The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution

Guo-Qiang Zhang1,*, Qing Xu2,*, Chao Bian3,*, Wen-Chieh Tsai4,5,6,*, Chuan-Ming Yeh7,*, Ke-Wei Liu8,*, Kouki Yoshida9,*, Liang-Sheng Zhang10,11, Song-Bin Chang1, Fei Chen11, Yu Shi11,12, Yong-Yu Su11,12, Yong-Qiang Zhang1, Li-Jun Chen1, Yaiyi Yin1, Min Lin1, Huixia Huang1, Hua Deng13, Zhi-Wen Wang14, Shi-Lin Zhu14, Xiang Zhao14, Cao Deng14, Shao-Nei Niu2, Jie Huang1, Meina Wang1, Guo-Hui Liu1, Hai-Jun Yang1,12, Xin-Ju Xiao1, Yu-Yun Hsiao1, Wan-Lin Wu1,5, You-Yi Chen4,5, Nobutaka Mitsuda15, Masaru Ohme-Takagi17,15, Yi-Bo Luo2, Yves Van de Peer16,17,18 & Zhong-Jian Liu1,11,12

Orchids make up about 10% of all seed plant species, have great economical value, and are of specific scientific interest because of their renowned flowers and ecological adaptations. Here, we report the first draft genome sequence of a lithophytic orchid, *Dendrobium catenatum*. We predict 28,910 protein-coding genes, and find evidence of a whole genome duplication shared with *Phalaenopsis*. We observed the expansion of many resistance-related genes, suggesting a powerful immune system responsible for adaptation to a wide range of ecological niches. We also discovered extensive duplication of genes involved in glucomannan synthase activities, likely related to the synthesis of medicinal polysaccharides. Expansion of MADS-box gene clades *ANR1*, *StMADS11*, and *MIKC*, involved in the regulation of development and growth, suggests that these expansions are associated with the astonishing diversity of plant architecture in the genus *Dendrobium*. On the contrary, members of the type I MADS box gene family are missing, which might explain the loss of the endospermous seed. The findings reported here will be important for future studies into polysaccharide synthesis, adaptations to diverse environments and flower architecture of Orchidaceae.

1Shenzhen Key Laboratory for Orchid Conservation and Utilization, The National Orchid Conservation Center of China and The Orchid Conservation and Research Center of Shenzhen, Shenzhen 518114, China. 2State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China. 3Shenzhen Key Lab of Marine Genomics, State Key Laboratory of Agricultural Genomics, Shenzhen 518083, China. 4Department of Life Sciences, National Cheng Kung University, Tainan 701, Taiwan. 5Orchid Research Center, National Cheng Kung University, Tainan 701, Taiwan. 6Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, Taiwan. 7Graduate School of Science and Engineering, Saitama University, Saitama 338-8570, Japan. 8The Center for Biotechnology and Biomedicine, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China. 9Technology Center, Taisei Corporation, Kanagawa 245-0051, Japan. 10Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou 350002, China. 11Fruit Crop Systems Biology Laboratory, College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China. 12College of Forestry, South China Agricultural University, Guangzhou, 510640, China. 13Chinese Academy of Forestry, Beijing, 100093, China. 14BioProduction Research Institute, National Institute of Advanced Industrial Science and Technology, Ibaraki 305-8562, Japan. 15Department of Plant Systems Biology, VIB, and Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. 16Bioinformatics Institute Ghent, Ghent University, Ghent B-9000, Belgium. 17Department of Genetics, Genomics Research Institute, Pretoria, South Africa. *These authors jointly supervised this work. †These authors contributed equally to this work. ‡These authors should be addressed to Y.-B.L. (email: luoyb@ibcas.ac.cn) or Y.V.d.P. (email: yves.vandepeer@psb.vib-ugent.be) or Z-J.L. (email: liuzj@sinicaorchid.org)
Orchids, constituting approximately 10% of all seed plant species, have enormous value for commercial horticulture, and are of specific scientific interest because of their spectacular flowers, ecological adaptations and secondary metabolites. *Dendrobium* is the third largest genus of Orchidaceae and contains approximately 1,450 species, characterised by a fleshy stem with abundant polysaccharides and growing in diverse habitats. A draft genome sequence of *Dendrobium officinale* Kimura & Migo has been reported before but the highly fragmented assembly and the presence of multiple peaks in K-mer analyses, suggesting that its sequence is likely derived from an artificial hybrid, seriously complicate correct interpretation of the genome. To complement the lack of a high quality, well assembled genome sequence for *Dendrobium*, we here present the genome of *D. catenatum* Lindl., a lithophytic orchid found in subtropical and temperate regions and commonly used as a health food in many Asian countries. Analysis of the *D. catenatum* genome sequence offers insights into flower development and polysaccharide synthesis, as well as its wide distribution.

**Results**

**Genome sequencing and genome characteristics.** *Dendrobium catenatum* (Supplementary Note 1) has thirty-eight (2N = 2X = 38) small chromosomes of approximately 2 μm (Supplementary Note 2 and Supplementary Fig. 1). To sequence its complete genome, we generated a total of 222.51 Gb of raw reads, with multiple insert libraries ranging in size from 180 bp to 20 Kb (Supplementary Table 1). A K-mer analysis estimated the genome size of *D. catenatum* at 1.11 Gb (Supplementary Fig. 2). Assembly was done with SOAPdenovo2 and Platanus, but completeness and N50 length of scaffolds were much better with the latter tool (Supplementary Fig. 3), the results of which were used in subsequent analyses. The total length of its assembly was 1.01 Gb (Supplementary Table 2). Mapping all of the paired-end reads to the assembly revealed that 97% of the sequence had a coverage depth greater than five (Supplementary Fig. 4). Further quality analysis indicated that 93% of the set of euchromatic core genes (CEGMA) were present and 97% were partially represented, suggesting near completeness of the euchromatin component. In addition, 93%–95% of the RNA seq data set could be mapped onto the assembled sequence (Supplementary Tables 3 and 4). These results suggest that our genome assembly is of high quality.

A total of 789 Mb of repetitive elements occupying more than 78.1% of the *D. catenatum* genome were annotated using a method combining structural and homology information. Retrotransposable elements, known to be the dominant form of repeats in angiosperm genomes, constituted a large part of the *D. catenatum* genome and included the most abundant subtypes, such as LTR/Copia (27.36%), LTR/Gypsy (18.49%), LINE/L1 (8.44%) and LINE/RTE (5.68%), among others. In addition, the percentage of de novo predicted repeats was notably larger than that obtained for repeats based on Repbase, indicating that *D. catenatum* has many unique repeats compared to other sequenced plant genomes (Supplementary Note 3 and Supplementary Table 5). Among these elements, long terminal repeats (LTRs) were the most dominant type, accounting for approximately 46% of the genome. After calculating their times of insertion, we discovered that a burst of LTR activity occurred during the last five million years (Supplementary Fig. 5) and therefore, we deduced that these LTRs were inserted into the genome after *D. catenatum* diverged from *Phalaenopsis* species (which is estimated to have occurred 22.6–59.6 million years ago, Fig. 1). We annotated 28,910 protein-coding genes (Supplementary Note 4), of which 22,394 had a coverage depth greater than five (Supplementary Fig. 4). Further quality analysis indicated that 93% of the RNA seq data set could be mapped onto the assembled sequence (Supplementary Tables 3 and 4). These results suggest that our genome assembly is of high quality.

We identified 5,758,781 heterozygous single nucleotide polymorphisms (SNPs) in the *D. catenatum* genome. The heterozygous SNP rate for the whole genome was estimated at 6.28 × 10⁻³, whereas the SNP rate in exons was as low as 4.98 × 10⁻³ (Supplementary Fig. 9). Of the 139,830 SNPs that were found in exons, 69,459 caused non-synonymous mutations, affecting 18,404 genes, and this suggested that *D. catenatum* is a high heterozygosity genome. We conducted a Gene Ontology (GO) and KEGG enrichment analyses of the affected genes and found enrichment of the KEGG pathways: ‘Biosynthesis of secondary metabolites’, ‘Plant hormone signal transduction’, ‘Metabolic pathways’ and ‘Isoprenoid biosynthesis’ (Supplementary Table 10), and the GO terms ‘ATP binding’, ‘protein tyrosine kinase activity’ and ‘transition metal ion binding’ (Supplementary Table 11).

**Genome evolution.** We constructed a highly supported phylogenetic tree and estimated the divergence times of 12 plants based on genes extracted from a total of 677 single-copy families (Fig. 1a). As expected, we found that the *D. catenatum* is most closely related to *P. equestris* from which it separated approximately 38 million years ago. Both the distribution of synonymous substitutions per synonymous site (Ks) across all paralogous genes (regardless of gene order, Fig. 1b) and for duplicated genes lying in synteny blocks (Fig. 1c) show two obvious
peaks at Ks values between 0.7–0.9 and 1.5–1.8, suggestive of two rounds of whole-genome duplication (WGDs) in the *D. catenatum* lineage (Supplementary Note 6). Dating of the WGDs suggests that the most recent WGD appeared near to the Cretaceous–Paleogene (K/Pg) boundary and is shared with the WGD event documented recently for *P. equestris* 12. Since it has been suggested that WGDs might facilitate species diversification 17,18, it would be interesting to see whether the WGD has also been shared with the species-rich subfamily Orchidoideae (3630 species), which diverged from the Epidendroideae (about 20,000 species, amongst which *D. catenatum* and *P. equestris*) about 59 million years ago 19, and with the subfamilies Apostasioideae, Vanilloideae and Cypripedioideae, which only include 17, 185 and 180 species, respectively 20. Cypripedioideae and the ancestor of Orchidoideae and Epidendroideae subfamilies are assumed to have diverged from each other about 68 million years ago19. Unfortunately, whole genome sequences, or extensive transcriptome data sets from members of these other subfamilies are not yet available. The older peak in the Ks age distribution probably points to one or more older WGD events.

**Gene family evolution.** We have also zoomed in on some specific gene families.

*Terpene synthase genes.* As secondary metabolites, most plant terpenes and their corresponding synthases have evolved selectively to increase fitness by adaptation to specific ecological niches22. Plant terpene synthase (TPS) genes can be divided into seven subfamilies (a, b, c, d, e/f, g and h)22. The TPS genes of *D. catenatum* and

---

**Figure 1.** Phyllogenetic position and Ks distributions for *D. catenatum*. (a) Phylogenetic tree showing the topology and divergence times for 12 plant species, including *D. catenatum*. Estimated divergence times are indicated by light blue boxes at internodes. Numbers at nodes indicate bootstrap values. The brown bar indicates the orchid-specific whole-genome duplication (WGD), while the green bar indicates a more ancient monocot-specific WGD (Thanks Li-Jun Chen for taking the images of species). (b) Distribution of synonymous substitutions per synonymous site (Ks) for the whole *D. catenatum* paranome. (c) Distribution of synonymous substitutions per synonymous site (Ks) for orthologous genes found in syntenic regions. Two consistent peaks highlighted by the dashed rectangles are considered to reflect the most recent and older WGD events.
**Figure 2.** Phylogeny of putative full-length TPSs from *D. catenatum* (green), *P. equestris* (purple), *O. sativa* (blue) and *A. thaliana* (cyan). See text for details.

*P. equestris* all fall into known angiosperm TPS clades, TPS-a, TPS-b, TPS-c/f, TPS-c, and TPS-g (Fig. 2). The genome of *D. catenatum* encodes 39 members of TPS, whereas there are only 21 in the genome of *P. equestris*. Notably, rapid expansion by tandem gene duplication is particularly common in the TPS-a subfamily of these two orchids. Furthermore, the specific placement of TPS-a genes for *D. catenatum* and *P. equestris* (Fig. 2; Supplementary Tables 12 and 13) implies that the expansion of this gene family has occurred in the ancestor of the Epidendroideae subfamily, or at least prior to the divergence of *D. catenatum* and *P. equestris*, and might have contributed to species radiation in this subfamily, containing over 20,000 species. Indeed, although this needs to be further investigated, a previous study has suggested that the expansion of the TPS-a subfamily might be linked to the radiation of the flowering plants.

**Disease resistance genes.** Plant disease resistance genes (*R* genes) play a key role in recognizing proteins expressed by specific avirulence genes of pathogens, and form various subfamilies, such as the TIR-domain-containing (for example, TOLL/INTERLEUKIN LIKE RECEPTOR/RESISTANCE PROTEIN) (TIR-NB-LRR), the non-TIR-domain containing (NB-LRR), and the non-TIR coiled-coil domain-containing (CC-NB-LRR) R-protein subfamilies. The genomes of *D. catenatum* and *P. equestris* possess 157 and 79 *R* genes, respectively (Supplementary Table 14). Although further investigation is required, the dramatic expansion of its *R* genes suggests that *D. catenatum* may possess a more powerful disease immune system than *P. equestris*.

**Heat-shock proteins.** As molecular chaperones, heat-shock proteins (Hsp) are ubiquitous in plant cells. Hsp genes are not only associated with stress caused by heat shock and other abiotic factors, but have recently also been found to be associated with response to biotic stress. Hsp genes function to manage the stress-induced denaturation of other proteins and can be classified into seven major families based on their molecular weight: small Hsps, Hsp20, Hsp40, Hsp60, Hsp70, Hsp90 and Hsp110. Of those, plants mainly contain Hsp20, Hsp70 and Hsp90 subfamilies. The genome of *D. catenatum* contains 20 members of Hsp70, whereas there are only 9 in that of *P. equestris*. Interestingly, in particular Hsp70 genes encoding proteins localizing in the cytoplasm have more members in *D. catenatum* than in *P. equestris* (11 vs. 3) (Supplementary Fig. 10). Due to the fact that *D. catenatum* is found in subtropical and temperate regions (Supplementary Fig. 11), can grow in wet and dry environments, can tolerate both low and high temperatures, and has a much wider distribution than *P. equestris* (Supplementary Fig. 11), it is interesting to speculate that the additional Hsp70 genes in the *D. catenatum* genome might have helped in the adaptation to a much wider variety of environments. However, more future work will be necessary to prove or disprove this hypothesis.

**Evolution of polysaccharide synthase gene families.** The fleshy stem of *D. catenatum* contains various types of polysaccharides, many of which have medicinal, such as anti-inflammatory, immuno-enhancing, antioxidant and anti-glycation activities. Among those, particularly glucomannan (GM) and galactoglucomannan (GGM) are two major medicinal polysaccharides in *D. catenatum*. Genes involved in GM and GGM biosynthesis were identified through their homology with genes in the Arabidopsis genome. A biosynthetic pathway was proposed and the tissue-specific expression patterns of GM and GGM biosynthesis genes were examined (Fig. 3 and Supplementary Note 7). The result suggests that the downstream genes of the biosynthesis pathway are highly expressed in stem tissues where high levels of GM and GGM accumulate. Therefore, we focussed on the analysis of these genes in the *D. catenatum* genome.

Since GM or GGM polysaccharides of *D. catenatum* are easily extracted with water, they may not be tightly bound to the cell wall and probably act as storage polysaccharides in specialized mucilage cells rather than being structural polysaccharides. Konjac glucomannan (KGM) is water-soluble and accumulates in storage tissues. It also has several bioactivities, such as reducing plasma cholesterol, removing free radicals and inhibiting tumor genesis and metastasis. In addition, the backbone structure of KGM is similar to that of GM. Therefore, we included konjac EST sequences responsible for GM synthesis to search for *D. catenatum* orthologs.
Previous studies showed that CslA (Cellulose synthase-like A) genes of glycosyltransferase (GT) family 2 are involved in GM backbone synthesis. We found 13 CslA genes in the D. catenatum genome, compared to only 6 copies in the P. equestris genome. This expansion of CslA genes in the D. catenatum genome is mainly due to tandem duplication (three arrays: Dca006365, Dca006366; Dca007032, Dca007033, Dca007034; Dca013434, Dca013437). Interestingly, these genes were grouped in the same clade with AkCslA3, a konjac GM synthase (Supplementary Fig. 12). In addition, two of these genes (Dca006366 and Dca007032) were significantly higher expressed in stem than in other tissues (root, leaf and flower, Supplementary Fig. 13). Therefore, these expanded CslA genes may act as GM or GGM synthases in D. catenatum. Although CslD genes were reported to synthesize mannan rather than GM in A. thaliana, a recent study showed that konjac CslD may also be involved in the synthesis of GM. Based on our phylogenetic analysis, two D. catenatum genes (Dca006366 and Dca007032) were significantly higher expressed in stem than in other tissues (root, leaf and flower, Supplementary Fig. 13). Therefore, these expanded CslA genes may act as GM or GGM synthases in D. catenatum. Although CslD genes were reported to synthesize mannan rather than GM in A. thaliana, a recent study showed that konjac CslD may also be involved in the synthesis of GM. Based on our phylogenetic analysis, two D. catenatum genes (Dca006366 and Dca007032) cluster with the konjac CslD EST clones (Supplementary Fig. 14). These two genes were highly expressed in stem and leaf, respectively (Supplementary Fig. 15) and suggest their potential roles in the synthesis of GM or GGM.

Arabidopsis CslD5 was reported to play an important role in osmotic stress tolerance. In addition, GM present in the pseudobulb of an epiphytic CAM orchid, Cattleya forbesii Lindl. × Laelia tenebrosa Rolfe, has been associated with drought tolerance. The large accumulation of starch, fructan and GM in storage organs of geophytes is critical for their survival in detrimental conditions. Because Dendrobium species accumulate high amounts of GM and GGM in their stems and/or leaves, it would be interesting to know whether this is also related to adaptation to environmental stresses, such as drought, a common condition experienced by epiphytic or lithophytic Dendrobium species in their natural environment. The online microarray data from Arabidopsis eFP Browser seems to support this because A. thaliana CslA (CslA7 and CslA10) and CslD (CslD2 and CslD3) genes are induced by drought, osmotic, salt or cold stress.

Figure 3. Proposed biosynthetic pathway of GM and GGM in Dendrobium stem. The biosynthetic pathway was modified according to the pathways proposed in Amorphophallus konjac. GM or GGM biosynthesis is supposed to be generated from sucrose, mainly produced by photosynthesis in the leaf tissue (Supplementary Note 7). The enzymes indicated in red are highly expressed in the stem. Only abbreviations of gene names are shown: Csl, Cellulose synthase like gene; FRK, fructokinase; Fru, fructose; Fru-6-P, Fructose-6-phosphate; GGP, GDP-glucose-pyrophosphorylase; GMPP, GDP-mannose pyrophosphorylase; GT, glycosyltransferase; HXK, hexokinase; INV, invertase; MSR, mannan synthesis-related; PGI, phosphoglucose isomerase; PGM, phosphoglucomutase; PMI, phosphomannose isomerase; PMM, phosphomannomutase; RW A, Reduced Wall Acetylation protein; SuS, sucrose synthase; UGE, UDP-galactose epimerase; UGP, UDP-glucose pyrophosphorylase.
Storage of carbohydrates in geophytes can serve as carbon and energy sources for the maintenance under adverse environments and for growth under favorable conditions. In addition, soluble sugars, such as glucose and sucrose, can act as osmolytes under osmotic stress. Accumulation of these metabolites is enhanced in response to environmental stresses and has been shown to contribute to drought and freezing tolerance. In Easter lily bulbs, when stored at -1.0 °C, large amounts of sucrose, mannose, fructose and oligosaccharides accumulated, suggesting that not only starch but also GM was degraded to soluble sugars during frozen storage.

| Species                  | Total | Type II | Type I |
|-------------------------|-------|---------|--------|
| Total                   |       | Total MIKCc | Total Mα |
| *Solanum tuberosum*     | 167   | 65      | 102    |
| *Populus trichocarpa*   | 105   | 64      | 41     |
| *Arabidopsis thaliana*  | 108   | 46      | 62     |
| *Brassica rapa*         | 160   | 95      | 65     |
| *Brachypodium distachyon* | 57   | 39      | 18     |
| *Oryza sativa*          | 75    | 43      | 32     |
| *Phalaenopsis equestris*| 51    | 29      | 22     |
| *Dendrobium catenatum*  | 63    | 35      | 28     |

Table 1. The number of MADS-box genes in some representative plant species.

**Evolution of MADS-box genes.** Given that orchids are a unique model system for flower development, we characterised their MADS-box genes, which hold diverse functions in many important processes during plant development, in greater detail. An investigation of the *D. catenatum* genome revealed 63 putative functional MADS-box genes and 12 pseudogenes (Table 1). As earlier reported for *P. equestris*, there seem to be fewer MADS box genes present in orchids than in most other angiosperms, such as rice (*Oryza sativa*; 75 genes) and *A. thaliana* (108 genes). *D. catenatum* has 35 type II MADS-box genes (Table 1), compared with 29 in *P. equestris*. Phylogenetic analysis (Supplementary Fig. 22) shows that most type II MADS-box genes have been duplicated in *D. catenatum*, except for those in the B-PI clade. Among these clades, *ANR1* (with three members), *StMADS11* (three members), MIKCc (three members), and Bs (two members) contain more members than in *P. equestris* (two members in *ANR1* and one member in other three clades, respectively) (Supplementary Fig. 23). The *ANR1* MADS-box gene in Arabidopsis is a key gene involved in regulating lateral development in response to external nitrate supply. Genes in the *StMADS11* clade have functions in controlling flowering time and inflorescence architecture.

In *D. catenatum*, the number of MADS-box genes is 63, which is less than in most other angiosperms. This lower number of MADS-box genes in *D. catenatum* may be due to the rapid evolution of floral development in orchids, which is reflected in the low number of MADS-box genes compared to other angiosperms. The lower number of MADS-box genes in *D. catenatum* may also be due to the rapid evolution of floral development in orchids, which is reflected in the low number of MADS-box genes compared to other angiosperms. This lower number of MADS-box genes in *D. catenatum* may also be due to the rapid evolution of floral development in orchids, which is reflected in the low number of MADS-box genes compared to other angiosperms.
were also found in the same scaffold7526. Interestingly, the D. catenatum genome does not contain any type I M(3) MADS-box genes, although these genes do exist in Arabidopsis, poplar and rice. Interactions among type I MADS-box genes are important for the initiation of endosperm development54. The failed development of endosperm in orchids might be related to the smaller number of type I MADS box genes in the D. catenatum genome.

In conclusion, Dendrobium represents a fascinating groups of orchids because of their fleshy stem, various flower architectures, and synthesis of many kinds of different polysaccharides and the D. catenatum genome sequence forms an important resource for further exploring orchid gene and genome evolution.

Method
Sample preparation and sequencing. For genome sequencing, we collected leaves, stems and flowers from an individual of wild D. catenatum (voucher specimen : CHINA.Yunnan: Guangnan county, on rock in evergreen broad-leaf forest, alt. 1350 m, 10 March, 2010, Z.J. Liu 4979, NOCC) and extracted genomic DNA using a modified CTAB protocol. Sequencing libraries with insert sizes ranging from 180 bp to 20 Kb (Supplementary Table 1) were constructed using a library construction kit (Illumina, San Diego, CA). These libraries were then sequenced using Illumina HiSeq 2000 platform. The raw reads generated were filtered according to the sequencing quality, presence of adaptor contamination and duplication. Thus, only high-quality reads were used for genome assembly.

Genome size estimation. To estimate the genome size of D. catenatum, we used reads from pair-end libraries to determine the distribution of K-mer values. According to the Lander–Waterman theory55, the genome size can be determined by the total number of K-mers that were divided by the peak value of the K-mer distribution. Given the high heterozygosity in the D. catenatum genome, we found two peaks in the distribution (Supplementary Fig. 2). Using the second peak as the expected K-mer depth, and the formula Genome size = Total K-mer/Expected K-mer depth, the size of the haploid genome was estimated to be 1.11 Gb (haploid).

Sequence assembly. Initially, we used SOAPdenovo28 to assemble the genome, which produced an assembly of 1.27 Gb with an N50 scaffold size of 80.56 Kb and a corresponding N50 contig size of 6.64 Kb (Supplementary Table 18). These figures suggest high fragmentation and redundancy. Therefore, to generate a better assembly for further analyses, Platanus9, which can effectively manage high-throughput data from heterogeneous datasets, was used for whole genome shotgun assembly. We subsequently used GapCloser (http://soap.genomics.cn) to fill gaps remaining after the Platanus built-in gap-filling module had been applied. The final assembly was 1.01 Gb in length, approximately 91% of the estimated genome size, with an N50 scaffold size of 391 Kb and a corresponding N50 contig size of 33.1 Kb (Supplementary Table 2).

Gene and non-coding RNA gene prediction. MAKER26 was used to generate a consensus gene set based on de novo prediction, homology annotation with CEGMA15 and other sequenced monocots, and RNA-seq gene prediction. These results were integrated into a final set of 28,910 protein-coding genes for genome annotation (Supplementary Table 19). We then generated functional assignments of the D. catenatum genome by aligning their CDS (protein-coding sequences) to sequences available in the public protein databases including KEGG15, SwissProt57, TrEMBL57 and InterProScan58 (Supplementary Table 20). tRNA genes were searched in the genome to identify possible rRNAs. Additionally, other types of non-coding RNAs, including miRNA and snRNA, were identified by utilizing INFERNAL60 to search the Rfam database.

Single nucleotide polymorphisms. We used the BWA program14 to remap the pair-end (500 bp) clean reads to the assembled scaffolds. After merging the BAM results, sorting the alignments by the leftmost coordinates and removing potential PCR duplicates, we used SAMtools15 ‘mpileup’ to identify single nucleotide polymorphisms (SNPs) and short INDELs. We rejected SNPs and InDels within reads with depths lower (<5 folds) or higher (>80 folds) than expected. Filtering was achieved using the vcfutils.pl varFilter tool in the SAMtools package, with parameters -Q 10 -d 5 -D 86. We estimated heterozygosity rates as the density of heterozygous SNPs from the whole genome, gene intervals, introns and exons.

Gene family identification. We downloaded genome and annotation data from Amborella (Am.) trichopoda (http://amborella.huck.psu.edu, version 1.0), Arabidopsis (A.) thaliana (TAIR 10), Brachypodium distachyon (purple false brome; Phytozone v9.0), Musaceae acuminata (http://ensemblgenomes.org, release-21), Oryza sativa (Nipponbare, IRGP-1.0), Phoenix (Ph. dactylifera) (http://qatar-weill.cornell.edu/research/datepalmGenome), Phyllostachys edulis (http://ftp.genomics.org.cn/from_BGisz/20130120/), Populus (Po.) trichocarpa (http://ensemblgenomes.org, release-21), Sorghum (S.) bicolor (sorghum; Phytozone v9.0), Spirodela (Sp.) polyrhiza (common duckweed; http://www.spirodelagenome.org) and Vitis vinifera (Phytozone v9.0), Zea mays (http://www.plantgdb.org/ZmGDB), Phyllostachys (Phy.) heterocyclica (http://www.bamboogdb.org). We chose the longest transcript to represent each gene, and removed gene models with an open reading frame (ORF) shorter than 150 bp. These protein sets were aligned and clustered using OrthoMCL61.

Phylogenomic dating. We conducted phylogenomic dating with PAML McMCTree62. The McM process was run for 1,500,000 iterations, with a sample frequency of 150 after a burn-in of 500,000 iterations. Other parameters used the default settings of McMCTree. Two independent runs were performed to check convergence. The following constraints were used for time calibrations:
1. 140–150 million years ago (MYA) for the monocot – dicot split\textsuperscript{63},
2. 94 MYA as the lower boundary for the Vitis – Eurosid split\textsuperscript{22},
3. 130 MYA as the lower boundary for the Alismatales – Acorales and core monocots (Commelinids, Asparagales, Liliales, etc.) split\textsuperscript{64}, and
4. 200 MYA as the upper boundary for basal angiosperms\textsuperscript{65}.

Based on these divergence time ranges and the inferred phylogenetic tree, the divergence times between the 12 species were estimated using McMCtree software.

**Identification of resistance genes.** HMMPER V3.0 was used to align the protein sequences of *D. catenatum* against the hidden Markov model of the Pfam NBS (NB-ARC). The TIR and LRR domains were detected by using the Pfam_Scan (E 0.01 – domE 0.01). MARCOIL\textsuperscript{66} and paircoil2\textsuperscript{67} were utilized for identification of the CC motif.

**Identification of polysaccharide-related genes.** We collected polysaccharide-related genes of Arabidopsis first by using the CAZY database and other information resources. Then, we performed TBLASTN search against all coding sequences (CDS) datasets of each plant species. These CDS datasets were downloaded from Phytozone (poplar, Selaginella and Physcomitrella), ConGenIE (Norway spruce), QATAR-WEILL. CORNELL (dates palm), RAP-DB (rice) and TAIR (Arabidopsis). In case of populus, Selaginella and Physcomitrella, ConGenIE (Norway spruce), QATAR-WEILL. CORNELL (dates palm), RAP-DB (rice) and TAIR (Arabidopsis). In case of

**Evolution of MADS box genes in *D. catenatum*.** The MADS-box domain is comprised of 60 amino acids, which we identified for all the potential MADS-box sequences of *D. catenatum*. Next, we aligned all the MADS-box genes with ClustalW. An un-rooted neighbour-joining phylogenetic tree was constructed in MEGA5\textsuperscript{69} with default parameters. Confidence on the tree branches was evaluated by bootstrap analysis (1000 replicates).

Associated references and supplementary information are available in the online version of the paper.

**References**

1. Robertson, D. L. D. & Orchids, K. W. *Curr Biol* 18, 325–329 (2008).
2. Pridegson A. M., Cribb, J. P., Chase W. M. & Rasmussen F. N. In *Genera Orchidacearum*, Vol. 6, Epidendroideae (Part Three), 3–544 (Oxford University Press, Oxford, UK, 2013).
3. Leitch, I. J. et al. Genome size diversity in orchids: consequences and evolution. *Ann Bot* 104, 469–481 (2009).
4. Ng, T. B. et al. Review of research on Dendrobium, a prized folk medicine. *Appl Microbiol Biotechnol* 93, 1795–1803 (2012).
5. Pan, L. H. et al. Comparison of hypoglycemic and antioxidative effects of polysaccharides from four different Dendrobium species. *Int J Macromol* 64, 420–427 (2014).
6. Dressler, R. L. *In Phylogeny and Classification of the Orchid Family*, 7–278 (Cambridge University Press, Australia, 1993).
7. Yan, L. et al. The genome of Dendrobium officinale illuminates biology of the important traditional Chinese orchid herb. *Mol Plant* doi:10.1016/j.molp.2014.12.011 (2014).
8. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1, 18 (2012).
9. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* 24, 1384–1395 (2014).
10. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067 (2007).
11. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009).
12. Cai, J. et al. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat Genet* 47, 65–72 (2015).
13. 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).
14. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25–29 (2000).
15. Ogata, H. et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27, 29–34 (1999).
16. Kawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci USA* 106, 5737–5742 (2009).
17. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10, 725–732 (2009).
18. Oderda, G. M. et al. Creating a path to the summit by thinking off the map: report of the 2008-2009 Academic Affairs Committee. *Am J Pharm Educ* 73 Suppl., S7 (2009).
19. Gustafsson, A. L., Verola, C. F. & Antonelli, A. Reassessing the temporal evolution of orchids with new fossils and a Bayesian relaxed clock, with implications for the diversification of the rare South American genus *Hoffmannseggia* (Orchidaceae: Epidendroideae). *BMC Evol Biol* 10, 177 (2010).
20. The Angiosperm Phylogeny Group. G. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* 161, 105–121 (2009).
21. Salse, J. *In silico* archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr Opin Plant Biol* 15, 122–130 (2012).
22. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J* 66, 212–229 (2011).
23. Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* 342, 1241089 (2013).
24. McHale, L., Tan, X., Koehl, P. & Michelmore, R. W. Plant NBS-LRR proteins: adaptable guards. *Genome Biol* 7, 212 (2006).
25. Wang, W., Vinocur, B., Shoseyov, O. & Altman, A. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci* 9(5), 244–252 (2004).
68. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics doi: 10.1002/0471250953.bi0203s00 (2002).
69. 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).
70. Potato Genome Sequencing, C. et al. Genome sequence and analysis of the tuber crop potato. Nature 475, 189–195 (2011).
71. Leseberg, C. H., Li, A., Kang, H., Duvall, M. & Mao, L. Genome-wide analysis of the MADS-box gene family in Populus trichocarpa. Genes 378, 84–94 (2006).
72. Duan, W. et al. Genome-wide analysis of the MADS-box gene family in Brassica rapa (Chinese cabbage). Mol Genet Genom 290, 239–255 (2015).
73. Wei, B. et al. Genome-wide analysis of the MADS-box gene family in Brachypodium distachyon. PLoS ONE 9, e84781 (2014).
74. Arora, R. et al. MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. BMC Genomics 8, 324 (2007).

Acknowledgements
The authors acknowledge support from the 948 programme from the State Forestry Administration P. R. China (no. 2011–4–53), the Funds for Forestry Science and Technology Innovation Project of Guangdong, China (no. 2011KJCX009; no. 2013KJCX014–05), the Funds for Environmental Project of Shenzhen, China (no. 2013–02), the Funds for Technology Research and Development Project of Shenzhen, China (no. CXZZ20120830103025851), the Funds for the Development of Strategic Emerging Industries of Shenzhen, China (no. NY20130205008) to Z.-J. L. The authors appreciate technical help of Dr. Saqib Muhummad (AIST) for assembling *Amorphophallus* (*Amo.* konjac) RNA-seq data. Y. V. d. P. acknowledges the Ghent University Multidisciplinary Research Partnership ‘Bioinformatics: From Nucleotides to Networks’ and support from the European Union Seventh Framework Programme (FP7/2007–2013) under European Research Council Advanced Grant Agreement 322739–DOUBLE-UP.

Author Contributions
Z.-J.L. and G.-Q.Z. managed the project. Z.-J.L., G.-Q.Z., C.B., Y.V.d.P., Y.-B.L., Q.X., K.-W.L., L.-S.Z., F.C., W.-C.T., Z.-W.W. and Y.-Y.H. planned and coordinated the project, and wrote the manuscript. Z.-J.L., Y.S., Y.-Y.S., Y.-Q.Z., L.-C.C., H.D., S.-C.N., J.H., M.W., G.-H.L., X.-J.X., H.H., Y.Y. and H.-J.Y. collected and grew the plant material. Q.X., Z.-J.L., W.-C.T., K.-W.L., L.-J.C., Y.S., Y.-Y.S., M.L. and Y.Y. prepared samples. C.B., G.-Q.Z., Z.-J.L., Y.-Q.Z. and K.-W.L. sequenced and processed the RAW data. Z.-W.W., S.-L.Z., X.Z., C.D., C.B. and G.-Q.Z. annotated the genome. Z.-J.L., Z.-W.W., S.-L.Z., Z.X., L.-S.Z., F.C. and C.D. analyzed gene family. Z.-J.L., Z.-W.W., K.-W.L., G.-Q.Z. and C.B. conducted genome evolution analysis. C.B., L.-S.Z., F.C., Z.-J.L., G.-Q.Z., K.Y., N.M., C.-M.Y., M.O.-T. and Q.X. conducted secondary metabolites and R gene analysis. W.-C.T., Y.-Y.H., Z.-J.L., S.-B.C., L.-S.Z., F.C., W.-L.W., Y.-Y.C. and K.-W.L. conducted the MADS-box gene analysis. C.B., H.D., L.-S.Z., Z.-J.L., G.-Q.Z., Y.-Q.Z. and K.-W.L. conducted transcriptome sequencing and analysis.

Additional Information
Accession codes: Genome sequences have been submitted to the National Center for Biotechnology Information (NCBI). Whole genome assemblies have been deposited in DDBJ/EMBL/GenBank under the accession codes JSDN00000000 (URL: http://www.ncbi.nlm.nih.gov/bioproject/262478).

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zhang, G.-Q. et al. *The Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. Sci. Rep. 6, 19029; doi: 10.1038/srep19029 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/