MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 1012–1021 doi: 10.1111/j.1755-0998.2011.03044.x

# DNA barcoding of the recently evolved genus *Holcoglossum* (Orchidaceae: Aeridinae): a test of DNA barcode candidates

XIAO-GUO XIANG,*‡ HAO HU,†‡ WEI WANG* and XIAO-HUA JIN*
*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China, †College of Life Sciences and Technology, Gansu Agricultural University, Lanzhou 730070, China

## Abstract

**Orchidaceae is one of the largest families of flowering plants. Many species of orchid are endangered, and all species are included in Conventions on International Trade of Endangered Species of Fauna and Flora (CITES) I and II, but it is very difficult to identify orchid species, even those with fertile parts. The genus *Holcoglossum* (Orchidaceae: Aeridinae) has long been problematic in taxonomy. It consists of both long-evolved and radiated species and is an excellent case to use for testing DNA barcodes for Orchidaceae. We investigated the power of a subset of proposed plant barcoding loci [*rbcL*, *matK*, *atpF-atpH*, *psbK-psbI*, *trnH-psbA* and internal transcribed spacer (ITS)] to discriminate between species in this genus. Our results showed that all these DNA regions, except *psbK-psbI* and *atpF-atpH*, can be amplified easily from *Holcoglossum* and sequenced with established primers. The DNA regions *matK* and ITS had the highest variability. Among the six loci, *matK* resolved eight of the 12 *Holcoglossum* species and had the highest discriminatory ability. However, the combination of *matK* and ITS showed a greater ability to identify species than *matK* alone. Single or combined DNA markers discriminated between *Holcoglossum* species distributed in tropical areas effectively, but had less ability to identify radiated species from the temperate Hengduan Mountains of China. In the study, *matK* proved to be a useful DNA barcode for the genus *Holcoglossum*; however, complementary DNA regions are still required to accelerate the investigation and preservation of radiated species of orchid.**

*Keywords*: barcoding, DNA markers, *Holcoglossum*, ITS, *matK*, *trnH-psbA*

*Received 14 October 2010; revised received 29 May 2011; accepted 3 June 2011*

## Introduction

Orchidaceae is one of the largest and most complex families of flowering plants, comprising approximately 22 500 species (Atwood 1986; Mabberley 2008). All of these species are listed in the Conventions on International Trade of Endangered Species of Fauna and Flora (CITES) Appendices I and II. Orchids are common targets for *ex situ* and *in situ* conservation, for example, in natural reserves, germplasm banks, and botanical gardens. However, it is well-known that the species to which an orchid belongs is difficult to identify, even when they are in flower (Dressler & Dodson 1960; Dodson 1962; Van den Berg *et al.* 2000; Gravendeel *et al.* 2001; Pridgeon *et al.* 2001; Koehler *et al.* 2002; Salazar *et al.* 2003; Cameron 2004; Fan *et al.* 2009). Recently, studies have indicated that the use of DNA barcodes might be a promising approach to identification (Hebert *et al.* 2003; Newmaster & Ragupathy 2009; Starr *et al.* 2009; Van de Wiel *et al.* 2009; Liu *et al.* 2010; Pryer *et al.* 2010; Ren *et al.* 2010).

DNA barcoding is a method of identifying and recognizing species that is based on the comparison of DNA sequence data (Hebert *et al.* 2003, 2004; Hogg & Hebert 2004). In a study that sampled more than 1000 species of Orchidaceae, Lahaye *et al.* (2008a) demonstrated that the chloroplast gene *matK* could be amplified and aligned easily and could be used to identify more than 90% of species correctly. In the light of this result, the authors proposed that *matK* should be adopted as a universal DNA barcode for flowering plants (Lahaye *et al.* 2008a). However, the use of *matK* has been shown to have limitations. For example, on the basis of results from a study of the Australian genus of orchid *Caladenia*, Farrington *et al.* (2009) indicated that *matK* was less informative at the subgeneric level and not suitable for discrimination below this level. They suggested that the *trnL* intron was a more useful barcode to use for identifying variation at species level. In yet another attempt to find a suitable barcode, Yao *et al.* (2009) demonstrated that the intergenic spacer region of *psbA-trnH* could be useful for

Correspondence: Xiao-Hua Jin, Fax: 8610 62590843;
E-mail: xiaohuajin@ibcas.ac.cn

‡These authors contributed equally to this work.

distinguishing medicinal *Dendrobium* species. It is difficult to compare these studies because they utilized completely different sampling strategies. Additional studies, which employ well-designed sampling methods, are needed to evaluate the potential of DNA barcodes to identify Orchidaceae species.

Historically, *Holcoglossum* and related genera (subtribe Aeridinae) are highly problematic in taxonomy (Garay 1972; Dressler 1981, 1993; Christenson 1987; Kocyan *et al.* 2008). However, these problems with the taxonomy have been resolved recently by a revision of the group on the basis of herbarium and field observations (Jin 2005). *Holcoglossum* comprises approximately 12 species that range from Malaysia, Thailand, Myanmar and Vietnam to southwestern China; eight of these species are endemic to China (Jin & Wood 2009). Using molecular data, Fan *et al.* (2009) found that the 12 *Holcoglossum* species fall into three groups: (i) a tropical group of four species that occur in southeast Asia, (ii) an alpine group of five species that are found in the Hengduan Mountains of China and (iii) a biogeographically intermediate group of three species that are distributed in the region between the first two groups. The identification of species within these groups, using either morphological or molecular data, remains difficult, particularly in the case of the alpine group (Fan *et al.* 2009). Given that *Holcoglossum* species are characterized by narrow distributions, they are vulnerable to threats, such as habitat fragmentation, and many of the species are listed as endangered (Averyanov & Averyanova 2003; Jin 2005). An appropriate DNA barcode for *Holcoglossum* would be of great value for further research and conservation efforts.

In the study reported herein, we investigated six DNA regions that have been proposed previously as DNA barcodes, namely, the chloroplast genes *rbcL*, *matK*, *atpF-atpH*, *psbK-psbI* and *trnH-psbA*, and the nuclear internal transcribed spacer (ITS) region (Kress *et al.* 2005; Chase *et al.* 2007; Kress & Erickson 2007; Fazekas *et al.* 2008; Lahaye *et al.* 2008a,b; Newmaster *et al.* 2008; CBOL Plant Working Group 2009) to establish a DNA barcoding protocol to identify species of *Holcoglossum*. Our goals were (i) to test the universality of six DNA regions in *Holcoglossum*, and (ii) to evaluate the potential of these barcodes to identify species of *Holcoglossum*.

## Materials and methods

### *Taxon sampling*

Multiple samples of each species of *Holcoglossum* that is recognized in Flora of China (Jin & Wood 2009) were included in this study to ensure that both the morphological and the geographical ranges of each taxon were covered. Except for the rare species *Holcoglossum omeiense*, more than three individuals of each species were collected. In total, we sampled 52 individuals, among which were represented the 12 species of *Holcoglossum*. *Ascocentrum ampullaceum* was used as an outgroup (Fan *et al.* 2009). Detailed information about the samples is included in Table S1 (Supporting information). Newly characterized sequences were submitted to GenBank and have the Accession nos HQ404384–HQ404499, JF763774–FJ763809, JF925118 (Table S1, Supporting information).

### *DNA extraction, amplification and sequencing*

Total DNA was isolated from leaves dried in silica gel in accordance with the protocol of Doyle & Doyle (1987). Specific DNA regions were amplified via a standard polymerase chain reaction (PCR). The primer sequences used for amplification and sequencing are presented in Table S2 (Supporting information). PCR products were sequenced using an ABI 3730 DNA sequencer (Applied Biosystems). Primers were not redesigned when problems were encountered in the PCR or sequencing, because one of the major aims of the study was to evaluate the universality of specific DNA regions with respect to amplification and sequencing. The primers for the ITS were as described by Baldwin (1992); the primers for *trnH-psbA* were obtained from http://www.kew.org/barcoding/protocols.html; the primers for *matK* were as described by Cuenoud *et al.* (2002); the primers for *atpF-atpH* and *psbK-psbI* were as described by Lahaye *et al.* (2008b); and the primers for *rbcL* were as described by Gastony & Rollo (1995). The regions *atpF-atpH* and *psbK-psbI* were omitted from further analysis because they could not be sequenced successfully (see Results).

### *Data analysis*

The sequences were aligned using ClustalX (Thompson *et al.* 1997), and then adjusted manually in Bioedit v. 7 (Hall 1999). To distinguish taxa from each other, the sequence character-based method (hereafter referred to as loci examination) (Rach *et al.* 2008) was implemented in DnaSP (Rozas *et al.* 2003), and the information at each site was treated as a character.

To estimate the variation in the DNA regions that we examined, we calculated pairwise genetic p-distances for each region using MEGA v. 4 (Kumar *et al.* 2008). *RbcL* was dropped from further analysis because of its low interspecific variation (see Results). *MatK*, ITS and *trnH-psbA* were retained for multivariate classification analysis. Initially, these distances were used to evaluate interspecific and intraspecific divergence in the samples for all regions with the Kimura 2-parameter model (K2P) (Kumar *et al.* 2008).

To assess whether sequences from the proposed barcode regions formed species-specific clusters, we conducted neighbour joining (NJ) analyses in MEGA v. 4 (Kumar *et al.* 2008) and used Bayesian inference (BI) in MrBayes v. 3.1 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003). We conducted the NJ analyses with distance options set to the best-fit model of evolution. We assessed node support using the bootstrap method (Felsenstein 1988), with 1000 pseudoreplicates of NJ run with the above-mentioned distance options. For the BI analyses, each DNA region (*matK*, *trnH-psbA* and ITS) and combinations of two or three regions (*matK* + ITS, *matK* + *trnH-psbA*, ITS + *trnH-psbA*, *matK* + ITS + *trnH-psbA*) were assigned its own model of nucleotide substitution, as determined by the Akaike information criterion (AIC) in Modeltest v. 3.06 (Posada & Crandall 1998). Four chains of Markov chain Monte Carlo were run, with one tree sampled every 1000 generations for 2 000 000 generations, starting with a random tree. Majority rule consensus trees (>50%) were constructed after the samples from the burn-in period (the first 10% of sampled trees) had been removed. Neighbour joining and BI trees for *matK*, ITS and *trnH-psbA*, and for combinations of two or three loci (*matK* + ITS, *matK* + *trnH-psbA*, ITS + *trnH-psbA*, *matK* + ITS + *trnH-psbA*), were constructed with the aim of determining node support values.

To assess the potential of the barcode regions to identify species accurately, we measured the proportion of correct identifications that were achieved using TAXONDNA (Meier *et al.* 2006). The 'Best Match', 'Best Close Match' and 'All Species Barcodes' tests in TAXONDNA were run for species that were represented by more than one individual, using both individual sequences and 'Species Barcodes' (Meier *et al.* 2006), which were constructed from the consensus of all conspecific sequences. The 'All Species Barcodes' test was only performed with individual sequences for species that were represented by two or more individuals. Single DNA regions and combinations of regions were assessed by TAXONDNA.

## Results

### *Amplification, sequencing and character analysis of DNA regions*

PCR amplification was achieved with all 52 DNA samples for all loci, except *matK*, for which only 92.3% of samples could be amplified. Success rates for bidirectional sequencing were highest for *rbcL* (100%) and *trnH-psbA* (100%), followed by ITS (84%) and *matK* (~75%). All ITS sequences were obtained using the primer ITS1, and 83% of the *matK* sequences were obtained using primer 390F (Table 1). However, *atpF-atpH* and *psbK-psbI* could

**Table 1** Evaluation of six DNA markers and combinations of the markers

| | *rbcL* | *matK* | ITS | *trnH-psbA* | *atpF-atpH* | *psbK-psbI* | *matK* + ITS | *matK* + *trnH-psbA* | ITS + *trnH-psbA* | Three loci |
|---|---|---|---|---|---|---|---|---|---|---|
| Universality of primers | Yes | Yes | Yes | Yes | No | No | – | – | – | – |
| Percentage PCR success (%) | 100 | 92.31 | 100 | 100 | Low | Low | – | – | – | – |
| Percentage sequencing success (%) | 100 | 82.70 | 100 | 100 | Low | Low | – | – | – | – |
| Length of aligned sequence (bp) | 1288 | 924 | 567 | 739 | – | – | 1391 | 1663 | 1306 | – |
| No. of indels (bp) | 0 | 0 | 2 (1) | 5 (6–28) | – | – | 2 (1) | 5 (6–28) | 7 (1–28) | – |
| No. of informative sites/variable sites | 3/8 | 25/33 | 26/27 | 9/10 | – | – | 51/60 | 34/43 | 35/37 | – |
| No. of species samples (individuals) | 12 (12) | 12 (41) | 12 (52) | 12 (52) | – | – | 12 (41) | 12 (41) | 12 (52) | 12 (41) |
| Interspecific distance (mean) | 0.0019 | 0.0081 | 0.0135 | 0.0030 | – | – | – | – | – | – |
| Intraspecific distance (mean) | – | 0.0003 | 0.0003 | 0.0007 | – | – | – | – | – | – |
| Ability to discriminate (NJ) | – | 6/12 | 5/12 | 5/12 | – | – | 7/12 | 6/12 | 6/12 | 7/12 |
| Ability to discriminate (BI) | – | 5/12 | 4/12 | 5/12 | – | – | 7/12 | 5/12 | 6/12 | 7/12 |
| Ability to discriminate (loci examination) | – | 8/12 | 6/12 | 5/12 | – | – | 8/12 | 8/12 | 8/12 | 8/12 |

BI, Bayesian inference; ITS, internal transcribed spacer; NJ, neighbour joining.

not be sequenced successfully because of the presence of mononucleotide repeat regions; hence, they were omitted from subsequent analyses.

From the 12 *Holcoglossum* species, we obtained 12 *rbcL*, 41 *matK*, 52 ITS and 52 *trnH-psbA* sequences, which gave a total of 157 new sequences generated in this study. The mean interspecific p-distances were low between all species for *rbcL* (0.0019), *trnH-psbA* (0.0030), *matK* (0.0081) and ITS (0.0135) (Table 1). *RbcL* was dropped from further analysis because it had the lowest interspecific variation. The results for each of the six regions with regard to the universality of primers and success of sequencing are listed in Table 1.

## matK

For the *matK* matrix (after alignment using ClustalX and manual adjustment in Bioedit), the length of the aligned sequence was 924 bp with no indel and it contained 25 informative sites. In the *matK* data set, the interspecific genetic distance was 30 times the intraspecific distance (Table 1, Fig. 1a). Five monophyletic groups with relatively strong support values were obtained from the BI tree and six from the NJ tree (Fig. S1, Supporting information); these groups corresponded to *H. amesianum* (BI/NJ support value: 100/96), *H. flavescens* (91/65), *H. kimballianum* (−/66), *H. quasipinifolium* (99/93), *H. subulifolium* (100/99) and *H. wangii* (100/99). On the basis of the TAXONDNA analysis, *matK* showed a higher success rate for correct identification of species (Best Match: 65.85%; Best Close Match: 60.97%; All Species Barcodes: 63.41%). Judging on the basis of loci examination, six species had unique characters for *matK* that differentiated them from the other species (Table S3, Supporting information). For instance, *H. amesianum* had two unique characters (position 93: A; position 335: T). The other species that were distinguished by unique characters were *H. flavescens*, *H. kimballianum*, *H. quasipinifolium*, *H. subulifolium and H. wangii*. *Holcoglossum rupestre* and *H. weixiense* were identified on the basis of character at a combination of sites. Two pairs of species, (i) *Holcoglossum sinicum* and *H. nujiangense* and (ii) *H. lingulatum* and *H. omeiense*, shared the same characters in the *matK* matrix, respectively (Table S3, Supporting information). Failure to amplify the *matK* locus from *Holcoglossum* occurred apparently at random; hence, we suggest that better primers may be needed for this region.

## ITS

For the ITS matrix, the length of the aligned sequence was 567 bp and it contained two indels, each of 1 bp. There were 26 informative sites and 27 variable sites scattered across the matrix. In the ITS data set, the mean interspecific genetic distance was more than 40 times the mean intraspecific distance (Table 1, Fig. 1b). Four monophyletic groups with strong or moderate support values were obtained from the BI tree and five from the NJ tree; these corresponded to *H. amesianum* (100/96),

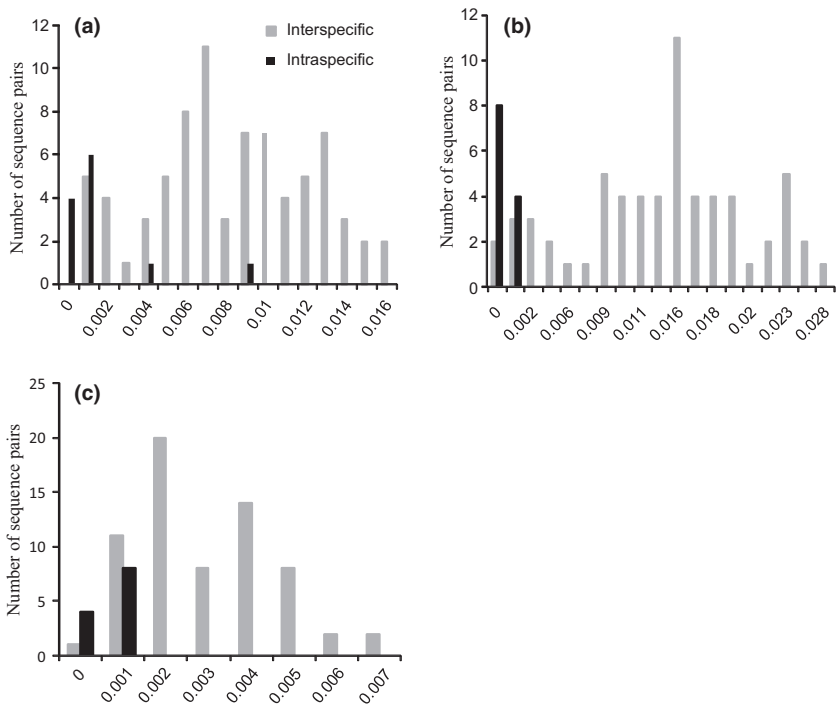


**Fig. 1** Relative distribution of interspecific and intraspecific distances: a. *matK*, b. internal transcribed spacer (ITS) and c. *trnH-psbA*.

*H. flavescens* (-/65), *H. kimballianum* (98/63), *H. quasipinifolium* (100/99) and *H. subulifolium* (84/83). On the basis of the TAXONDNA analysis, the ITS exhibited a high success rate for the correct identification of species (Best Match: 67.30%; Best Close Match: 65.38%; All Species Barcodes: 80.76%) (Table 2). Judging on the basis of loci examination, four species had unique characters that differentiated them from other *Holcoglossum* (Table S4, Supporting information). For example, *H. amesianum* had six unique characters (position 6: G; position 302: T; position 320: C; position 337: G; position 385: G; position 560: C). The three other species with unique characters were *H. kimballianum* (position 417: G), *H. quasipinifolium* (position 58: G; position 309: C; position 503: C) and *H. subulifolium* (position 307: G). *Holcoglossum wangii* and *H. flavescens* had three and four combined sites that discriminated them from the other species, respectively. Two groups of species, namely (i) *H. lingulatum* and *H. omeiense*, (ii) *H. nujiangense, H. weixiense, H. flavescens* and *H. sinicum*, shared the same character states, respectively (Fig. S2, Table S4, Supporting information).

## trnH-psbA

For the *trnH-psbA* matrix, the length of the aligned sequence was 739 bp with five indels that were 6–28 bp long and it contained nine informative sites. For *trnH-psbA*, the distribution of interspecific and intraspecific distances is shown in Fig. 1c. The mean interspecific p-distance was more than four times the intraspecific distance (Table 1). Only three monophyletic groups with moderate support values were obtained from the NJ and BI analyses (Fig. S3, Supporting information); the groups corresponded to *H. amesianum* (63/55), *H. quasipinifolium* (80/56) and *H. wangii* (92/66).The TAXONDNA analysis showed that *trnH-psbA* had the lowest success rate for identifying species among the loci analysed (Best Match: 48.07%; Best Close Match: 44.23%; All Species Barcodes: 44.23%) (Table 2). Loci examination showed that four species had unique *trnH-psbA* character states (Table S5, Supporting information). *Holcoglossum quasipinifolium* displayed the unique character G at site 407, and *H. wangii* had the unique character T at site 356. *Holcoglossum rupestre* had a unique long indel (11 bp) from site 233 to 243, and *H. amesianum* also had a unique long indel (16 bp) from site 307 to 322. *Holcoglossum kimballianum* was discriminated from the other species by a combination of sites (Table S5, Supporting information). In addition, two groups of species, namely (i) *H. subulifolium, H. lingulatum* and *H. omeiense*, (ii) *H. nujiangense, H. weixiense, H. flavescens* and *H. sinicum*, shared the same character states, respectively (Table S5, Supporting information).

**Table 2** Ability of DNA barcode regions to discriminate species as assessed using TAXONDNA*

| Region | Best match | | | Best close match | | | All species barcodes† | | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct (%) | Ambiguous (%) | Incorrect (%) | Correct (%) | Ambiguous (%) | Incorrect (%) | Correct (%) | Ambiguous (%) | Incorrect (%) |
| *matK* | 65.85 | 31.70 | 2.43 | 60.97 | 24.39 | 2.43 | 63.41 | 21.95 | 2.43 |
| ITS | 67.30 | 30.76 | 1.92 | 65.38 | 28.84 | 1.92 | 80.76 | 11.53 | 3.84 |
| *trnH-psbA* | 48.07 | 48.07 | 3.84 | 44.23 | 46.15 | 3.84 | 44.23 | 48.07 | 1.92 |
| *matK* + ITS | 63.41 | 34.14 | 2.43 | 56.09 | 19.51 | 2.43 | 58.53 | 17.07 | 2.43 |
| *matK* + *trnH-psbA* | 73.07 | 21.15 | 5.76 | 59.61 | 15.38 | 5.76 | 53.84 | 25.00 | 1.92 |
| ITS + *trnH-psbA* | 70.73 | 24.39 | 4.87 | 60.97 | 17.07 | 4.87 | 56.09 | 24.39 | 2.43 |
| *matK* + ITS + *trnH-psbA* | 70.73 | 19.51 | 9.75 | 51.21 | 7.31 | 9.75 | 46.34 | 19.51 | 2.43 |

ITS, internal transcribed spacer.
*Percentage identification success for each of the three tests is given for individual sequences.
†The 'All Species Barcodes' test with individual sequences was only performed for species represented by >2 individuals.

### *Combinations of DNA regions*

The results for the combinations of DNA regions, namely *matK* + ITS, *matK* + *trnH-psbA*, ITS + *trnH-psbA* and *matK* + ITS + *trnH-psbA*, are listed in Table 1. All four combinations identified the same six species that had been identified as monophyletic groups in this study, namely *H. amesianum*, *H. flavescens*, *H. kimballianum*, *H. quasipinifolium*, *H. subulifolium* and *H. wangii* (Fig. 2, Fig. S4-S6, Supporting information).

The combination of *matK* + ITS provided the highest resolving power. Seven species showed strong or moderate support values in the NJ and BI analyses (Fig. 2). The TAXONDNA analysis showed that *matK* + ITS had a high success rate (Best Match: 63.41%; Best Close Match: 56.09%; All Species Barcodes: 58.53%) (Table 2). Loci examination showed six of the species could be discriminated by both *matK* and ITS (Tables S3 and S4, Supporting information). *Holcoglossum rupestre* and *H. weixiense* shared the same characters in the ITS data, but the two species could be differentiated from each other using the *matK* data (Tables S3 and S4, Supporting information).

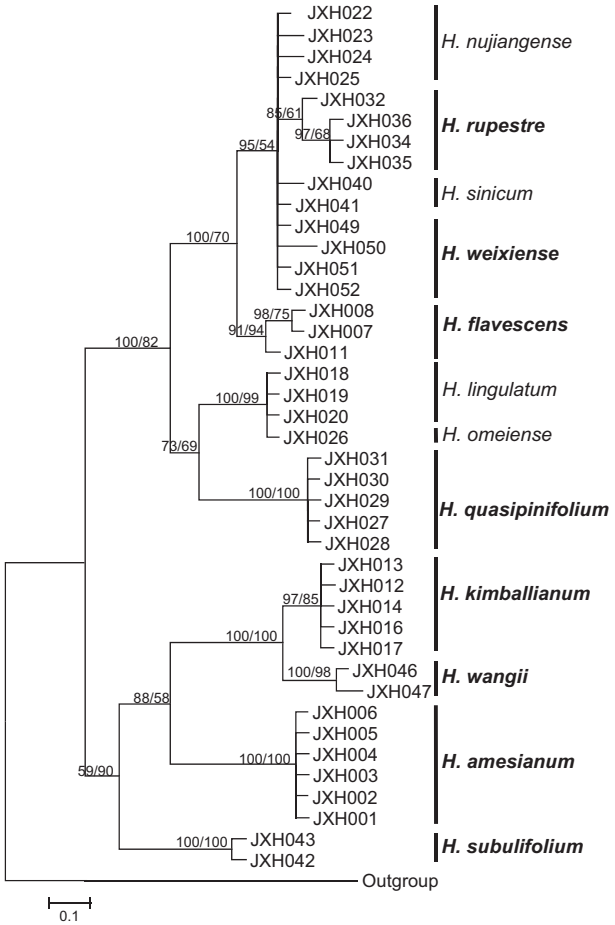


**Fig. 2** Bayesian inference (BI) tree based on the combined *matK* + internal transcribed spacer (ITS) matrix for 12 *Holcoglossum* species. Numbers on branches represent BI and NJ support values, respectively. Every individual is shown with its DNA number. The species in bold could be identified on the basis of loci examination.

For the *matK* + *trnH-psbA* matrix, six species showed strong or moderate support values in the NJ and BI trees (Fig. S4, Supporting information). The TAXONDNA analysis showed that *matK* + *trnH-psbA* had a high success rate (Best Match: 73.07%; Best Close Match: 59.61%; All Species Barcodes: 53.83%) (Table 2). Eight species could be discriminated using loci examination. Five of these species could be differentiated by both *matK* and *trnH-psbA* (Table S3 and S5, Supporting information). *Holcoglossum flavescens*, *H. subulifolium* and *H. weixiense* shared the same characters in the *trnH-psbA* sequence alignment, but the three species could be differentiated from each other using the *matK* data (Tables S3 and S5, Supporting information).

For the ITS + *trnH-psbA* matrix, six species could be discriminated with strong or moderate support values on the basis of the NJ and BI trees (Fig. S5, Supporting information). The TAXONDNA analysis showed that ITS + *trnH-psbA* had a high success rate (Best Match: 70.73%; Best Close Match: 60.97%; All Species Barcodes: 56.09%) (Table 2). The ITS and *trnH-psbA* sequences complemented each other: *H. rupestre*, *H. nujiangense*, *H. sinicum* and *H. weixiense* shared the same DNA characters in the ITS sequence matrix, but in the *trnH-psbA* sequence matrix, *H. rupestre* had an indel of 11 bp, which enabled it to be discriminated. *Holcoglossum subulifolium*, *H. omeiense* and *H. lingulatum* shared the same character states in the *trnH-psbA* data set (Table S4 and S5, Supporting information), but *H. subulifolium* differed from the other two species with respect to the ITS sequence (position 307: G) (Table S4, Supporting information).

When *matK*, ITS and *trnH-psbA* were combined, the NJ and BI trees distinguished seven monophyletic species with strong or moderate support values, and eight species could be discriminated using loci examination (Table 1, Fig. S6, Supporting information). The TAXONDNA analysis showed that *matK* + ITS + *trnH-psbA* had a high success rate (Best Match: 70.73%; Best Close Match: 51.21%; All Species Barcodes: 46.34%) (Table 2).

## Discussion

### *Fitness of DNA markers and species identification*

We assessed the potential of six proposed DNA regions as barcodes for orchid species in terms of three main criteria: (i) universality of amplification and sequencing, (ii) pattern of intraspecific vs. interspecific variation and

(iii) power to identify species (http://www.Barcoding.si.edu/PDF/). However, the six DNA barcode candidates were unable to identify fully the 12 species of *Holcoglossum* on the basis of the analysis of either single regions or combinations of regions.

Difficulties in sequencing suggested that the *atpF-atpH* and *psbK-psbI* regions are not practical barcodes for identifying *Holcoglossum* or other land plants, even though these regions have been suggested as potential DNA barcodes for plants (Lahaye *et al.* 2008b). Our results showed that *rbcL* had the lowest interspecific variation (0.0019) among the four remaining DNA markers, and it was not suitable for the DNA barcoding of *Holcoglossum* species. Three of the markers that we tested (*matK*, ITS and *trnH-psbA*) warranted further consideration as barcodes for *Holcoglossum* because of their easy amplification, sequencing and alignment, and higher interspecific variation.

In the study, the coding gene *matK* was easy to align and had a high interspecific p-distance, as has been shown in other land plants (Kress *et al.* 2005; Newmaster *et al.* 2006, 2008; Kress & Erickson 2007; Kelly *et al.* 2010). Among the single region barcodes, *matK* exhibited the highest ability to identify species on the basis of loci examination (8/12), and TAXONDNA analysis also showed that its discriminatory ability was high; however, the tree methods indicated a lower discriminatory ability (NJ tree: 6/12; BI tree: 5/12). Lahaye *et al.* (2008a) analysed 1084 plant species (nearly 96% of which were orchid species) and showed that a portion of the plastid *matK* gene could be a universal DNA barcode for flowering plants. Other authors have proposed *matK* as a core universal plant barcode (Newmaster *et al.* 2006; Kress & Erickson 2007). Our results indicated that, although amplification and sequencing of the *matK* region might not be entirely straightforward, the higher ability of *matK* to discriminate between the species means that this region has a greater potential as a barcode for *Holcoglossum* species than ITS or *trnH-psbA*.

Bidirectional sequences for the nuclear ITS region could not be obtained reliably, but accurate unidirectional sequences (mainly with primer ITS1) were obtained for all samples in our study. High variability of the ITS region has also been observed in *Alnus* (Ren *et al.* 2010) and Euphorbiaceae (Pang *et al.* 2010). The ITS showed the highest rate of species identification on the basis of TAXONDNA analysis, but displayed a more limited ability to identify *Holcoglossum* species on the basis of loci examination and tree methods. Six of the 12 *Holcoglossum* species could be discriminated from the ITS data on the basis of loci examination. In addition, for the NJ and BI analyses of the ITS data, five species of *Holcoglossum* had well-supported values (Fig. S3, Supporting information Table 1). ITS2 (part of the ITS) has shown a high discriminatory ability at species level and has been suggested previously as a universal DNA barcode for plants and animals (Yao *et al.* 2009; Chen *et al.* 2010). Pang *et al.* (2010) collected 1183 plant samples from 871 species in 66 diverse genera of Euphorbiaceae, and DNA barcoding with ITS/ITS2 identified >90% and 100% successfully to species and genus levels, respectively. However, in our study, ITS alone had a low ability to discriminate between *Holcoglossum* species.

Amplification and sequencing of *trnH-psbA* were straightforward across *Holcoglossum* species, as in other plants (Kress *et al.* 2005; Newmaster *et al.* 2008; Ren *et al.* 2010). To date, *trnH-psbA* has been evaluated as a potential barcode region because of its universality and ease of amplification and alignment at the generic level (Kress *et al.* 2005; Lahaye *et al.* 2008a; Newmaster *et al.* 2008; Ren *et al.* 2010). The *trnH-psbA* region might be a promising locus for DNA barcoding on the basis of studies on liquorice (Kondo *et al.* 2007), *Compsoneura* (Newmaster *et al.* 2008), orchids (Lahaye *et al.* 2008a), filmy ferns (Nitta 2008), *Dendrobium* (Yao *et al.* 2009) and *Alnus* (Ren *et al.* 2010). In our study, *trnH-psbA* had low interspecific variability and was able to discriminate <50% of *Holcoglossum* species (Table 1). As a consequence, we suggest that this region alone would not be an appropriate DNA barcoding marker for *Holcoglossum*.

The NJ and BI tree methods indicated that the combination of *matK* with ITS or *trnH-psbA* could increase the rate of species identification significantly (33% vs. 58% or 50%, respectively). However, the combination of all three DNA regions did not increase the percentage success of species identification. The combination of *matK* + ITS showed a moderate success rate for species identification on the basis of the TAXONDNA analysis (Table 2). With respect to loci examination, *matK* as a single region barcode identified the largest number of species (Table 1), and the rate of species identification was not improved using combinations of DNA regions (*matK* + ITS, *matK* + *trnH-psbA*, ITS + *trnH-psbA* or *matK* + ITS + *trnH-psbA*). The lack of improvement in the rate of species identification when using combinations of DNA regions has also been reported by Lahaye *et al.* (2008a,b) and Starr *et al.* (2009).

Molecular phylogeny of *Holcoglossum* has shown that the genus contains three clades: the tropical clade (four species), the alpine temperate clade (five species) and a third clade that is distributed between the regions occupied by the first two groups (three species) (Fan *et al.* 2009). The four species of the tropical clade, which are well isolated from each other on the basis of floral and molecular evidence, were also identified readily by DNA barcoding with the single or combined DNA regions. *Holcoglossum quasipinifolium*, which is in the third intermediate clade, could also be discriminated on the basis of

loci examination and tree analyses, whereas the sister groups (*H. lingulatum* and *H. omeiense*) could not. In the alpine temperate clade, *matK* could discriminate three species on the basis of loci examination, but even in combination with ITS and/or *trnH-psbA* the remaining two species could not be distinguished. These results suggest that the radiated *Holcoglossum* species are difficult to identify using the single or combined DNA regions analysed in this study.

### *Suggestions for the DNA barcoding of radiated species*

The identification of plant species using DNA barcoding remains a challenge. Recently, there has been intense debate with respect to the identification of the optimal DNA barcode for land plants (Sass *et al.* 2007; Edwards *et al.* 2008; Farrington *et al.* 2009; Kress *et al.* 2009; Seberg & Petersen 2009; Pang *et al.* 2010). However, very few studies have focused on the discrimination of species that have emerged as a result of radiation, even though this is one of the most practical applications for DNA barcodes. Our results indicated that the *matK* locus had more variation and greater power to discriminate between *Holcoglossum* species than the other five DNA regions that we tested. To identify all species, other DNA regions are required. In recent studies, the chloroplast gene *trnL-F* was shown to be easy to amplify and sequence and displayed sufficient variation and a high rate of discrimination (e.g. Edwards *et al.* 2008; Kocyan *et al.* 2008). As a result, it was proposed as another promising barcode (Farrington *et al.* 2009). DNA barcodes that incorporate *matK* in combination with other highly variable regions, such as *trnL-F*, might have the potential to facilitate the investigation and preservation of radiated orchid species.

## Acknowledgements

The authors thank Ji-Hong Li, Wei-Tao Jin and Hai-Lang Zhou (Institute of Botany, Chinese Academy of Sciences) for revising the manuscript. This research was supported by the State Key Laboratory of Systematic and Evolutionary Botany, the National Natural Science Foundation of China (Grant No. 30600037), the large-scale Scientific Facilities of the Chinese Academy of Sciences (2009-LSF-GBOWS-01) and the Research Fund of the Chinese Academy of Sciences (13214G1014).

## References

Atwood JT (1986) The size of the Orchidaceae and the systematic distribution of epiphytic orchids. *Selbyana*, **9**, 171–186.

Averyanov LV, Averyanova AL (2003) *Updated Checklist of the Orchids of Vietnam*. Vietnam National University Publishing House, Hanoi.

Baldwin BG (1992) Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositae. *Molecular Phylogenetics and Evolution*, **1**, 3–16.

Cameron KM (2004) Utility of plastid *psaB* gene sequences for investigating intrafamilial relationships within Orchidaceae. *Molecular Phylogenetics and Evolution*, **31**, 1157–1180.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA*, **106**, 12794–12797.

Chase MW, Cowan RS, Hollingsworth PM *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Chen SL, Yao H, Han JP *et al.* (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE*, **5**, e8613.

Christenson EA (1987) An infrageneric classification of *Holcoglossum* Schltr. *(Orchidaceae: Sarcanthinae) with a key to the genera of the Aerides-Vanda alliance. Notes in Royal Botanical Garden Edinburgh*, **44**, 249–256.

Cuenoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW (2002) Molecular phylogenetics of the Caryophylles based on 18SrDNA, *rbcL, atpB* and *matK* sequences. *American Journal of Botany*, **89**, 132–144.

Dodson CH (1962) The importance of pollination in the evolution of the orchids of tropical America. *American Orchid Society Bulletin*, **31**, 525–534.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, **19**, 11–15.

Dressler RL (1981) *The Orchids: Natural History and Classification*. Harvard University Press, Cambridge, Massachusetts, USA.

Dressler RL (1993) *Phylogeny and Classification of the Orchid Family*. Cambridge University Press, Cambridge.

Dressler RL, Dodson CH (1960) Classification and Phylogeny in the Orchidaceae. *Annals of the Missouri Botanical Garden*, **47**, 25–68.

Edwards D, Horn A, Taylor D, Savolainen V, Hawkins JA (2008) DNA barcoding of a large genus, *Aspalathus* L.(Fabaceae). *Taxon*, **57**, 1317–1314.

Fan J, Qin HN, Li DZ, Jin XH (2009) Molecular phylogeny and biogeography of *Holcoglossum* (Orchidaceae: Aeridinae) based on nuclear ITS, and chloroplast *trnL-F* and *matK*. *Taxon*, **58**, 849–861.

Farrington L, MacGillivray P, Faast R, Austin A (2009) Investigating DNA barcoding options for the identification of Caladenia (Orchidaceae) species. *Australian Journal of Botany*, **57**, 276–286.

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG *et al.* (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, **3**, e2802.

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics*, **22**, 521–565.

Garay LA (1972) On the systematics of the monopodial orchids I. *Botanical Museum Leaflets*, **23**, 149–212.

Gastony GJ, Rollo DR (1995) Phylogeny and generic circumscriptions of cheilanthoid ferns (Pteridaceae: Cheilanthoideae) inferred from *rbcL* nucleotide sequences. *American Fern Journal*, **85**, 341–360.

Gravendeel B, Chase MW, de Vogel EF, Roos MC, Mes THM, Bachmann K (2001) Molecular phylogeny of *Coelogyne* (Epidendroideae; Orchidaceae) based on plastid RFLPs, *matK*, and nuclear ribosomal ITS sequences: evidence for polyphyly. *American Journal of Botany*, **88**, 1915–1927.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleus Acids Symp Series*, **41**, 95–98.

Hebert PDN, Cywinska A, Ball SL, Dewaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **270**, 313–322.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences, USA*, **101**, 14812–14817.

Hogg ID, Hebert PDN (2004) Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Canadian Journal of Zoology*, **82**, 749–754.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Jin XH (2005) Generic delimitation and a new infrageneric system in the genus *Holcoglossum* (Orchidaceae: Aeridinae). *Botanical Journal of the Linnean Society*, **149**, 465–468.

Jin XH, Wood JJ (2009) *Holcoglossum*. In: *Flora of China*, Vol. 25. (eds Wu ZY & Raven PH), pp. 499–502. Science Press, Beijing, Missouri Botanical Garden, St. Louis.

Kelly LJ, Ameka GK, Chase MW (2010) DNA barcoding of African Podostemaceae (river-weeds): a test of proposed barcode regions. *Taxon*, **59**, 251–260.

Kocyan A, Vogel EF, Conti E, Gravendeel B (2008) Molecular phylogeny of *Aerides* (Orchidaceae) based on nuclear and two plastid markers: a step forward in understanding the evolution of the Aerdinae. *Molecular Phylogenetics and Evolution*, **48**, 422–443.

Koehler S, Williams NH, Whitten WM, Maria do Carmo E (2002) Phylogeny of the *Bifrenaria* (Orchidaceae) complex based on morphology and sequence data from nuclear rDNA internal transcribed spacers (ITS) and chloroplast *trnL\trnF* region. *International Journal of Plant Sciences*, **163**, 1055–1066.

Kondo K, Shiba M, Yamaji H *et al.* (2007) Species identification of licorice using nrDNA and cpDNA genetic markers. *Biological & Pharmaceutical Bulletin*, **30**, 1497–1502.

Kress W, Erickson D (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS ONE*, **2**, e508.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences, USA*, **102**, 8369–8374.

Kress WJ, Erickson DL, Jones FA *et al.* (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences, USA*, **106**, 18621–18626.

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, **9**, 299–306.

Lahaye RRY, van der Bank M, Bogarin D *et al.* (2008a) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences, USA*, **105**, 2923–2928.

Lahaye RRY, Savolainen V, Duthoit S, Maurin O, van der Bank M (2008b) A test of *psbK-psbI* and *atpF-atpH* as potential plant DNA barcodes using the flora of the Kruger National Park (South Africa) as a model system. *Nature Precedings* http://hdl.handle.net/10101/npre. 12008.11896.10101.

Liu J, Möller M, Gao LM, Zhang DQ, Li DZ (2010) DNA barcoding for the discrimination of Eurasian yews (*Taxus* L., Taxaceae) and the discovery of cryptic species. *Molecular Ecology Resources*, **11**, 89–100.

Mabberley JD (2008) *Mabberley's Plant-Book. A Portable Dictionary of Plants, their Classification and Uses*, 3rd edn. Cambridge University Press, Cambridge.

Meier R, Kwong S, Vaidya G, Ng PKL (2006) DNA Barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

Newmaster SG, Ragupathy S (2009) Testing plant barcoding in a sister species complex of pantropical *Acacia* (Mimosoideae, Fabaceae). *Molecular Ecology Resources*, **9**, 172–180.

Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Botany*, **84**, 335–341.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2008) Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources*, **8**, 480–490.

Nitta JH (2008) Exploring the utility of three plastid loci for biocoding the filmy ferns (Hymenophyllaceae) of Moorea. *Taxon*, **57**, 725–722.

Pang XH, Song JY, Zhu YJ, Xie CX, Chen SL (2010) Using DNA barcoding to identify species within Euphorbiaceae. *Planta Medica*, **76**, doi: 10.1055/s-0030-1249806.

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Pridgeon AM, Solano R, Chase MW (2001) Phylogenetic relationships in Pleurothallidinae (Orchidaceae): combined evidence from nuclear and plastid DNA sequences. *American Journal of Botany*, **88**, 2286–2308.

Pryer KM, Schuettpelz E, Huiet L *et al.* (2010) DNA barcoding exposes a case of mistaken identity in the fern horticultural trade. *Molecular Ecology Resources*, **10**, 979–985.

Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **275**, 237–247.

Ren BQ, Xiang XG, Chen ZD (2010) Species identification of *Alnus* (Betulaceae) using nrDNA and cpDNA genetic markers. *Molecular Ecology Resources*, **10**, 594–605.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Salazar GA, Chase MW, Soto Arenas MA, Ingrouille M (2003) Phylogenetics of Cranichideae with emphasis on Spiranthinae (Orchidaceae, Orchidoideae): evidence from plastid and nuclear DNA sequences. *American Journal of Botany*, **90**, 777–795.

Sass C, Little DP, Stevenson DW, Specht CD (2007) DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. *PLoS ONE*, **2**, e1154.

Seberg O, Petersen G (2009) How many loci does it take to DNA barcode a *Crocus*? *PLoS ONE*, **4**, e4598.

Starr JR, Naczi RFC, Chouinard BN (2009) Plant DNA barcodes and species resolution in sedges (*Carex*, Cyperaceae). *Molecular Ecology Resources*, **9**, 151–163.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, **25**, 4876–4882.

Van de Wiel CC, van der Schoot J, van Valkenburg JLCH, Duistermaat H, Smulders MJM (2009) DNA barcoding discriminates the noxious invasive plant species, floating pennywort (*Hydrocotyle ranunculoides* L.f.), from non-invasive relatives. *Molecular Ecology Resources*, **9**, 1086–1091.

Van den Berg C, Higgins WE, Dressler RL *et al.* (2000) A phylogenetic analysis of Laeliinae (Orchidaceae) based on sequence data from internal transcribed spacers (ITS) of nuclear ribosomal DNA. *Lindleyana*, **15**, 96–114.

Yao H, Song JY, Ma XY *et al.* (2009) Identification of *Dendrobium* species by a candidate DNA barcode sequence: the chloroplast *psbA-trnH* intergenic region. *Planta Medica*, **75**, 667–669.

## Data Accessibility

DNA sequences: GenBank Accessions HQ404384–HQ404499, JF763774–FJ763809, JF925118.

## Supporting Information

Additional supporting information may be found in the online version of this article.

**Table S1** Samples and voucher information for the *Holcoglossum* species used in this study.

**Table S2** PCR and sequencing primers used in this study.

**Table S3** Character-based DNA database for *Holcoglossum* species from the *matK* region.

**Table S4** Character-based DNA database for *Holcoglossum* species from the ITS region.

**Table S5** Character-based DNA database for *Holcoglossum* species from the *trnH-psbA* region.

**Fig. S1** Bayesian inference tree (BI, left) and neighbour joining tree (NJ, right) based on the *matK* matrix for 12 *Holcoglossum* species. Numbers on branches represent BI and NJ support values, respectively. Every individual is shown with its DNA number. White circles indicate that a single diagnostic site was used; black ones mean that a combination of candidate sites was used; stars and asterisks represent those species that had the same sequences. The species in bold was identified only through the NJ tree.

**Fig. S2** Bayesian inference tree (BI, left) and neighbour joining tree (NJ, right) based on the ITS matrix for 12 *Holcoglossum* species. Numbers on branches represent BI and NJ support values, respectively. Every individual is shown with its DNA number. White circles indicate that a single diagnostic site was used; black ones mean that a combination of candidate sites was used; stars and asterisks represent those species that had the same sequences. The species in bold was identified only on the NJ tree.

**Fig. S3** Bayesian inference tree (BI) based on the *trnH-psbA* matrix for 12 *Holcoglossum* species. Numbers on branches represent BI and NJ support values, respectively. Every individual is shown with its DNA number. White circles indicate that a single diagnostic site was used; black ones mean that a combination of candidate sites was used; stars and asterisks represent those species that had the same sequences.

**Fig. S4** Bayesian inference tree (BI, left) and neighbour joining tree (NJ, right) based on the *matK* + *trnH-psbA* combined matrix for 12 *Holcoglossum* species. Numbers on branches represent BI and NJ support values, respectively. Every individual is shown with its DNA number. The species in bold could be identified on the basis of loci examination.

**Fig. S5** Bayesian inference tree (BI) based on the ITS + *trnH-psbA* combined matrix for 12 *Holcoglossum* species. Numbers on branches represent BI and NJ support values, respectively. Every individual is shown with its DNA number. The species in bold could be identified on the basis of loci examination.

**Fig. S6** Bayesian inference tree (BI) based on the *matK* + *trnH-psbA* + ITS combined matrix for 12 *Holcoglossum* species. Numbers on branches represent BI and NJ support values, respectively. Every individual is shown with its DNA number. The species in bold could be identified on the basis of loci examination.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.