenetic and expression analysis of orchid B-AP3 genes. Ash, A. shenzhenica; Dca, D. catenatum; Hf, H. forrestii; Mc, M. capitulata; Peq, P. equestris; Pm, P. malipoense; Vs, V. shenzhenica. Expressions of B-class genes derived from H. forrestii are not shown, because only a flower sample was collected from H. forrestii. The expression levels (FPKM value) are represented by the colour bar.
Extended Data Figure 8 | Phylogenetic tree of MIKC*-type genes. The red boxes indicate MADS-box genes from A. shenzhenica. Ash, A. shenzhenica; Dca, D. catenatum; Hi, H. forrestii; Mc, M. capitulata; Peq, P. equestris; Pm, P. malipoense; Vs, V. shenzhenica. MIKC* sequences of the other species were retrieved from GenBank based on Liu et al.20.
Extended Data Figure 9 | Expression patterns of MIKC* MADS-box genes. Ash, *A. shenzhenica*; Dca, *D. catenatum*; Mc, *M. capitulata*; Peq, *P. equestris*. The expression levels (FPKM value) are represented by the colour bar.
Extended Data Figure 10 | Expression of type I M\(^\gamma\) MADS-box genes in *M. capitulata*, *A. shenzhenica* and *P. equestris*. As, *A. shenzhenica*; Mc, *M. capitulata*; Pe, *P. equestris*. The expression levels (FPKM value) are represented by the colour bar.