The *Gastrodia elata* genome provides insights into plant adaptation to heterotrophy

Yuan Yuan1, Xiaohua Jin2, Juan Liu1, Xing Zhao3, Junhui Zhou1, Xin Wang1, Deyi Wang2, Changjiangsheng Lai1, Wei Xu3, Jingwen Huang1, Liangping Zha4, Dahui Liu5, Xiao Ma2, Li Wang6, Menyan Zhou3, Zhi Jiang3, Hubiao Meng1, Huasheng Peng4, Yuting Liang1, Ruiqiang Li3, Chao Jiang1, Yuyang Zhao1, Tiegui Nan1, Yan Jin1, Zhilai Zhan1, Jian Yang1, Wenkai Jiang3 & Luqi Huang1

We present the 1.06 Gb sequenced genome of *Gastrodia elata*, an obligate mycoheterotrophic plant, which contains 18,969 protein-coding genes. Many genes conserved in other plant species have been deleted from the *G. elata* genome, including most of those for photosynthesis. Additional evidence of the influence of genome plasticity in the adaptation of this mycoheterotrophic lifestyle is evident in the large number of gene families that are expanded in *G. elata*, including glycoside hydrolases and urease that likely facilitate the digestion of hyphae are expanded, as are genes associated with strigolactone signaling, and ATPases that may contribute to the atypical energy metabolism. We also find that the plastid genome of *G. elata* is markedly smaller than that of green plant species while its mitochondrial genome is one of the largest observed to date. Our report establishes a foundation for studying adaptation to a mycoheterotrophic lifestyle.
Symbiotic associations between plants and fungi (mycorrhizae) began about 450 million years ago. Most mycorrhizal associations are mutualistic, such that the host plant and mycorrhizal fungi exchange nutrients with each other. However, mycoheterotrophs have evolved a special type of plant–fungi symbiosis in which a plant gets fixed carbon and other nutrients from fungal partners, rather than from photosynthesis. One of the most interesting characteristics of orchids is the reliance on fungi for seed germination and nutrient absorption, for example, through formation of mycorrhiza with fungi. Over 99% of orchids show partial mycoheterotrophy in which young plants obtain carbon (C) nutrients from fungi prior to the development of green leaves, while adult plants are autotrophic. The extreme type of mycoheterotrophy in orchids is obligate mycoheterotrophy, in which plants are achlorophyllous (lack chlorophyll) throughout their life cycle and therefore fully dependent on fungi for nutrition.

Gastrodia elata (Orchidaceae) is an orchid popularly used in traditional Chinese medicine that has a fully mycoheterotrophic lifestyle with highly reduced leaves and bracts in scape, although field guides and systematists often refer to the plants as leafless. During its life cycle, in associates with at least two types of fungi: Mycena for seed germination and Armillaria mellea for plant growth. To obtain nutrition, it forms an association with A. mellea and more than 80% of its ~36-month lifespan is spent underground as a tuber (Fig. 1a). These features of the plant are putative adaptations to its obligate mycoheterotrophic lifestyle. G. elata thus offers the possibility of obtaining a valuable insight into the genetic basis of mycoheterotrophy. Here we present a high-quality reference genome assembly of G. elata (Orchidaceae), and use it to investigate the molecular basis of its full mycoheterotrophic life cycle. The observations presented here will be of value for functional ecological studies seeking to understand the mechanisms and evolutionary basis of plant–fungi associations.

**Results**

**Sequencing and annotation.** The genome of a G. elata individual was sequenced using a whole-genome shotgun (WGS) approach (Supplementary Table 1). Through K-mer distribution analysis, the genome size was estimated to be 1.18 Gb (Supplementary Fig. 1). The assembly consisted of 3779 scaffolds, with a scaffold N50 of 4.9 Mb (total length = 1061.09 Mb) and contig N50 of 68.9 kb (total length = 1025.5 Mb) (Supplementary Table 2). Overall, 98.51% of the raw sequence reads could be mapped to the assembly, suggesting that our assembly results contained comprehensive genomic information (Supplementary Table 3). Gene region completeness was evaluated by RNA-Seq data (Supplementary Table 4): of the 80,646 transcripts assembled by Trinity, 98.66% could be mapped to our genome assembly, and 94.41% were considered as complete (more than 90% of the transcript could be aligned to one continuous scaffold). The completeness of gene regions was further assessed using CEGMA (conserved core eukaryotic gene mapping approach): 239 of 248 (96.37%) conserved core eukaryotic genes from CEGMA were captured in our assembly, and 217 (87.5%) of these were complete (Supplementary Table 5).

Much of the G. elata genome (66.18%) was occupied by transposable elements (TEs). Class I (retrotransposons) and Class II (DNA transposons) TEs accounted for 55.94% and 4.38% of the genome, respectively (Supplementary Table 6). Long terminal repeats (LTRs) formed the most abundant category of TE, with LTR/Gypsy and LTR/Copia occupying 45.04% and 7.10% of the genome, respectively (Supplementary Table 7). Global activity of LTRs was similar between G. elata and Phalaenopsis equestris, while Dendrobium officinale presented a recent burst of LTR activities (Supplementary Fig. 2a). Compared to P. equestris, all LTR families in G. elata had fewer members, except a substantive expansion of del family (Supplementary Table 8 and Supplementary Fig. 2b). Through a combination of ab initio prediction, homology search, and RNA sequence-aided prediction, 18,969 protein-coding genes were predicted in the G. elata genome. Of these genes, 81.6% were functionally annotated (Supplementary Table 9) and 88.69% had detectable transcripts in an RNA-seq analysis of protocorms, tubers (juvenile, immature, and mature tubers), and scapes (Supplementary Table 10). Our transcriptomics analysis revealed that there were 10,548 differentially expressed genes among the five growth stages; these differentially expressed genes clustered into five distinct groups that were representative of the particular stages of growth of G. elata (Supplementary Fig. 3, Supplementary Table 11 and Supplementary Note).

**Phylogeny and whole-genome duplication.** Comparison of the sequenced genomes of the orchid species G. elata, P. equestris, and D. officinalis indicated that they diverged approximately 67 million years ago (Fig. 1b and Supplementary Fig. 4). Two ancient whole-genome duplication (WGD) events are evident in the G. elata genome; these events can also be discerned in the genomes of P. equestris and D. officinalis suggesting they occurred prior to the divergence of the three orchid species (Supplementary Fig. 5). The older WGD event might represent the heWGD event shared by most monocots, while the younger WGD event was likely shared by all extant orchids and might contribute to the divergence of orchid, as suggested in Apostasia shenzhenica genome.

**Extensive gene lost in G. elata genome.** Compared to P. equestris (29,431 protein-coding genes) and D. officinalis (28,910 protein-coding genes), G. elata has a relative small proteome size (18,969 protein-coding genes). The estimated proteome size of G. elata is the smallest theoretical proteome so far identified among angiosperm genomes (Supplementary Table 12). Comparison of G. elata, P. equestris, and D. officinalis genes that have functional annotation information revealed global gene set reduction in the G. elata genome. For example, almost all second level gene ontology (GO) categories had fewer genes in G. elata than in the other two species, and 9 of these categories (16.7%) were significantly reduced (Fisher’s Exact test, \( p < 0.05 \), Supplementary Fig. 6 and Supplementary Table 13). We also found that several Pfam domain families were significantly reduced in the G. elata genome (Supplementary Table 14). Among the 14 angiosperm used in the phylogenetic analysis, G. elata had the lowest number of gene families; moreover, G. elata had on average the lowest number of genes in each gene family (Fig. 1c, and Supplementary Table 15). This consistently low number of genes and gene families suggests that many gene families have been eliminated from the G. elata genome, and further suggests that many of the remaining gene families have contracted. Gene family expansion and contraction analysis based on maximum likelihood modeling of gene gain and loss confirmed that 3586 gene families had undergone contraction in G. elata, much more compared to the other two orchid genomes (Supplementary Fig. 7 and Supplementary Table 16). A Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis, which assessed 956 orthologous groups with genes present as single-copy in at least 90% of plant genomes, revealed that 195 (20.4%) highly-conserved genes were missing from the G. elata genome. This rate of absence is much higher than in the genomes of the 13 land species that were included in this analysis. All of these analyses indicate that G. elata has undergone extensive gene
Fig. 1 Gastrodia elata life cycle and gene-family contraction. **a** The main developmental stages of G. elata. Seeds develop into a protocorm stage without requiring *A. mellea* (A). The protocorm then differentiates into a corm stage after commencing its association with *A. mellea*; note that lateral buds develop into juvenile tubers (B). Young and immature tubers (C). Mature tuber with an emergent young scape (D). Scape (stem and inflorescence) of mature plants (E). **b** Phylogenetic tree of 14 plant species including *G. elata*. The red dot represents a calibration point determined from the timetree website. **c** Bar graph of the number of protein-coding genes in each of the species. Single-copy orthologs include common orthologs with one copy in specific species. Multi-copy orthologs include common orthologs with multiple copy numbers in specific species. Other orthologs include genes from families shared in 2–13 species. ‘Eudicot’ clusters with eudicots. ‘Monocots’ clusters with monocotyledonous plants. ‘Orchid’ clusters with *G. elata*, *P. equestris*, and *D. officinale*. ‘Unclustered’ include genes that cannot be clustered into gene families. **d** Analysis of gene numbers in the genomes of four species for the nucleotide-binding site gene family (NBS), the pathogenesis-related protein (PR) family (Gel *G. elata*, Peq *P. equestris*, Dof *D. officinale*, Aco *A. comosus*). Numbers in circles represent the number of family members in each genome, and numbers with plus or minus signs indicate, respectively, the number of duplicated or deleted genes. **e** The plastid genomes of *P. equestris* (outer circle) and *G. elata* (inner circle). Red, protein; orange, rRNA; green, tRNA; gray, genes lost from the plastid genome of *G. elata*.
losses, even for genes that were conserved in other plant species that have also undergone extensive lost events.

The absence of these genes is unlikely to be due to genome assembly problems because 98.66% of the transcripts assembled from transcriptome data could be mapped to the assembly. Another possibility is that several genes were missed due to gene prediction problems. By mapping RNA reads onto the annotated genome, we found that the majority of RNA reads (>86%) from all G. elata tissues could be mapped to annotated exon regions (Supplementary Table 18). This rate of mapping was comparable to that achieved in the well-annotated rice genome and higher than in the P. equestris genome (Supplementary Table 18). Through analysis of gene synteny among G. elata and P. equestris and D. officinale, we detected 2961 gene deletion events in G. elata versus P. equestris, and 3120 gene deletion events in G. elata versus D. officinale (Supplementary Table 19). Further TBLASTN searches of these deleted genes recovered less than 3% of them. Of these genes, fewer than 15% were supported by RNA-seq data (Supplementary Table 19). Both the RNA mapping results and the synteny deletion analysis confirmed that our gene prediction was comprehensive; thus, the possibility of missing gene annotations was low. Finally, PCR amplification of 18 lost genes...
orchid species with sequenced genomes. The plastid genomes of these two species comprise two single-copy regions (a large and a small single-copy region) and the two identical large inverted repeats (IRs) encode 75 and 76 genes, respectively, that are most associated with photosynthesis. The G. elata plastid genome has lost one IR and encodes only 19 protein-coding genes (Fig. 1e), suggesting that G. elata is an ancient mycoheterotroph and that its plastid genome is in the last stage of a ‘degradation ratchet’, i.e., retention and loss of the five core nonbioenergetic genes17–18. Excluding the possibility that these genes were missed by our genome assembly, the transcriptome sequencing analysis indicated that none of the deleted plastid or nuclear encoded genes were expressed in G. elata, while the five core nonbioenergetic genes, trnE, aacD, clpP, ycf1, and ycf2, were moderately to highly expressed in all five stages in G. elata (Supplementary Tables 30, 31). These results clearly show that both the plastid and nuclear genomes of G. elata have lost most of the genes required for photosynthesis, although the highly degraded plastome is still essential for this full mycoheterotroph.

Expansion of mitochondrial genome. Although the G. elata genome has clearly undergone extensive gene loss, we found that 430 gene families (19 by a significant margin), containing 1532 genes (184 by a significant margin), showed expansion in G. elata compared to P. equestris, D. officinale, and A. comosus (Supplementary Fig. 7 and Supplementary Tables 32, 33). These genes are enriched for GO terms related to several metabolic processes (Fig. 2a, Supplementary Table 32). We speculate that these expanded genes are related in some way to the functional requirements of the obligate mycoheterotrophic lifestyle of G. elata. We first sequenced and assembled the mitochondrial genome to explore this idea, and the mitochondrial genome G. elata is markedly expanded in size (1339 kb, Fig. 2b) compared to the mitochondrial genomes of most other seed plants19. Thirty-seven protein-coding genes were annotated, and one subunit of mitochondrial ATP synthase, atp4, had two copies in the mitochondrial genome of G. elata and was highly expressed in the cortex layer (Supplementary Table 34). In addition, 36 of the genes had detectable expression in mature tubers (epidermis, cortex, and parenchymal cell) using a tissue-specific qPCR-based analysis (Supplementary Table 34).

Management of symbiotic microbials. We next explored how gene expansion in G. elata may have contributed to its association and interactions with fungal microbiota. The monocot mannose-binding lectin antifungal protein family (GAFP) of G. elata contains 20 genes, compared to only 3 in A. comosus and 0 in A. thaliana (Supplementary Table 27). GAFP proteins have been documented to inhibit the growth of both ascomycete and basidiomycete fungal plant pathogens in vitro20. More than 80% of the GAFP genes were highly expressed in protocorms and juvenile tubers, the growth stages that occur before G. elata establishes

Fig. 2 Gene expansion in G. elata and microbial community analysis. a REVIGO semantic similarity scatter plot of Biology Process Gene Ontology terms for expanded genes in G. elata. In semantic spaces, the proximity between circles represents relatedness (similarity) of the GO terms. Similar GO terms are close together in the plot. The axes in the plot have no intrinsic meaning, but were used to measure pairwise similarities between GO terms. Color indicates degree of enrichment for each process presented as the p-value from the hyper-geometric test. b The draft mitochondrial genome of G. elata. Nineteen contigs are manually displayed as a circle, including 12 circular contigs in orange (ornamented with stars) and 7 linear contigs (in blue). The genes are indicated in the middle circle, and are color coded as follows: tm (blue), mm (light blue), atp (red), and other protein-coding genes (black). The duplicated atp genes and their fragments are detailed in the inner circle. The duplicated genes are suffixed with ‘b’, and the gene fragments are suffixed with ‘fragment’. c Gene expression heat map of the normalized RNA-Seq data for genes encoding the monocot mannose-binding lectin antifungal proteins (GAFP) in G. elata20. The units indicate the expression levels of different gene members of GAFP in the protocorm, juvenile tuber, immature tuber, mature tuber, and scape of G. elata (only shown where the gene expression level RPKM > 1, n = 3). d Venn diagrams showing the number of shared and unique fungal and bacterial operational taxonomic units (OTUs) based on the ITS and 16S sequence analyses in protocorms, juvenile tubers, immature tubers, and mature tubers of G. elata. OTUs showed the composition and abundance of the microbe species, which were defined at 3% dissimilarity.
a stable symbiotic association with *A. mellea* (Fig. 2c). 4-Hydroxybenzyl alcohol (*p*-PA), the precursor of the phytoalexin gastrodin is a major phenolic compound of *G. elata*21. The expression of *p*-PA biosynthesis genes (e.g., cinnamate 4-hydroxylase, *C4H*, alcohol dehydrogenase, *ADH*, hydroxybenzaldehyde synthase, *HBS*)21 was relatively high in protocorms and juvenile tubers. Ultra-performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry (UPLC-ESI-Q-TOF-MS) metabolite analysis revealed that *p*-HA is present in *G. elata* tubers but not in *A. mellea* hyphae (sampled from tuber, wood, and PDA medium). S-(*p*-HA)-glutathione was detected in both *G. elata* tubers and in *A. mellea* hyphae sampled from tubers (Supplementary Fig. 12, Supplementary Table 35), which putatively suggests that *G. elata* may transport this phytoalexin to *A. mellea* and prevent the excessive growth of *A. mellea*. To investigate the effect of *A. mellea* on microbial management in *G. elata*, we performed a 16S ribosomal (rRNA) and rDNA ITS sequencing analysis and found that the diversity of bacterial and microbial species was significantly lower during the protocorm stage than at other growth stages (*p < 0.05*), which was consistent with the pattern of gene expression of *GAFP* (Fig. 2c, d, Supplementary Tables 36, 37). This increased diversity of bacteria and fungi during the juvenile tuber to mature tuber periods implies that a compatible mycorrhizal fungus (*A. mellea*) can affect the structure of the microbial community associated with its host and greatly reduce the antifungal and antibacterial activities as a symbiotic association with *A. mellea* is established.

**Signaling and nutrition transfer in *G. elata***. Without the ability to perform photosynthesis, *G. elata* depends completely on its symbiotic fungus for nutrition. It is thus obvious that the signaling pathways related to the establishment of this symbiotic relationship are crucial for *G. elata*. Some of the mechanisms underlying the symbiotic interaction between *G. elata* and *A. mellea* are similar to those for interactions between other plants and arbuscular mycorrhizal (AM) fungi22. The *G. elata* genome contains many of the genes known to participate in AM associations (Fig. 3a, Supplementary Table 38). Key genes for biosynthesis and secretion of strigolactone were expanded in *G. elata* (e.g., carotenoid cleavage dioxygenases, CCDs, for biosynthesis23 and ABC transporters, PDRs, for secretion24) (Supplementary Table 38). It is known that strigolactone can stimulate hyphal branching and development of arbuscular mycorrhizal fungi, which increases the chances of an encounter with a host plant24. We conducted growth assays and confirmed that strigolactone had similar branch-inducing effects in *A. mellea* (Fig. 3b, Supplementary Fig. 13). The expanded number of genes encoding CCDs and PDRs suggests that *G. elata* has enhanced its ability to interact with *A. mellea* to increase the efficiency of the establishment of the symbiotic relationship essential for its nutrition and metabolism. Calmodulin-dependent protein kinase gene of the *does-not-make-infections* 3 subfamily (DMI3) were also doubled or tripled in *G. elata* (10 genes) compared to *P. equestris* (3 genes), *D. officinale* (5 genes), and *A. comosus* (4 genes); these genes participate in the Ca<sup>2+</sup> spiking process that has been shown to regulate the colonization of plants by fungi25.
After A. mellea colonizes G. elata, fungal growth is restricted to its cortex layer (Fig. 3c, Supplemental Fig. 14). We performed a tissue-specific qPCR-based analysis of 10 genes in G. elata tubers and found that PDR transcripts, which mediate secretion of strigolactone to the extracellular space, were highly abundant in the cortex layer (Supplemental Fig. 15). This finding suggests that G. elata may preferentially guide A. mellea to colonize its cortex layer. Similarly to ATP synthases, we found that some glycoside hydrolases from gene families that have expanded in the G. elata genome were also highly expressed in the cortex layer, supporting the idea that A. mellea hyphal walls are digested in the cortex layer of G. elata tubers (Supplementary Table 39). The expanded endo-β-1,4-d-xylanase and β-glucosidase may have become neo-functionalized to cleave fungal glycan substrates during the digestion of hyphal walls of A. mellea (Fig. 3a, Supplementary Table 40).26,27

Given that the ANTI-like aromatic and neutral amino acid transporters (ANT)28 are known to translocate arginine (Arg), which is a key component in nitrogen translocation in arbuscular mycorrhiza fungi20, it seems likely that Arg in G. elata is related to pathogen metabolism such as symbiotic (Fig. 3a, Supplementary Table 41). It is known that arginases can hydrolyze Arg into urea in mycelia, which is further hydrolyzed to ammonium and carbonic acid by ureases.30 Although P. equestris, D. officinale, A. comosus and A. thaliana have only one copy of glutamate N-acetyltransferase (Arg), an enzyme of the arginine biosynthesis pathway31, G. elata has three copies (Supplementary Fig. 11 and Supplementary Table 41). The number of genes encoding ureases is drastically expanded in G. elata (9 genes) compared with P. equestris (2 genes), D. officinale (2 genes), A. comosus (1 gene), and A. thaliana (1 gene) (Supplementary Table 41). This suggests that urea metabolism might be an important source of nitrogen for G. elata (Fig. 3a).

Conclusion. The extensive deletion and expansion of genes, especially the global reduction of gene complements in almost all functional categories in the G. elata genome, provides a powerful example of how a plant with a fully heterotrophic life cycle has made use of genome plasticity to achieve extensive neo-functionalization and gene loss. Our results establish a unique opportunity for researchers to understand how plants that have abandoned photosynthesis continue to persist and thrive.

Methods
Plant materials and DNA preparation. The experimental materials of Gastrodia elata were harvested from Xiaoxiachun in Yunnan Province (latitude 27.79°N longitude 104.24°E) located in the southwestern China. Genome sequencing and assembly was done on the scope of beige-scape G. elata. Five transcriptiones were sequenced from five different G. elata tissues (protoctum, juvenile tuber, immature tuber, mature tuber, scape). Four different G. elata tissues (protoctum, juvenile tuber, immature tuber, mature tuber) were collected to investigate the diversity of microbial communities. High-quality genomic DNA was extracted using the Qiagen DNeasy Plant Mini Kit.

Genome sequencing and assembly. Multiple paired-end and mate-pair libraries were constructed with a spanning size that ranged from 180 bp to 20 kb. Sequencing was conducted on an Illumina HiSeq 2500 platform. In total, 179.1 Gb raw sequencing reads were produced (Supplementary Table 1). Raw sequencing reads were subjected to filtering to remove (1) low quality reads with low quality bases (>50% bases with Q-value ≤5); (2) reads with Ns >10% of the read length; (3) reads with adapter contamination; and (4) duplicated reads caused by PCR during library construction. Filtered data were assembled using ALLpaths-LG (version 44080)52, where overlapping paired-end reads with an insert size of 230 nucleotides were used as fragment libraries, and all other libraries (>230 nucleotide insert size) were used as jumping libraries. The Allpaths-LG assembly was run with default settings, then a gap filling step was carried out using GapCloser based on the paired-end information of the paired-end reads that had one end mapped to the unique contig and the others located in the gap region (http://sourceforge.net/projects/soapdenovo2/files/GapCloser).

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03423-5 | www.nature.com/naturecommunications

Genome-quality evaluation. To evaluate the completeness of the assembly and the uniformity of the sequencing, all the paired-end reads were mapped to the genome using BWA.53 The number of genes encoding Arg into urea in mycelia, which is further hydrolyzed to ammonium and carbonic acid by ureases.30 Although P. equestris, D. officinale, A. comosus and A. thaliana have only one copy of glutamate N-acetyltransferase (Arg), an enzyme of the arginine biosynthesis pathway31, G. elata has three copies (Supplementary Fig. 11 and Supplementary Table 41). The number of genes encoding ureases is drastically expanded in G. elata (9 genes) compared with P. equestris (2 genes), D. officinale (2 genes), A. comosus (1 gene), and A. thaliana (1 gene) (Supplementary Table 41). This suggests that urea metabolism might be an important source of nitrogen for G. elata (Fig. 3a).

Conclusion. The extensive deletion and expansion of genes, especially the global reduction of gene complements in almost all functional categories in the G. elata genome, provides a powerful example of how a plant with a fully heterotrophic life cycle has made use of genome plasticity to achieve extensive neo-functionalization and gene loss. Our results establish a unique opportunity for researchers to understand how plants that have abandoned photosynthesis continue to persist and thrive.

Dynamics of long terminal-repeat retrotransposons. Intact Long terminal-repeat retrotransposons (LTR) were identified by searching the genomes of G. elata, D. officinale and P. equestris from LTRharry59 from Genome Tools v1.5.1. The candidate sequences were filtered by two-step procedure to reduce false positives. First, LTRdigest was used to identify the primer binding sites (PBS) in sequences. Then, the predicted sequences from the whole elements contained PBS were retained; then protein domains (pol, gag and env) in candidate LTR retrotransposons were identified by searching against HMM profiles collected by Gypsy Databases (GyDB)61. Elements contained gag domain, protease domain, reverse transcriptase (RT) domain and integrase domain, which were considered as intact. Second, all families of these intact LTR retrotransposons were clustered using the previously described method47. Finally, LTRs that did not contain protein domains or that belonged to families with less than 5 members were discarded. The EMBOSS program distmat was used to estimate LTR divergence rates between the 5′- and 3′-LTR sequences of the intact LTRs (Supplementary Fig. 2).

Gene prediction. Gene prediction was conducted through a combination of homology-based prediction, ab initio prediction and transcript-based prediction methods. Protein repertoire of plants including A. comosus10, Amborella trichopoda32, Arabidopsis thaliana (phytozome10), Brachypodium distachyon (phytozome10), D. officinale1, O. sativa (phytozome10), P. equestris, Vitus vinifera (phytozome10), Sorghum bicolor (phytozome10) and Zea mays (phytozome10) were downloaded and mapped to the G. elata genome using TRALASTN (E-value ≤5e-5). The BLAST hits were conjoined by Solar software46. GeneWise (version 2.4.1)65 was used to predict the exact gene structure of the corresponding genomic region on each BLAST hit. Homology predictions were denoted as “Homology-set”. RNA-seq data derived from protocorn, juvenile tuber, immature tuber, mature tuber, and scape (Fig. 1a) were assembled by Trinity (version 2.0)41. The Trinity assembly included 183,515 contigs with an average length of 592 bp. These assembled sequences were aligned against the G. elata genome by PASA (Program to Assemble Spliced Alignment)44. Valid transcript alignments were clustered based on genome mapping location and assembled into gene structures. Gene models generated by PASA were annotated as GlimmerHMM66. Besides, RNA-seq reads were directly mapped to the genome using Tophat (version 2.0.8)69 to identify putative exon regions and splice junctions; Cufflinks (version 2.1.1) was then used to assemble the mapped reads into gene models (Cufflinks-set). Augustus (version 2.3.5)50, GeneID (version)51, GeneScan (version 1.0)53, GlimmerHMM version 3.0.115 and SNAP (version)48 were also used to predict coding regions in the repeat-masked genome. These of Augustus, SNAP and GlimmerHMM were trained by PASA-H-set gene models. Gene models generated from all the methods were integrated by EvidenceModeler (EVM)48. Weights for each type of evidence was set as follows: PASA-T set > Homology-set > Cufflinks-set > Augustus > GeneID = SNAP = GlimmerHMM = GeneScan. The gene models were further updated by PASA2 to generate UTRs, alternative splicing variation information, which generated 26,872 gene models. Gene models only supported by ab initio evidence were filtered out. To reduce the possibility of missing and poorly annotated genes, we invested additional effort in annotating some gene families that could be missed by automated genome annotation, such as NBS-encoding genes. In total, 1943 protein sequences containing an NB-ARC domain were searched against the G. elata genome using TBLASTN with a threshold of 1e-5. All
BLAST hits in the genome, together with 5000 bp flanking regions on both sides, were annotated by the GeneWise program. The resulting predictions were surveyed to verify whether they encoded NLS or LRR motifs using Pfam. We also focused on other genes, such as those related to photosynthesis, and transporter, and these were manually annotated through a combination of BLAST search and motif verification. Ultimately, a comprehensive non-redundant reference gene set was produced that contained 18,969 protein-coding gene models. Functional annotation of the protein-coding genes was carried out using BLASTP (E-value cut-off 1e−5). Before the BLAST search, regions of the 18,969 true genes were masked. The BLAST hits were conjoined by Solar software. GeneWise was used to predict the pseudo gene structures with the ‘pseudo’ parameter. Pseudo genes were then classified by PseudoPipe. The PseudoPipe program applies a set of sequence identity and completeness cut-off to report a final set of good-quality pseudo gene sequences. We used the following cutoffs: amino acid (AA) sequence identity >80% and match length >50 AA to filter out false positives. GeneWise results that fulfilled the cut-off criteria were denoted as high-confidence pseudo genes. High-confidence pseudo genes were then assigned to three categories. (1) Processed/retrotransposed pseudo gene sequences were formed through retrotransposition. Retrotransposition occurred by reintegration of a cDNA, a reverse transcribed mRNA into the genome of a new location. (2) Duplicated pseudo gene sequences (DUPS), which formed through gene duplication, following by decay of genes, include frameshifts or premature stop codons. (3) Pseudo gene fragments (FRAGs), which were fragments with identities of at least 80% and length greater than 200 bp, were filtered out as reliably assessed as processed or duplicated. We used the following criteria to classify PSSDs, DUPS, and FRAGs: (i) PSSDs, exon number = 1, 0.7 < align ratio ≤ 0.95, 0.3 ≤ identity ≤ 0.95; (ii) DUPS, exon number > 1, 0.3 ≤ identity ≤ 0.95, and existing insertion, deletion, termination, or frameshift; (iii) FRAGs, exon number = 1, align ratio ≤ 0.7, 0.3 ≤ identity ≤ 0.95.

**Identification of pseudo genes.** Pseudogenes in the *G. elata* genome were identified by searching against *G. elata* intergenic regions using *D. officinale* or *P. equestris* protein sequences as the seed sequences (TLASTN, E-value cut-off 1e−5). Before the BLAST search, regions of the 18,969 true genes were masked. The BLAST hits were conjoined by Solar software. GeneWise was used to predict the pseudo gene structures with the ‘pseudo’ parameter. Pseudo genes were then classified by PseudoPipe. The PseudoPipe program applies a set of sequence identity and completeness cut-off to report a final set of good-quality pseudo gene sequences. We used the following cutoffs: amino acid (AA) sequence identity >80% and match length >50 AA to filter out false positives. GeneWise results that fulfilled the cut-off criteria were denoted as high-confidence pseudo genes. High-confidence pseudo genes were then assigned to three categories. (1) Processed/retrotransposed pseudo gene sequences were formed through retrotransposition. Retrotransposition occurred by reintegration of a cDNA, a reverse transcribed mRNA into the genome of a new location. (2) Duplicated pseudo gene sequences (DUPS), which formed through gene duplication, following by decay of genes, include frameshifts or premature stop codons. (3) Pseudo gene fragments (FRAGs), which were fragments with identities of at least 80% and length greater than 200 bp, were filtered out as reliably assessed as processed or duplicated. We used the following criteria to classify PSSDs, DUPS, and FRAGs: (i) PSSDs, exon number = 1, 0.7 < align ratio ≤ 0.95, 0.3 ≤ identity ≤ 0.95; (ii) DUPS, exon number > 1, 0.3 ≤ identity ≤ 0.95, and existing insertion, deletion, termination, or frameshift; (iii) FRAGs, exon number = 1, align ratio ≤ 0.7, 0.3 ≤ identity ≤ 0.95.

**Phylogenetic tree reconstruction.** Protein sequences from 74 single-copy gene families were used for phylogenetic tree reconstruction. MUSCLE was used to generate multiple sequence alignment for protein sequences in each single-copy family with default parameters. Then, the alignments of each family were concatenated to a super alignment matrix. The super alignment matrix was used for phylogenetic tree reconstruction through maximum likelihood (ML) methods. Before ML reconstruction, we used ProtTest to select the best substitution models. The JTT + I + G + F model was selected as the best-fit model, and RAXML was used to reconstruct the phylogenetic tree.

**Species divergence time estimation.** Divergence time between 14 species was estimated using MCMC in PAML with the options ‘correlated molecular clock’ and ‘JC69’ model. A Markov Chain Monte Carlo analysis was run for 20,000 generations, using a burn-in of 1000 iterations. Five calibration points were applied in the present study (Fig. 1): *P. equestris* and *D. officinale* divergence time (47–52.9 million years ago) was used to calibrate the *P. equestris* and *D. officinale* *mitochondrial* genomes; *Z. mays* divergence time (24–84 million years ago) was used to calibrate the *Z. mays* *mitochondrial* genome; *A. thaliana* and *P. trichocarpa* divergence time (65–89 million years ago) was used to calibrate the *A. thaliana* and *P. trichocarpa* *mitochondrial* genome; *G. max* divergence time (56–89 million years ago) was used to calibrate the *G. max* *mitochondrial* genome, and root of land plants (407–557 million years ago) was used to calibrate land plants (407–557 million years ago).
Auckland, New Zealand: http://www.geneious.com) for initial assembly. The contigs generated by the initial assembly were used as seeds for further iterative mapping and extension processes. Velvet and Geneious were alternatively used during assembly with multiple combinations of k-mer lengths. In most cases, the extension process of the assembly worked well. In particular, when the head and tail of a contig had an overlapping region and could not be further extended, this contig could be reasonably connected into a circle. Although several circles were produced during assembly, some problems did arise in the extension process. For example, some contigs were displayed as single lines because their boundaries were too difficult to determine due to poly structures or repeats in the mitogenome. The final assembled results were verified by remapping and some ambiguous regions with low coverage were further checked by PCR. Overall, 19 local database with an the mitogenome. The several circles were produced during assembly, some problems did arise in the mapping and extension processes. Velvet and Geneious were alternatively used between groups using R software (http://www.r-project.org/).

RNA extraction and quantitative PCR
μg of total RNA was reverse transcribed at 42 °C using TransScript manufacturer's recommendations. Prior to use in qPCR, cDNA was diluted 1:5 with H2O.

The qPCR reactions were performed in duplicate for each condition using the KAPA SYBR® FAST qPCR Master Mix (KapaBiosystems, USA) and LightCycler® 480 Real-Time PCR System (Roche, Switzerland). Each reaction consisted of 20 μL of cDNA and 200 nM of each primer (Supplementary Tables 59 and 60). The cycling conditions were: denaturation at 95 °C for 3 min; followed by 45 two-segment cycles of amplification at 95 °C for 10 s, and 60 °C for 30 s in which ΔΔCt was selected as the internal control and the expression levels of tested genes were determined using the comparative Ct (2−ΔΔCt) method.

Diversity of microbial communities.
Four different G. elata tissues (protocorm, juvenile tuber, immature tuber, mature tuber) were collected and total genomic DNA were extracted using hexadecyl trimethyl ammonium bromide (CTAB). The 16S V4 and ITS1 genes in all sample were amplified using the universal primers 515F-806R and ITS5-1737F with a barcode as a marker for distinguishing samples. The PCR was performed with Phusion® High-Fidelity PCR Master Mix (New England Biolabs, Ipswich, MA, UK). PCR products were mixed in equidensity ratios. Then, the mixed PCR products were purified with QIAquick Gel Extraction Kit (Qiagen, Germany). Sequencing libraries were generated using a TruSeq® PCR-FREE Sample Preparation Kit (Illumina, USA) following the manufacturer’s recommendations and index codes were added. The library quality was assessed on a Qubit® 2.0 Fluorometer (Thermo Scientific) and Agilent Bioanalyzer 2100 system. The library was sequenced on an Illumina HiSeq 2500 platform and 250 bp paired-end reads were generated.

High quality sequences were clustered into OTUs defined at 97% similarity. These OTUs were applied for diversity, richness and rarefaction curve analyses using MOTHUR. Taxonomic assignments of OTUs that reached the 97% similarity level were made using the QIME (quantitative insights into microbial ecology) software package through comparison with the SILVA, Greengene, and RDP databases. Venn diagrams were generated to identify the mutual and specific taxons between groups using R software (http://www.r-project.org/).

Microscopy.
For identification and analysis of G. elata infected by A. mella, hand sections were cut through the infection point of an immature tuber, and the sections were then embedded in agar plates. Images were captured using a Zeiss AX10 fluorescence microscope with ×10 water immersion lenses. Owing to the spontaneous blue fluorescence, both visible and DAPI filters were used to observe the hyphae of A. mella with fluorescent microscopy (Zeiss AX10). To analyze tissue and cell structures of G. elata uninfected by A. mella, paraffin sections (10 μm thickness) were obtained using a Thermo Scientific MicRoM HM 325 sliding microtome. For light microscopic observations of the highly lignified hyphae of A. mella, sections were stained with Fast Green stain reagent to investigate the infected cells of G. elata. After staining, sections were washed by PBS three times, dehydrated through an alcohol (50%, 80%, 90%, 100%), cleared in xylene, sealed with neutral gum, and observed using a fluorescence microscope (Zeiss AX10).

Quantification of p-HA and S-(p-HA)-glutathione. About 0.35 g of frozen fresh G. elata were homogenized and ultrasonically extracted for 30 min in four volumes (g mL−1) water. After centrifugation (13,000 rpm, 10 min), 100 μL aliquots of supernatant were mixed with 10 μL of rutin (101.0 μg mL−1) for quantification of p-HA (A 270 nm, Rt 6.8 min) and S-(p-HA)-glutathione (m/z 412.12, Rt 8.9 min) using an UPLC-PDA-ESI-Q TOF-MSE method72. The injection volume was 1 μL. The contents were determined by the peak intensity ratios of the analyte to rutin (m/z 609.14, 15.5 min).

Hyphal-branching assay.
Hyphal branching in A. mella fungus was evaluated in vitro by the paper disk diffusion method74. Primary hyphae were cultured in PDA medium containing 20 g L−1 glucose, 4 g L−1 potato powder and 14 g L−1 agar. The dishes were cultured in the dark for 5–7 days at 23 °C. Secondary hyphae emerge from primary hyphae and grow upward in a negative geotropic manner in the growth of secondary hyphae was determined by mitosis. The control was on the opposite direction of the paper without 5-deoxy-striogly. Hyphal branch patterns were analyzed at 24 h and 48 h after treatment. The sample was scored as positive for hyphal branching if new hyphal branches formed from the treated secondary hyphae. The assay was repeated at least twice, using between three and five dishes for each concentration.

Data availability. Genome data were deposited in GenBank under accession number PVEL00000000 and transcriptome sequence reads were deposited in the Sequence Read Archive (SRA) under accession number SRX2879747. The standard flowgram format (SFF) files related with bacterial and fungal communities were also deposited in the SRA under study accession SRX2876148. Plastid genome data were deposited in GenBank under accession number MF163256. Mitochondrial genome data were deposited in GenBank under accession numbers MF070084-MF070102.

Received: 20 June 2017 Accepted: 12 February 2018

Published online: 24 April 2018

References
1. Beckam, D. S. et al. Molecular evidence for the early colonization of land by fungi and plants. Science 293, 1129-1133 (2001).
2. And, S. E. S. & Gianinazzi-Pearson, V. Physiological interactions between symbionts in vesicular-arbuscular mycorrhizal plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. 39, 221-244 (1988).
3. Hyvönen, N. A., Press, K., Gebauer, G. & Bruns, T. D. Isotopic evidence of full and partial myco-heterotrophy in the plant tribe Pyroleae (Ericaceae). New Phytol. 182, 719 (2009).
4. Leake & Jonathan, R. Plants parasitic on fungi: unraveling the fungi in myco-heterotrophy and debunking the ‘saprophytic’ plant myth. Mycol. 113, 113-122 (2005).
5. Tsai, C. C. et al. Comparative transcriptome analysis of Gastrodia elata (Orchidaceae) in response to fungus symbiosis to identify gastrodin biosynthesis-related genes. BMC Genom. 17, 1-16 (2016).
6. Zhang, W. J. & Li, B. F. The biological relationship of Gastrodia elata and Armillaria mellea. Bull. Bot. 22, 57-62 (1980).
7. Cai, J. et al. The genome sequence of the orchid Phalaenopsis equestris. Nat. Genet. 47, 65 (2014).
8. Yan, L. et al. The genome of Dendrobium officinale illuminates the biology of the important traditional Chinese orchid Herb. Mol. Plant 8, 922–934 (2015).
9. Zhang, G. Q. et al. The Dendrobium catenulatum Lindl. genome sequence provides insights into polyaccharide synthesis, floral development and adaptive evolution. Sci. Report. 6, 19029 (2016).
10. Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. Nat. Genet. 47, 1435–1442 (2015).
Acknowledgements
This research was funded by the Fundamental Research Funds for the Central public welfare research institutes (ZZ10-008) and Key project at central government level for the ability establishment of sustainable use for valuable Chinese medicine resources (2060302).

Author contributions
L.H. is the lead investigator. Y.Y. and W.I. coordinated the project and directed work on the reference genome of G. elata. X.J., D.W., and X.M. assembled the plastid and mitochondrial genomes and analyzed gene duplication. X.Z., W.X., M.Z., Z.J., R.L. assembled the reference genome and analyzed genome characterization. J.L. performed experiments of diversity of microbial communities in G. elata during different periods. Y.Y., J.L., X.Z., and L.Z. performed bioinformatic analyzes to annotate the reference genome. J.Z. and C.J. extracted high-quality RNA and qPCR analysis. C.L., T.N., Z.Z., and J.Y. conceived and conducted the analyses of the metabolomics. X.W. and H.P. observed the microscopic characterization of G. elata. J.H., Y.Y., and Y.L. worked on culture of Armillaria mellea and the hyphal-branching assay. D.L., L.W., Y.Y., and Y.Z. collected samples. Y.J. and H.M. verified the gene loss using PCR.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-03423-5.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018