J Ind Microbiol Biotechnol (2013) 40:365–378
DOI 10.1007/s10295-012-1228-4

GENETICS AND MOLECULAR BIOLOGY OF INDUSTRIAL ORGANISMS

On the reliability of DNA sequences of *Ophiocordyceps sinensis* in public databases

Shu Zhang · Yong-Jie Zhang · Xing-Zhong Liu · Hong Zhang · Dian-Sheng Liu

Received: 18 October 2012 / Accepted: 24 December 2012 / Published online: 9 February 2013

Abstract Some DNA sequences in the International Nucleotide Sequence Databases (INSD) are erroneously annotated, which has lead to misleading conclusions in publications. *Ophiocordyceps sinensis* (syn. *Cordyceps sinensis*) is a fungus endemic to the Tibetan Plateau, and more than 100 populations covering almost its distribution area have been examined by us over recent years. In this study, using the data from authentic materials, we have evaluated the reliability of nucleotide sequences annotated as *O. sinensis* in the INSD. As of October 15, 2012, the INSD contained 874 records annotated as *O. sinensis*, including 555 records representing nuclear ribosomal DNA (63.5 %), 197 representing protein-coding genes (22.5 %), 92 representing random markers with unknown functions (10.5 %), and 30 representing microsatellite loci (3.5 %). Our analysis indicated that 39 of the 397 internal transcribed spacer entries, 27 of the 105 small subunit entries, and five of the 53 large subunit entries were incorrectly annotated as belonging to *O. sinensis*. For protein-coding sequences, all records of serine protease genes, the mating-type gene *MAT1-2-1*, the DNA lyase gene, the two largest subunits of RNA polymerase II, and elongation factor-1α gene were correct, while 14 of the 73 β-tubulin entries were indeterminate. Genetic diversity analyses using those sequences correctly identified as *O. sinensis* revealed significant genetic differentiation in the fungus although the extent of genetic differentiation varied with the gene. The relationship between *O. sinensis* and some other related fungal taxa is also discussed.

Keywords *Ophiocordyceps sinensis* · Sequence reliability · Public nucleotide database · Nuclear ribosomal DNA · Protein-coding gene · Genetic diversity · Cryptic species · Synonym

Abbreviations

INSD	International Nucleotide Sequence Databases
nrDNA	Nuclear ribosomal DNA
ITS	Internal transcribed spacer
nrLSU	Nuclear ribosomal large subunit DNA
nrSSU	Nuclear ribosomal small subunit DNA
K2P	Kimura 2-parameter
rpb1	The largest subunit of RNA polymerase II
rpb2	The second largest subunit of RNA polymerase II
tef1	Elongation factor-1α gene
tub	β-tubulin gene
OSRC	*O. sinensis* random clone

Shu Zhang and Yong-Jie Zhang contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10295-012-1228-4) contains supplementary material, which is available to authorized users.

S. Zhang · D.-S. Liu
Institute of Applied Chemistry, Shanxi University,
Taiyuan 030006, China

S. Zhang · X.-Z. Liu (✉)
State Key Laboratory of Mycology, Institute of Microbiology,
Chinese Academy of Sciences, NO. 3, 1st West Beichen Road,
Chaoyang District, Beijing 100101, China
e-mail: liuxz@im.ac.cn

Y.-J. Zhang
School of Life Sciences, Shanxi University,
Taiyuan 030006, China

H. Zhang
Shanxi Academy of Analytical Science, Taiyuan 030006, China

Introduction

International Nucleotide Sequence Databases (INSD) such as GenBank, EMBL, and DDBJ are critical resources for molecular biology, evolutionary biology, and ecology. There has been dramatic increase in nucleotide records in these databases [1, 2], in part because most scientific journals require that relevant sequences be submitted to one of the databases before manuscript submission. Quality control of the raw data, however, often depends solely on the submitters. When reading scientific journals, one often finds papers with erroneous conclusions based on DNA sequence data obtained from organisms that were misidentified. Evaluations of fungal nuclear ribosomal DNA (nrDNA) sequences indicated that ~20 % of the sequences in public databases may be unreliable [5, 33]. Annotation errors of INSD sequences were also reported in many other groups of organisms [21, 28–30, 37]. The most common causes of errors include misidentification or mislabeling of original materials, contamination by other organisms, or technical faults (e.g., PCR-generated chimeric sequences) [44]. Error in nucleotide records is a serious problem that threatens the utility of the sequence databases.

The current report concerns the reliability of DNA sequences reported for the fungus *Ophiocordyceps sinensis* (syn. *Cordyceps sinensis*). *O. sinensis* parasitizes and mummifies underground caterpillars within the family Hepialidae, and both the fungus and its host insects are endemic to alpine regions on the Tibetan Plateau. The joint fungus–insect structure resulting from fungal parasitism of insect larvae has been termed the “natural *O. sinensis* specimen” and is commonly called “dong chong xia cao (冬虫夏草)” in Chinese; the recommended term in English is currently “Chinese cordyceps” [61]. Chinese cordyceps has been widely used in traditional Chinese medicine for the treatment of asthma, bronchial and lung inflammations, and other diseases [10, 68]. Because the demand for Chinese cordyceps has increased but its distribution in nature is limited and large-scale cultivation of its sexual fruiting body has not been successful, the harvest of Chinese cordyceps in recent years has been heavy and is depleting this natural resource [59]. This is unfortunate because *O. sinensis* acts as a flagship species for its ecosystem and has been nominated as the national fungus of China [6, 61]. *O. sinensis* has also been listed as an “endangered species for protection” in China [61].

Because *O. sinensis* is important ecologically, economically, and medicinally [61], recent research has been conducted on its morphology [50, 55], anamorph determination [18, 48], mycelial fermentation [12], sexual reproduction [58], genetic differentiation [11, 23, 63], rapid detection [19, 66], pharmacology [10], chemical components [14, 20], investigation of natural resources [22, 51], artificial cultivation [53], associated microbial community [65], and insect hosts [47]. As a result, many sequences annotated as *O. sinensis* are available in the INSD, but some records contain random as well as systematic errors [41]. During literature searches, we often came across papers with unexpected results. For example, some papers reported that *O. sinensis* was obtained from soils other than the Tibetan Plateau [3, 8, 31], from foxing spots on old paper artifacts [36], and from plants [25, 39]. Because *O. sinensis* is a specialized parasite of insects, we suspect that the fungi in these reports were misidentified. In addition, some sequences were obtained from Chinese cordyceps or from fungal isolates recovered from Chinese cordyceps, but they actually represent other fungi that were associated with the parasitized insects [13, 17]. The existence of significant genetic variation among *O. sinensis* isolates [63] makes the situation even more complex. Therefore, it is vital to verify the reliability of all sequences annotated as *O. sinensis* in the INSD.

Over recent years, we have collected and evaluated the DNA sequences from more than 100 populations of *O. sinensis*, covering almost its entire distribution. These data represent a useful basis for evaluating the reliability of DNA sequences annotated as *O. sinensis*. The four aims of the work are: to summarize sequences annotated as *O. sinensis* in the INSD as of October 15, 2012; to determine the reliability of these INSD sequences; to explore the intraspecific genetic variations of *O. sinensis* based on correct INSD sequences; and finally to elucidate the relationship between *O. sinensis* and some other related fungal taxa.

Materials and methods

Sequence download and arrangement

Nucleotide sequences of *O. sinensis* in the INSD were obtained by searching with the keyword “*O. sinensis*” in GenBank (http://www.ncbi.nlm.nih.gov/). These sequences were recorded according to type of genes, submitter, country affiliation of submitters, submission year, publication status, etc. All sequences representing the same gene were compiled as an individual file in fasta format.

Amplification of *O. sinensis* genes from authentic cultures

Five Chinese cordyceps samples were collected from different, widely separated regions on the Tibetan Plateau, and one isolate tentatively identified as *O. sinensis* was obtained from each sample (Table 1). Based on morphological analysis and sequence analysis of the internal

Table 1 *Ophiocordyceps sinensis* isolates used in this study and their GenBank accession numbers for various gene fragments

Strain	Origin	Latitude (north)	Longitude (east)	GenBank accession no.						
				nrDNA ITS	nrSSU	nrLSU	*tub*	*tef1*	*rpb1*	*rpb2*
QH09-201	Gonghe, Hainan, Qinghai, China	36.43	99.47	JQ325080	JX968024	JX968029	JX968019	JX968014	JX968004	JX968009
QH06-197	Yushu, Yushu, Qinghai, China	32.58	96.51	FJ654228[a]	JX968025	JX968030	JX968020	JX968015	JX968005	JX968010
XZ06-44	Gyalsa, Lhoka, Tibet, China	29.29	92.72	FJ654218[a]	JX968026	JX968031	JX968021	JX968016	JX968006	JX968011
YN07-8	Deqin, Diqing, Yunnan, China	28.34	99.07	FJ654237[a]	JX968027	JX968032	JX968022	JX968017	JX968007	JX968012
YN09-64	Lanping, Nujiang, Yunnan, China	26.37	99.43	JQ325141	JX968028	JX968033	JX968023	JX968018	JX968008	JX968013

[a] These accession numbers are sequences from Zhang et al. [63]. Other sequences were reported in this study

transcribed spacer (ITS) region of nrDNA, the identity of these isolates as *O. sinensis* was confirmed. These five isolates were cultivated as described previously [62], and fresh mycelia cultivated on cellophane papers were used for genomic DNA extraction with the cetyltrimethylammonium bromide-based approach [64].

DNA fragments of nrDNA ITS, the large and small subunits (nrLSU and nrSSU) of nrDNA, two largest subunits of RNA polymerase II (*rpb1* and *rpb2*), elongation factor-1α (*tef1*), and β-tubulin (*tub*) were amplified from the five *O. sinensis* isolates (Table 1). Primers and annealing temperatures used in this study are listed in Table 2. PCR was performed in a $\mathrm{T}_{\mathrm{GRADIENT}}$ thermocycler (Biometra, Göttingen, Germany). After confirmation of the PCR products by agarose gel electrophoresis, the fragments were cleaned with the 3S Spin PCR Product Purification Kit (Biocolor Bioscience & Technology Company, China). The purified PCR products of nrDNA ITS, nrSSU, nrLSU, and *tub* were directly sequenced with PCR primers using an ABI 3730 XL DNA sequencer with BigDye 3.1 Terminators (Applied Biosystems, Carlsbad, CA, USA). PCR products of *rpb1*, *rpb2*, and *tef1* were inserted into the plasmid pMD18-T (TaKaRa, Japan) and then transformed into *Escherichia coli* DH5α before being sequenced using the primers M13-47 (5′-CGCCAGGGTTTTCCCAGT CACGAC-3′) and/or RV-M (5′-GAGCGGATAACAAT TTCACACAGG-3′). The obtained sequences were used as "reference sequences" to verify the reliability of INSD sequences as described in the next subsection.

Reliability analysis of nucleotide records in the INSD

The reliability of *O. sinensis* sequences in the INSD was examined by two approaches. In the first and most important approach, genetic distances were compared after they were calculated with the Kimura 2-parameter (K2P) model as implemented in *MEGA* version 5 [43]. In this approach, the K2P distances among the five "reference sequences" for each of the seven genes were first calculated, and the criteria that would be used to evaluate the reliability of INSD sequences were determined; these criteria were generally the maximum value for the K2P distances for each gene (Table 3). Then, for each of the seven genes, the K2P distances between INSD sequences and the five "reference sequences" were calculated and compared with the criteria values. Those INSD entries with distance values smaller than or equal to the criteria values with at least three of the "reference sequences" for each gene were directly considered as correct *O. sinensis* sequences. Otherwise, the INSD sequences were subjected to further analyses by the second approach. In the second approach, BLAST analyses were performed against the nucleotide database or against the non-redundant protein sequence database at NCBI. If the BLAST result for an INSD entry reported strong hits with non-*O. sinensis* fungi, that entry was regarded as incorrect. If the BLAST result for an INSD entry reported strong hits with correctly identified *O. sinensis* sequences, that entry was generally regarded as indeterminate. One exception was for ITS sequences; an ITS entry was also regarded as incorrect if the identity values with correct sequences were below 90 %, an identity threshold largely beyond a fungal species in most cases [32]. In addition, if different parts (>100 bp for each part) of an INSD entry had significant hits with different fungal taxa, that entry was regarded as chimeric.

Genetic diversity in *O. sinensis*

Intraspecific genetic diversity of *O. sinensis* was investigated based on the sequences of each of the seven genes in the INSD that were herein determined to be correctly identified as belonging to *O. sinensis* (see section

Table 2 Primers for amplification of *O. sinensis* genes

Gene fragment	Primer name	Primer sequence (5´–3´)	Direction[a]	Annealing temperature (°C)[b]	Expected size of amplified fragments (bp)	Reference
nrDNA ITS	ITS5	GGAAGTAAAAGTCGTAACAAGG	F	54	~550	[45]
	ITS4	TCCTCCGCTTATTGATATGC	R			
nrSSU	NS1	GTAGTCATATGCTTGTCTC	F	50	~1,100	[45]
	NS4	CTTCCGTCAATTCCTTTAAG	R			
	SR9R	YAGAGGTGAAATTCT	F	42	~850	Vilgalys lab[c]
	SR6	TGTTACGACTTTTACTT	R			
nrLSU	LR0R	ACCCGCTGAACTTAAGC	F	51	~1,100	Vilgalys lab
	LR6	CGCCAGTTCTGCTTACC	R			
	LR17R	TAACCTATTCTCAAACTT	F	51 → 42	~1,100	Vilgalys lab
	LR9	AGAGCACTGGGCAGAAA	R			
tub	T1	AACATGCGTGAGATTGTAAGT	F	55 → 51	~1,600	[34]
	T22	TCTGGATGTTGTTGGGAATCC	R			
tef1	983F	GCYCCYGGHCAYCGTGAYTTYAT	F	65 → 57	~1,100	[42]
	2218R	ATGACACCRACRGCRACRGTYTG	R			
rpb1	CRPB1	CCWGGYTTYATCAAGAARGT	F	58 → 52	~700	[7]
	RPB1Cr	CCNGCDATNTCRTTRTCCATRTA	R			
rpb2	fRPB2-5F	GAYGAYMGWGATCAYTTYGG	F	58 → 52	~1,200	[26]
	fRPB2-7cR	CCCATRGCTTGTYYRCCCAT	R			

[a] *F* forward, *R* reverse

[b] For fragments with two separate annealing temperatures indicated by a *right-directed arrow*, PCR was first performed at a higher annealing temperature for five cycles, then at a lower annealing temperature for the remaining 30 cycles. For other fragments, PCR was performed at a constant annealing temperature for 35 cycles

[c] Vilgalys R Conserved primer sequences for PCR amplification and sequencing from nuclear ribosomal RNA. http://www.biology.duke.edu/fungi/mycolab/primers.htm

Table 3 Nucleotide variations among reference sequences from *Ophiocordyceps sinensis* and the criteria used for verifying the INSD sequences

Gene	No. isolates	Length (bp)	No. variable sites	No. indel sites	Mean K2P	Max. K2P	Criterion
nrDNA ITS	5	538	21	3	0.0186	0.0306	0.0430[a]
nrSSU	5	1,666	11	0	0.0028	0.0048	0.0048
nrLSU	5	2,050	11	0	0.0026	0.0039	0.0039
tub	5	1,339	22	11	0.0074	0.0129	0.0129
tef1	5	988	22	0	0.0095	0.0185	0.0185
rpb1	5	717	15	0	0.0085	0.0199	0.0199
rpb2	5	1,194	15	0	0.0052	0.0093	0.0093

[a] The criterion for nrDNA ITS is based on Zhang et al. [63]; that paper reported a K2P value of 0.043 based on ITS sequences from 56 *O. sinensis* isolates. This value is also correct when more than 100 *O. sinensis* isolates from different areas were evaluated in our laboratory

"Results"). Numbers of variable nucleotide sites and alleles were calculated with DnaSP version 5.10 [24]. K2P distance values were calculated as described above.

Genetic difference between *O. sinensis* and related fungi

Nucleotide sequences of *O. sinensis*-related taxa were downloaded from the INSD. These included *O. robertsii*, *O. nepalensis*, *O. multiaxialis*, *O. crassispora*, *O. gansuensis*, and groups B and C identified by Stensrud et al. [41]. K2P values between *O. sinensis* reference sequences and these related taxa were calculated as described above.

Phylogenetic analysis

Two different sequence alignments were constructed based on (1) the ITS1-5.8S-ITS2 region (521 characters), and (2) the 5.8S coding region alone (153 characters). Each sequence alignment included 61 ITS sequences of *O. sinensis* submitted by our research group and 33 (for the former alignment) or 37 (for the latter alignment) INSD entries that were erroneously accessioned. Minimum evolutionary phylogenetic analyses were conducted with the K2P model using *MEGA* version 5 [43]. The validity of the phylogenetic relationships was assessed by bootstrap tests with 1,000 resamplings.

Nucleotide sequence accession numbers

The nucleotide sequence data reported in this paper were deposited in GenBank under accession numbers JX968004 to JX968033 (Table 1).

Results

Statistics of INSD records

With "*O. sinensis*" as the keyword query, 874 DNA entries were retrieved from the INSD on October 15, 2012. In addition, there were 254 genome survey sequences of the fungus in GenBank submitted by our research group; they all originated from a shotgun genomic library constructed using the *O. sinensis* isolate YN07-8 [60] and were not analyzed in the following sections. The 874 DNA entries include sequences of nrDNA (ITS, nrLSU, and nrSSU), the mating-type gene *MAT1-2-1*, the DNA lyase gene, serine protease genes (*csp1* and *csp2*), *rpb1*, *rpb2*, *tef1*, *tub*, random marker sequences (OSRCs) with unknown functions, and microsatellite loci. Sequences of nrDNA (555 entries) accounted for 63.5 % of the entries, and ITS sequences (397 entries) accounted for 45.4 % of the entries or 71.5 % of nrDNA entries (Table 4). For protein-coding genes, *MAT1-2-1* and *tub* had the largest number of entries (>60 entries) while *csp1*, *csp2*, and *rpb2* had only 1–2 entries (Table 4). The lengths of sequences of each individual gene in the INSD varied greatly, especially those of ITS, nrSSU, nrLSU, and *tub* (Table 4).

The first *O. sinensis* sequence was submitted to the INSD in 1996 [17], although this entry was most likely erroneously accessioned according to our subsequent analysis. Since that time, rapid increases in *O. sinensis* entries in the INSD occurred during two time periods: from 1999 to 2002 and from 2008 to the present (Fig. 1). During the latter period, *O. sinensis* entries in INSD tripled. All sequences in the INSD were exclusively representative of nrDNA before 2006, but sequences included those for the protein-coding genes of *O. sinensis* after 2006 (Table 4; Fig. S1).

Although *O. sinensis* is mainly distributed in China, scientists from more than ten countries submitted *O. sinensis* sequences to the INSD (Table 5), suggesting a worldwide interest in the fungus. Given the medicinal, cultural, and economic importance of Chinese cordyceps in China, it is not surprising that Chinese researchers submitted 78.8 % (689 records) of the total entries. In China, more than 24 research groups or institutes have submitted *O. sinensis* sequences to the INSD, and this number accounted for 60.0 % of all research groups or institutes submitting *O. sinensis* sequences to the INSD (Table 5). However, 62.0 % (542 records) of the overall entries were marked as unpublished even though some (ca. 71 entries as far as we know) were published.

Reliability of sequences deposited in the INSD

As of October 15, 2012, all sequences of *MAT1-2-1*, DNA lyase gene, *csp1*, *csp2*, and OSRCs represented in the INSD were deposited by our research group [58, 60, 62]. They originated from authentic *O. sinensis* isolates, i.e., isolate identification was based on morphological analysis and sequence analysis of the ITS region of nrDNA, and so their reliability was not suspected in this study. The recently reported microsatellite sequences were not analyzed in this study, and they represent true *O. sinensis* sequences judging from the qualification of the publication [46]. Because sequences of nrDNA ITS, nrLSU, nrSSU, *rpb1*, *rpb2*, *tef1*, and *tub* in INSD were submitted by various independent research groups, their reliabilities were carefully examined herein.

The ITS sequences were the most abundant among all the INSD records (Table 4), and novel entries have been submitted to the INSD almost every year since 1998 (Fig. S1). Most ITS entries included the full-length ITS region (i.e., ITS1, 5.8S, and ITS2) while some were incomplete, consisting of either the ITS1 and 5.8S gene or the ITS2 and 5.8S gene only. Of the 397 ITS entries, 348 were determined to be correct *O. sinensis* sequences, 39 represented other fungi, seven may be chimeric, and three were indeterminate (Table S1). The 39 incorrect sequences had K2P values of 0.1051–0.4581 with *O. sinensis* reference sequences and actually represent species of *Fusarium*, *Tolypocladium*, *Coniochaeta*, *Eurotium*, *Truncatella*, or others according to BLAST results. They clustered into more than six groups (B–G) according to phylogenetic analysis (Fig. 2). In addition, some INSD entries had faulty annotations. For example, AJ488236–AJ488275, AJ488278, and EF378610 were originally submitted using minus strand, and different ITS regions (i.e., ITS1, 5.8S, and ITS2) were wrongly annotated. Similarly, BD167299, BD167302, BD167305, BD167308, BD167311, BD167314, BD167317, BD167320, BD167323, and BD167325 were annotated as the ribosomal RNA gene alone, and HM594290 and HM637742 were submitted as mitochondrial genes, but they

Table 4 Statistics of genes annotated as *O. sinensis* in the INSD

Gene	No. INSD entries	Percentage (%)	Accession no., author, and year of the first entry in INSD	Length (bp)
nrDNA ITS	397	45.4	AF056582, Kang (1998) (AF122027, Chao and Li 1999)[a]	261–687[b]
nrSSU	105	12.0	D86053, Ito and Hirano (1996) (AJ007566, Chen and Hseu 1997)[a]	260–1,791
nrLSU	53	6.1	AB067704, Kinjo (2001)	461–3,315
tub	73	8.4	HM804956, Pandey et al. (2010)	433–1,201
tef1	31	3.5	EF468767, Sung et al. (2007)	533–674
rpb1	29	3.3	EF468874, Sung et al. (2007)	521–703
rpb2	2	0.2	EF468924, Sung et al. (2007)	890–1,040
MAT1-2-1	60	6.9	FJ654150, Zhang et al. (2009)	877–882 (7,996)[c]
DNA lyase gene	1[d]	–	HM212637, Zhang et al. (2010)	898
csp1	1	0.1	EU282382, Zhang et al. (2007)	2,143
csp2	1	0.1	EU282383, Zhang et al. (2007)	2,566
OSRC1	4	0.5	JQ277357, Zhang et al. (2012)	604
OSRC2	4	0.5	JQ277361, Zhang et al. (2012)	665
OSRC3	4	0.5	JQ277365, Zhang et al. (2012)	570–573
OSRC4	4	0.5	JQ277369, Zhang et al. (2012)	717
OSRC7	4	0.5	JQ277373, Zhang et al. (2012)	461–464
OSRC9	4	0.5	JQ277377, Zhang et al. (2012)	377
OSRC11	4	0.5	JQ277381, Zhang et al. (2012)	471–478
OSRC13	4	0.5	JQ277385, Zhang et al. (2012)	584–588
OSRC14	4	0.5	JQ277389, Zhang et al. (2012)	577–590
OSRC16	4	0.5	JQ277393, Zhang et al. (2012)	513
OSRC17	4	0.5	JQ277397, Zhang et al. (2012)	480
OSRC18	4	0.5	JQ277401, Zhang et al. (2012)	378
OSRC19	4	0.5	JQ277405, Zhang et al. (2012)	525
OSRC21	4	0.5	JQ277409, Zhang et al. (2012)	455
OSRC22	4	0.5	JQ277413, Zhang et al. (2012)	485
OSRC23	4	0.5	JQ277417, Zhang et al. (2012)	716–721
OSRC24	4	0.5	JQ277421, Zhang et al. (2012)	534
OSRC25	4	0.5	JQ277425, Zhang et al. (2012)	563–564
OSRC26	4	0.5	JQ277429, Zhang et al. (2012)	559
OSRC27	4	0.5	JQ277433, Zhang et al. (2012)	572–575
OSRC28	4	0.5	JQ277437, Zhang et al. (2012)	723
OSRC31	4	0.5	JQ277441, Zhang et al. (2012)	400
OSRC32	4	0.5	JQ277445, Zhang et al. (2012)	614–638
Microsatellites	30	3.4	JF781128, Wang et al. (2012)	56–204
Total	874			

[a] The first records of nrDNA ITS and nrSSU in the INSD were erroneously annotated. The earliest correct ones are in *parentheses*

[b] Partial sequences of nrLSU and nrSSU were included in some nrDNA ITS entries. The actual length of the ITS region (including ITS1, 5.8S, and ITS2) was ~500 bp

[c] Fifty-nine of *MAT1-2-1* entries were 877–882 bp long except for HM212637, which was 6,996 bp long because both flanking sequences were included in this entry

[d] The sequence of the DNA lyase gene was embedded in a *MAT1-2-1* entry, and they share the accession no. of HM212637

all belong to nrDNA ITS sequences. In another example, AF122030 was annotated as complete ITS2 and partial 18S and 5.8S, but it included the whole ITS region. Partial sequences of some entries may be incorrect, such as the first 25 nucleotides of AJ488255, the first 80 nucleotides of AJ488256, the first 30 nucleotides of AJ488257, the 50th to 70th nucleotides of AJ488275, the 90th to 120th nucleotides of GU246277, the last 50 nucleotides of GU246288. JQ936581 seems to have lost 66 bp of sequences in the ITS2 region.

The nrSSU sequences were the second most abundant among all INSD records (Table 4). Of the 105 nrSSU

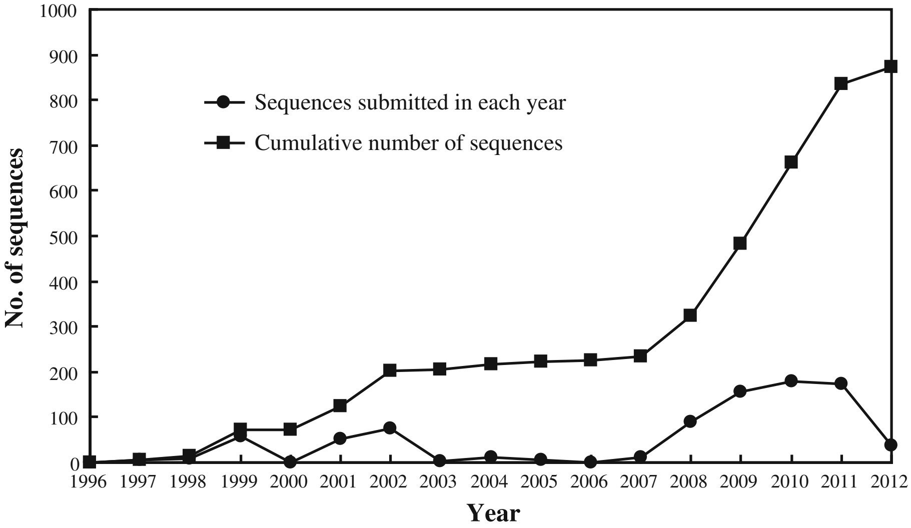

Fig. 1 Increase in INSD entries annotated as *O. sinensis* between 1996 and 2012. The entries for 2012 are probably incomplete because the data were collected up to October 15, 2012 and the entries submitted in 2012 may need some time to become open to public

Table 5 INSD entries submitted by scientists from different countries

Country	No. INSD entries	Percentage (%)	No. research groups	Percentage (%)
China[a]	689	78.8	24	60.0
India	97	11.1	2	5.0
Japan	41	4.7	3	7.5
USA	8	0.9	3	7.5
Germany	3	0.3	1	2.5
Switzerland	3	0.3	1	2.5
Korea	2	0.2	2	5.0
Norway	1	0.1	1	2.5
South Africa	1	0.1	1	2.5
Spain	1	0.1	1	2.5
Unknown	28	3.2	1	2.5

[a] This included scientists from China mainland, Hong Kong, and Taiwan

sequences, 78 represented *O. sinensis*, and 27 were incorrectly attributed to *O. sinensis* (Table S1). These false entries had K2P values of 0.0117–0.0877 with reference sequences, and BLAST analyses showed that they represent species of *Fusarium*, *Paecilomyces/Isaria*, *Cordyceps s.s.*, or others. Among all nrSSU sequences, BD167298, BD167301, BD167304, BD167307, BD167310, BD167313, BD167316, BD167319, and BD167322 were submitted as representing the ribosomal RNA gene, but they actually represent nrSSU sequences.

Of the 53 nrLSU sequences, 43 were correct *O. sinensis* sequences, five were incorrect, and five were indeterminate (Table S1). These indeterminate records had K2P values of 0.0044–0.0322 with reference sequences. These incorrect records had K2P values of 0.0410–0.1751 with reference sequences and represent other fungal species within the order of Hypocreales. Among all nrLSU sequences, BD167300, BD167303, BD167306, BD167309, BD167312, BD167315, BD167318, BD167321, and BD167324 were submitted as representing the ribosomal RNA gene, but they actually represented nrLSU sequences.

Of the 73 *tub* sequences, 59 represented *O. sinensis*, and 14 were indeterminate. The latter might be chimeric or represent different copies of *O. sinensis tub* genes (Table S1). All 31 *tef1* sequences, 29 *rpb1* sequences, and two *rpb2* sequences in the INSD until today were correct *O. sinensis* sequences.

Genetic diversity within *O. sinensis*

Intraspecific genetic variations of *O. sinensis* were determined using all correct INSD sequences identified above. The extent of genetic variations differed depending on the gene and number of sequences (Table 6). High genetic diversity was detected in nrDNA ITS region; within the nrDNA ITS region, the 5.8S region was the most conserved, and the ITS2 region was the most variable (Table 6; Fig. 3). Nucleotide diversity was also high for *MAT1-2-1*, OSRC14, OSRC17, OSRC19, OSRC22, OSRC27, and OSRC32 (Table 6). Among INSD sequences correctly identified as *O. sinensis*, the maximum K2P values slightly higher than the criteria values used in the reliability analyses were detected for nrDNA ITS and nrLSU (Tables 3

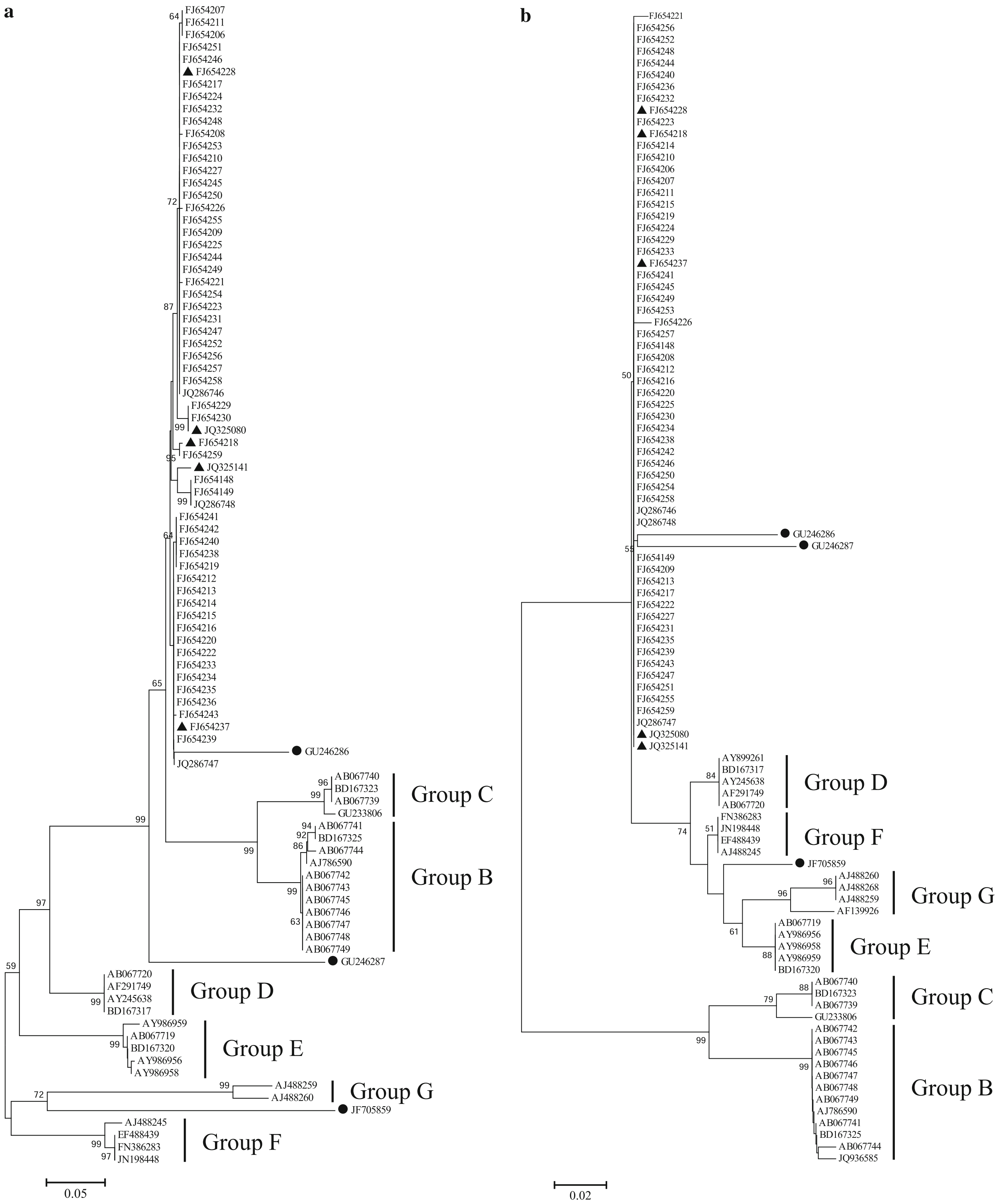

Fig. 2 Phylogeny derived from the ITS1-5.8S-ITS2 data set (**a**) and the 5.8S data set (**b**) based on erroneously annotated sequences in the INSD and correct sequences from our research group. A total of 39 ITS entries were identified as wrongly accessioned entries, whereas only 33 or 37 were used to perform phylogenetic analyses due to incomplete sequence for some ITS entries. The three entries marked with *black balls* were erroneously accessioned but did not form a clade with other entries. The remaining entries were correct sequences of *O. sinensis* from our research group, and the five "reference sequences" used in this study were marked with *black triangles*

Table 6 Nucleotide variations among correct *O. sinensis* sequences in the INSD

Gene fragment	No. entries	Aligned length	No. variable sites (%)[a]	No. phylogenetically informative sites (%)	No. phylogenetically uninformative sites	No. indel sites	No. of alleles (Frequency of the most common allele)[b]	Mean K2P	Max. K2P	Position[c]
nrDNA ITS	304	502	137 (27.3)	49 (9.8)	37	51	103 (113)	0.0107	0.0513	
ITS1	318	162	50 (30.9)	22 (12.9)	16	12	45 (143)	0.0098	0.0808	1–162
5.8S	336	157	22 (14.0)	10 (6.4)	10	2	23 (269)	0.0035	0.0395	163–319
ITS2	331	183	67 (36.6)	19 (10.4)	15	33	55 (130)	0.0187	0.0812	320–502
nrSSU	17	1,727	7 (0.4)	5 (0.3)	1	1	5 (10)	0.0010	0.0029	
	41	381	3 (0.8)	0 (0.0)	2	1	4 (38)	0.0003	0.0053	601–981
	41	268	2 (0.7)	1 (0.4)	0	1	3 (19)	0.0019	0.0038	1,110–1,376
	34	335	2 (0.6)	1 (0.3)	1	0	3 (31)	0.0007	0.0060	1,393–1,727
nrLSU	18	1,360	10 (0.7)	9 (0.7)	1	0	5 (6)	0.0021	0.0052	
	42	790	21 (2.7)	7 (0.9)	7	7	15 (16)	0.0017	0.0077	10–798
tub	33	1,201	11 (0.9)	9 (0.7)	2	0	6 (14)	0.0035	0.0075	
	59	424	12 (2.8)	10 (2.4)	0	2	8 (24)	0.0039	0.0119	1–423
tef1	31	513	12 (2.3)	8 (1.6)	4	0	9 (15)	0.0051	0.0158	
rpb1	29	521	8 (1.5)	5 (0.9)	3	0	7 (21)	0.0015	0.0077	
rpb2	2	870	5 (0.6)	0 (0.0)	5	0	2 (1)	0.0058	0.0058	
MAT1-2-1	59	882	34 (3.9)	23 (2.6)	6	5	13 (32)	0.0044	0.0244	
OSRC1	5	604	3 (0.5)	3 (0.5)	0	0	2 (3)	0.0030	0.0050	
OSRC2	5	665	0 (0.0)	0 (0.0)	0	0	1 (5)	0.0000	0.0000	
OSRC3	5	573	6 (1.1)	2 (0.4)	1	3	3 (2)	0.0028	0.0035	
OSRC4	5	717	3 (0.4)	2 (0.3)	1	0	3 (2)	0.0022	0.0042	
OSRC7	5	464	7 (1.5)	0 (0.0)	4	3	3 (3)	0.0035	0.0087	
OSRC9	5	377	7 (1.9)	3 (0.8)	4	0	3 (2)	0.0091	0.0188	
OSRC11	5	478	8 (1.7)	0 (0.0)	1	7	3 (3)	0.0008	0.0021	
OSRC13	5	591	16 (2.7)	1 (0.2)	5	10	3 (2)	0.0045	0.0104	
OSRC14	5	590	32 (5.4)	7 (1.2)	12	13	3 (2)	0.0159	0.0283	
OSRC16	5	513	7 (1.4)	4 (0.8)	3	0	3 (2)	0.0071	0.0098	
OSRC17	5	481	11 (2.3)	3 (0.6)	6	2	3 (2)	0.0088	0.0169	
OSRC18	5	378	5 (1.3)	4 (1.1)	1	0	3 (3)	0.0075	0.0134	
OSRC19	5	525	29 (5.5)	28 (5.3)	1	0	3 (2)	0.0340	0.0574	
OSRC21	5	454	1 (0.2)	1 (0.2)	0	0	2 (3)	0.0013	0.0022	
OSRC22	5	485	10 (2.1)	1 (0.2)	9	0	3 (2)	0.0088	0.0211	
OSRC23	5	721	7 (0.9)	1 (0.1)	1	5	3 (2)	0.0014	0.0028	
OSRC24	5	534	9 (1.7)	2 (0.4)	7	0	4 (2)	0.0076	0.0151	
OSRC25	5	564	10 (1.8)	1 (0.2)	8	1	3 (2)	0.0068	0.0162	
OSRC26	5	559	5 (0.9)	1 (0.2)	4	0	3 (2)	0.0040	0.0090	
OSRC27	5	575	17 (2.9)	0 (0.0)	14	3	2 (4)	0.0100	0.0250	
OSRC28	5	723	8 (1.1)	2 (0.3)	6	0	4 (2)	0.0050	0.0112	
OSRC31	5	400	3 (0.8)	1 (0.3)	2	0	3 (2)	0.0035	0.0075	
OSRC32	5	638	88 (13.8)	43 (6.7)	21	24	3 (2)	0.0592	0.0864	

[a] Variable sites consist of both indel (insertion/deletion) sites and substitution sites, and the latter contain both phylogenetically informative and uninformative sites

[b] An allele refers to any variant of DNA sequence observed at a given locus (gene) with indel sites considered. “Frequency of the most common allele” indicates the number of individuals represented by the dominate allele

[c] Positions are given for shorter fragments of a gene relative to the longest fragment

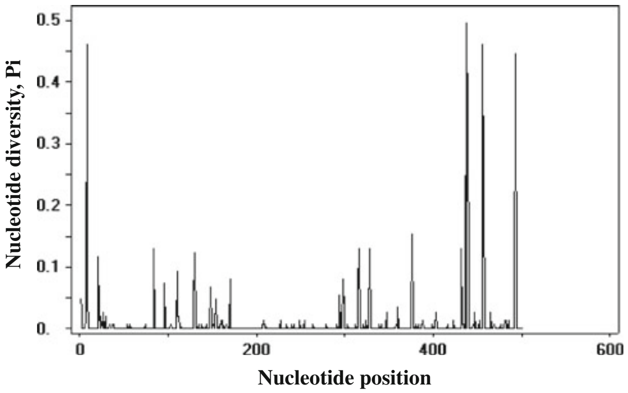

Fig. 3 Nucleotide diversity of different nucleotide sites of nrDNA ITS from *O. sinensis*. This figure was plotted by DnaSP software based on 304 correct *O. sinensis* INSD sequences with the full-length ITS region (i.e., ITS1, 5.8S, and ITS2). The 5.8S region is located between the 164th and 319th position. Almost all sites in the 5.8S region had nucleotide diversity indices lower than 0.1, whereas several sites in the ITS1 and ITS2 regions had nucleotide diversity indices larger than 0.1

and 6). The maximum K2P values for other genes as well as the mean K2P values for all genes did not exceed the criteria values used in reliability analyses (Tables 3 and 6).

Genetic difference between *O. sinensis* and other related taxa

Four records annotated as *O. robertsii* were found in the INSD, and they included two nrDNA ITS records, one nrLSU record, and one *tef1* record. The two nrDNA ITS records had high sequence dissimilarity (K2P = 0.1396). According to BLAST analyses, one (AJ309335) represents *O. robertsii* and the other (AJ309336) represents *Trichoderma* sp. Based on correct *O. robertsii* sequences, we found that the genetic distances between *O. sinensis* and *O. robertsii* were larger than the intraspecific variation of *O. sinensis* (Tables 3, 6, and 7).

Based on morphological characters, six species similar to *O. sinensis* have previously been reported on the Tibetan Plateau [9, 54, 56, 57]. No sequences of *O. kangdingensis* and *O. laojunshanensis* were accessible in the INSD, and the other four species (*O. crassispora*, *O. gansuensis*, *O. nepalensis*, and *O. multiaxialis*) had one to four entries in the INSD. The K2P values between *O. sinensis* and sequences annotated as *O. nepalensis* or *O. multiaxialis* were smaller than the maximal K2P values among *O. sinensis* sequences (Tables 3, 6, and 7). This is consistent with previous conclusions that *O. nepalensis* and *O. multiaxialis* are synonyms of *O. sinensis* [27, 40]. The K2P values between *O. sinensis* and sequences annotated as *O. crassispora* or *O. gansuensis*, however, were much larger than the maximal K2P values among *O. sinensis* sequences (Tables 3, 6, and 7). BLAST analyses showed that the two records annotated as *O. gansuensis* represent *Penicillium* sp. and that the four entries annotated as *O. crassispora* represent other Hypocreales species (Table 7). In addition, INSD sequences of *O. nepalensis* and *O. multiaxialis* but not of *O. crassispora* or *O. gansuensis* originated from type specimens. Therefore, determining whether *O. crassispora* and *O. gansuensis* are synonyms of *O. sinensis* will require more evaluation.

Groups B and C in the ITS phylogenetic trees (Fig. 2) were controversial. They were regarded as either cryptic species of *O. sinensis* [41] or different species [49] by different researchers. These sequences showed strong BLAST hits with correct *O. sinensis* although their similarity to *O. sinensis* was lower than 90 %. Further analysis indicated that even the minimum K2P values (0.1181–0.1495) between *O. sinensis* and groups B and C were much larger than the maximum K2P values between *O. sinensis* sequences (Tables 3, 6, and 7). Therefore, groups B and C are unlikely to be cryptic species of *O. sinensis*.

Discussion

Nucleic acid sequences deposited in public databases are widely used by systematic and evolutionary biologists to construct systematic frameworks and to model evolutionary pathways. According to the evaluation of fungal nrDNA sequences, as many as 20 % of the sequences in public databases may be unreliable [5, 33]. Because certain sequences in the INSD have been annotated erroneously [5, 33, 41] and because intraspecific genetic diversity is substantial in many organisms [32, 38, 63], the reliability of sequences annotated as a given species in INSD should be examined by those with substantial experience with the organism in question. The current study examined the reliability of all INSD sequences annotated as *O. sinensis*. We used a method that should be useful for evaluating the sequences of other organisms. In this method, genetic distances between INSD sequences and sequences from authenticated materials were carefully compared with maximal genetic distances among different authentic sequences as the criteria. In addition, BLAST searches were performed as a supplement.

As a valuable medicinal fungus, *O. sinensis* has become a focus of scientific research and commercialization in recent years [61]. The fungus is difficult to study, however, because it evidently grows only in harsh alpine environments on the Tibetan Plateau. *O. sinensis* is also difficult to study because in nature it is only found as a parasite of insects, and the parasitized insect (Chinese cordyceps) supports a complex microbial community [65]. These difficulties have generated confusion in the identification of *O. sinensis*. Some of this confusion has been reduced by

Table 7 Comparison of genetic distance between *O. sinensis* and related fungi

Fungal taxa	Accession no.	Gene fragment	Original length (bp)	Aligned length (bp)	Min. K2P	Max. K2P	Comment
O. robertsii	AJ309335	nrDNA ITS	623	557	0.0567	0.0651	
	AJ309336	nrDNA ITS	625	576	0.1847	0.1964	May represent *Trichoderma* sp.
	EF468826	nrLSU	894	863	0.0237	0.0261	
	EF468766	*tef1*	689	687	0.0298	0.0360	
O. crassispora	AB067714	nrDNA ITS	647	544	0.2013	0.2098	May represent *Neonectria* sp.
	FJ025150	nrDNA ITS	634	537	0.2043	0.2098	May represent *Neonectria* sp.
	AB067697	nrSSU	1,728	1,667	0.0226	0.0251	May represent *Myrothecium* sp.
	AB067706	nrLSU	1,413	1,387	0.0659	0.0683	May represent Hypocreales sp.
O. gansuensis	AF056583	nrDNA ITS	301	315	0.4743	0.5021	May represent *Penicillium* sp.
	AF139929	nrDNA ITS	260	286	0.2116	0.2277	May represent *Penicillium* sp.
O. nepalensis	AJ309358	nrDNA ITS	601	538	0.0038	0.0209	
O. multiaxialis	AJ309359	nrDNA ITS	602	538	0.0000	0.0229	
Group B	AB067741	nrDNA ITS	685	507	0.1308	0.1535	
	AB067742	nrDNA ITS	604	507	0.1181	0.1403	
	AB067743	nrDNA ITS	604	507	0.1206	0.1429	
	AB067744	nrDNA ITS	604	507	0.1308	0.1535	
	AB067745	nrDNA ITS	562	507	0.1206	0.1429	
	AB067746	nrDNA ITS	562	507	0.1206	0.1429	
	AB067747	nrDNA ITS	561	507	0.1209	0.1432	
	AB067748	nrDNA ITS	685	507	0.1228	0.1452	
	AB067749	nrDNA ITS	687	507	0.1206	0.1429	
	AB067750	nrDNA ITS	505	507	0.1231	0.1455	
	BD167325	nrDNA ITS	685	507	0.1308	0.1535	
	JQ936585	nrDNA ITS	359	360	0.1301	0.1561	
Group C	AB067739	nrDNA ITS	686	537	0.1495	0.1697	
	AB067740	nrDNA ITS	686	537	0.1495	0.1697	
	BD167323	nrDNA ITS	686	537	0.1495	0.1697	
	GU233806	nrDNA ITS	604	537	0.1495	0.1697	

recent research. For example, a recent study reported that of the 91 insect species associated with *O. sinensis* in the literature, 57 are potential hosts, eight are indeterminate hosts, and 26 are non-hosts [47]. Of the 203 distribution sites for *O. sinensis* recorded in publications, 106 are considered as confirmed distribution sites, 65 as possible distribution sites, 29 as “non-distribution” sites, and three as “suspicious” distribution sites [22]. Analyses of fungal materials used in 152 papers on *O. sinensis* from PubMed since 1998 showed that at least 116 papers (over 75 %) used unreliable, uncertain, or unspecified materials [13]. Errors in INSD sequences have been reported [41], and the reliability of all INSD sequences deposited as from *O. sinensis* therefore deserved careful evaluation.

Of the 874 nucleotide sequences that were in the INSD as of October 15, 2012 and that were evaluated in this study, 774 are considered as *O. sinensis* sequences, 71 as sequences of other fungi, seven as chimeras, and 22 as indeterminate sequences. Results of this study have provided essential reference for INSD staff and for scientists studying *O. sinensis* and related species. Because nrDNA ITS sequences were the most abundant sequences associated with *O. sinensis* in the INSD and also had the most annotation errors, we strongly recommend that, if they intend to report *O. sinensis* from new environments or to submit an ITS sequence under the name *O. sinensis*, researchers compare their sequences with our reference sequences and adhere to the criteria we have established for ITS sequences in this study (Table 3). We also strongly recommend that the quality of a sequence be carefully checked (i.e., the chromatograms should be carefully examined) before it is submitted to INSD.

Annotation error on newly generated sequences can be avoided if enough attention is paid, but how to rectify pre-existing erroneously accessioned entries in the INSD? One simple solution to this problem is that such correction can be done by submitters of those sequences. The reality, however, is that these submitters often have moved on to other projects and never get around to making the changes [35]. Some researchers have submitted a sequence of *O. sinensis* to INSD, but they may not want to perform an in depth research on this fungus. Another solution is that researches who have discovered inaccuracies should append corrections [4]. However, currently, third-party annotation tools are poorly developed in the INSD [35]. Anyway, appropriate measures need to be worked out as early as possible in order to prevent error propagation, which would degrade the quality of the INSD. It is obvious that these discussions have been beyond the scope of this study, and herein, we want to alert subsequent INSD users to the presence of such defective data of *O. sinensis*.

Why have so many sequences been erroneously accessioned in INSD? One possible reason is that researchers have incorrectly assumed that sequences obtained from Chinese cordyceps represent *O. sinensis*. This is a poor assumption because a natural Chinese cordyceps specimen harbors a complex fungal community in which *O. sinensis* is the dominant but not the only member [65]. Another possible reason is that the original material was misidentified. Too often materials have been identified based only on BLAST results of nrDNA sequences without a careful check of their morphological characteristics. Under these conditions, if a sequence had strong BLAST hits with an INSD entry erroneously annotated as *O. sinensis*, that sequence would have been incorrectly submitted to the INSD as *O. sinensis*. Considering these problems, it is understandable why many publications have reported *O. sinensis* from unexpected substrates and locations such as from plants and from soil outside the Tibetan Plateau.

Although *O. sinensis* is genetically different from other taxa, the sequence variation for many of its genes is high. Given their high sequence variation, genes like ITS, *MAT1-2-1*, OSRC14, OSRC17, OSRC19, OSRC22, OSRC27, and OSRC32 have the potential to serve as markers for the analysis of genetic diversity in *O. sinensis*. Genes like the partial nrSSU fragment, *rpb1*, OSRC2, and OSRC11, in contrast, are relatively conserved, and therefore have the potential to serve as markers for identification and DNA barcoding of *O. sinensis*.

Stensrud et al. [41] analyzed the nucleotide variation of 71 ITS sequences annotated as *O. sinensis* in the INSD and placed these sequences in five groups (A–E). Based on the additional ITS sequences submitted to the INSD in the following years, additional groups (F and G) were identified in the current study (Fig. 2). Group A represents true *O. sinensis* sequences. Groups D–G and JF705859 represent other fungi according to BLAST results. Groups B and C, GU246286, and GU246287 had high scoring hits with correct *O. sinensis* sequences when BLAST was performed, but their identities with correct *O. sinensis* sequences were less than 90 %, and they were thus treated as incorrect sequences in this study. Sequences in groups B and C, which were originally submitted by Kinjo, have been reported by different researchers, and have been detected by our laboratory during an analysis of an ITS clone library of Chinese cordyceps (unpublished data). Unfortunately, sequences of groups B and C were all reported from natural Chinese cordyceps samples rather than from isolated fungal cultures. Determining whether they are cryptic species of *O. sinensis* [41] or different species [49] will require the isolation of fungi with those sequences. Solving this problem seems to be urgent, considering that several papers have been published by a research group who regarded groups B and C as different mutant genotypes during maturation of *O. sinensis* in nature [15, 16, 52, 67, 69].

Acknowledgments This study was supported by grants from the National Plans of Sciences and Technology (2007BAI32B04), the cooperative project between Guangdong Province and Chinese Academy of Sciences (2009B091300015), the National Science Foundation of China (81102759; 31140013), and the Specialized Research Fund for the Doctoral Program of Higher Education (20101401120007). We also thank Prof. Bruce Jaffee (University of California at Davis) for serving as a pre-submission technical editor.

References

1. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. Nucl Acids Res 40(Database issue): D48–53
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucl Acids Res 39(Database issue):D32–37
3. Berg G, Zachow C, Lottmann J, Gotz M, Costa R, Smalla K (2005) Impact of plant species and site on rhizosphere-associated fungi antagonistic to *Verticillium dahliae* Kleb. Appl Environ Microbiol 71(8):4203–4213
4. Bidartondo MI, 255 others (2008) Preserving accuracy in GenBank. Science 319(5870):1616
5. Bridge PD, Roberts PJ, Spooner BM, Panchal G (2003) On the unreliability of published DNA sequences. New Phytol 160 (1):43–48
6. Cannon PF (2010) The caterpillar fungus, a flagship species for conservation of fungi. Chin J Grassland 32(Supp.):86–88
7. Castlebury LA, Rossman AY, Sung GH, Hyten AS, Spatafora JW (2004) Multigene phylogeny reveals new lineage for *Stachybotrys chartarum*, the indoor air fungus. Mycol Res 108(08):864–872
8. Chee-Sanford J (2008) Weed seeds as nutritional resources for soil Ascomycota and characterization of specific associations between plant and fungal species. Biol Fert Soils 44(5):763–771
9. Chen JY, Cao YQ, Yang DR, Li MH (2011) A new species of *Ophiocordyceps* (Clavicipitales, Ascomycota) from southwestern China. Mycotaxon 115(1):1–4
10. Chen JY, Lee SW, Cao YQ, Peng YQ, Winkler D, Yang DR (2010) Ethnomycological use of medicinal Chinese caterpillar fungus, *Ophiocordyceps sinensis* (Berk.) G. H. Sung et al. (Ascomycetes) in Northern Yunnan Province, SW China. Int J Med Mushrooms 12(4): 427–434

11. Chen YJ, Zhang YP, Yang YX, Yang DR (1999) Genetic diversity and taxonomic implication of *Cordyceps sinensis* as revealed by RAPD markers. Biochem Genet 37(5–6):201–213
12. Dong CH, Yao YJ (2005) Nutritional requirements of mycelial growth of *Cordyceps sinensis* in submerged culture. J Appl Microbiol 99(3):483–492
13. Dong CH, Yao YJ (2010) On the reliability of fungal materials used in studies on *Ophiocordyceps sinensis*. J Ind Microbiol Biotechnol 38(8):1027–1035
14. Dong CH, Yao YJ (2012) Isolation, characterization of melanin derived from *Ophiocordyceps sinensis*, an entomogenous fungus endemic to the Tibetan Plateau. J Biosci Bioeng 113(4):474–479
15. Gao L, Li XH, Zhao JQ, Lu JH, Zhao JG, Zhu JS (2012) Maturation of *Cordyceps sinensis* associates with alterations of fungal expressions of multiple *Ophiocordyceps sinensis* mutants in stroma of *Cordyceps sinensis*. J Peking Univ 44(3):454–463
16. Gao L, Li XH, Zhao JQ, Lu JH, Zhu JS (2011) Detection of multiple *Ophiocordyceps sinensis* mutants in premature stroma of *Cordyceps sinensis* by MassARRAY SNP MALDI-TOF mass spectrum genotyping. J Peking Univ 43(2):259–266
17. Ito Y, Hirano T (1997) The determination of the partial 18 S ribosomal DNA sequences of *Cordyceps* species. Lett Appl Microbiol 25(4):239–242
18. Jiang Y, Yao YJ (2002) Names related to *Cordyceps sinensis* anamorph. Mycotaxon 84:245–253
19. Jin GS, Wang XL, Li Y, Wang WJ, Yang RH, Ren SY, Yao YJ (2012) Development of conventional and nested PCR assays for the detection of *Ophiocordyceps sinensis*. J Basic Microbiol 52:1–9
20. Li C, Li Z, Fan M, Cheng W, Long Y, Ding T, Ming L (2006) The composition of *Hirsutella sinensis*, anamorph of *Cordyceps sinensis*. J Food Compos Anal 19(8):800–805
21. Li J, zu Dohna H, Miller J, Cardona CJ, Carpenter TE (2010) Identifying errors in avian influenza virus gene sequences and implications for data usage of public databases. Genomics 95 (1):29–36
22. Li Y, Wang X, Jiao L, Jiang Y, Li H, Jiang S, Lhosumtseiring N, Fu S, Dong C, Zhan Y, Yao Y (2011) A survey of the geographic distribution of *Ophiocordyceps sinensis*. J Microbiol 49(6):913–919
23. Liang HH, Cheng Z, Yang XL, Li S, Ding ZQ, Zhou TS, Zhang WJ, Chen JK (2008) Genetic diversity and structure of *Cordyceps sinensis* populations from extensive geographical regions in China as revealed by inter-simple sequence repeat markers. J Microbiol 46(5):549–556
24. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25 (11):1451–1452
25. Lin X, Wang ZF, Huang YJ, Qian XM (2010) A primary study of *Cordyceps* sp. about “host jumping” through plants. J Xiamen Univ (Nat Sci) 49(5):717–720
26. Liu Y, Whelen S, Hall B (1999) Phylogenetic relationships among ascomycetes: evidence from an RNA polymerse II subunit. Mol Biol Evol 16:1799–1808
27. Liu ZY, Yao YJ, Liang ZQ, Liu AY, Pegler DN, Chase MW (2001) Molecular evidence for the anamorph–teleomorph connection in *Cordyceps sinensis*. Mycol Res 105(7):827–832
28. Longo MS, O’Neill MJ, O’Neill RJ (2011) Abundant Human DNA contamination identified in non-primate genome databases. PLoS ONE 6(2):e16410
29. Lu FL, Jiang HY, Ding JH, Mu JB, Valenzuela JG, Ribeiro JMC, Su XZ (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. BMC Genomics 8(1):255
30. Mendonça R, Navia D, Diniz I, Auger P, Navajas M (2011) A critical review on some closely related species of *Tetranychus* sensu stricto (Acari: tetranychidae) in the public DNA sequences databases. Exp Appl Acarol 55(1):1–23
31. Nakamiya K, Hashimoto S, Ito H, Edmonds J, Morita M (2005) Degradation of 1,4–dioxane and cyclic ethers by an isolated fungus. Appl Environ Microbiol 71(3):1254–1258
32. Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson KH (2008) Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. Evol Bioinform 4:193–201
33. Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson KH, Kõljalg U (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLoS ONE 1 (1):e59
34. O’Donnell K, Cigelnik E (1997) Two divergent intragenomic rDNA ITS2 types within a monophyletic lineage of the fungus *Fusarium* are nonorthologous. Mol Phylogenet Evol 7(1):103–116
35. Pennisi E (2008) Proposal to ‘wikify’ GenBank meets stiff resistance. Science 319(5870):1598–1599
36. Rakotonirainy M, Heude E, Lavedrine B (2007) Isolation and attempts of biomolecular characterization of fungal strains associated to foxing on a 19th century book. J Cult Herit 8(2):126–133
37. Rothe J, Nagy M (2012) Strategies for excluding false Y-chromosomal SNP entries from human genome databases. Electrophoresis 33(9–10):1488–1491
38. Ryberg M, Nilsson RH, Kristiansson E, Topel M, Jacobsson S, Larsson E (2008) Mining metadata from unidentified ITS sequences in GenBank: a case study in Inocybe (Basidiomycota). BMC Evol Biol 8:50
39. Sánchez Márquez S, Bills G, Domínguez Acuña L, Zabalgogeazcoa I (2010) Endophytic mycobiota of leaves and roots of the grass *Holcus lanatus*. Fungal Divers 41(1):115–123
40. Shrestha B, Zhang WM, Zhang YJ, Liu XZ (2010) What is the Chinese caterpillar fungus *Ophiocordyceps sinensis* (Ophiocordycipitaceae)? Mycology 1(4):228–236
41. Stensrud Ø, Schumacher T, Shalchian-Tabrizi K, Svegården IB, Kauserud H (2007) Accelerated nrDNA evolution and profound AT bias in the medicinal fungus *Cordyceps sinensis*. Mycol Res 111(4):409–415
42. Sung GH, Sung JM, Hywel-Jones NL, Spatafora JW (2007) A multi-gene phylogeny of Clavicipitaceae (Ascomycota, Fungi): identification of localized incongruence using a combinational bootstrap approach. Mol Phylogenet Evol 44(3):1204–1223
43. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28(10):2731–2739
44. Vilgalys R (2003) Taxonomic misidentification in public DNA databases. New Phytol 160(1):4–5
45. White TJ, Bruns T, Lee S, Taylor JW (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis M, Gelfand DH, Sninsky JJ, White TJ (eds) PCR Protocols: a guide to methods and applications. Academic Press, New York, pp 315–322
46. Wang XL, Yang RH, Yao YJ (2011) Development of microsatellite markers for *Ophiocordyceps sinensis* (Ophiocordycipitaceae) using an ISSR–TAIL–PCR method. Am J Bot 98(12):e391–e394
47. Wang XL, Yao YJ (2011) Host insect species of *Ophiocordyceps sinensis*: a review. ZooKeys 127:43–59
48. Xl Wei, Yin XC, Guo YL, Shen NY, Wei JC (2006) Analyses of molecular systematics on *Cordyceps sinensis* and its related taxa. Mycosystema 25(2):192–202
49. Xiao W, Yang JL, Zhu P, Cheng KD, He HX, Zhu HX, Wang Q (2009) Non-support of species complex hypothesis of *Cordyceps*

sinensis by targeted rDNA-ITS sequence analysis. Mycosystema 28(5):724–730

50. Xiao YY, Chen C, Dong JF, Li CR, Fan MZ (2011) Morphological observation of ascospores of *Ophiocordyceps sinensis* and its anamorph in growth process. J Anhui Agric Univ 38(4):587–591
51. Yang DR (2008) Notes on field investigations of *Ophiocordyceps sinensis* resources in the Tibetan Plateau. China Nature 1:36–39
52. Yao YS, Zhou YJ, Gao L, Lu JH, Wu ZM, Zhu JS (2011) Dynamic alternations of the differential fungal expressions of *Ophiocordyceps sinensis* and its mutant genotypes in stroma and caterpillar during maturation of natural *Cordyceps sinensis*. J Fungal Res 9(1):37–49, 53
53. Yu YX (2004) Studies on artificial culture of *Cordyceps sinensis*. J Fungal Res 2:42–46
54. Zang M, Kinjo N (1998) Notes on the alpine *Cordyceps* of China and nearby nations. Mycotaxon 66:215–229
55. Zang M, Noriko K (1996) Type study on the *Cordyceps sinensis*. Acta Bot Yunnan 18(2):205–208
56. Zang M, Yang DR, Li CD (1990) A new taxon in the genus *Cordyceps* from China. Mycotaxon 37:57–62
57. Zhang KY, Wang CJ, Yan MS (1989) A new species of *Cordyceps* from Gansu, China. Trans Mycol Soc Jpn 30:295–299
58. Zhang S, Zhang YJ, Liu XZ, Wen HA, Wang M, Liu DS (2011) Cloning and analysis of the *MAT1-2-1* gene from the traditional Chinese medicinal fungus *Ophiocordyceps sinensis*. Fungal Biol 115(8):708–714
59. Zhang YJ (2012) Biology of the Chinese Caterpillar Fungus *Ophiocordyceps sinensis*. Science Press, Beijing
60. Zhang YJ, Bai FR, Zhang S, Liu XZ (2012) Determining novel molecular markers in the Chinese caterpillar fungus *Ophiocordyceps sinensis* by screening a shotgun genomic library. Appl Microbiol Biotechnol 95:1243–1251
61. Zhang YJ, Li EW, Wang CS, Li YL, Liu XZ (2012) *Ophiocordyceps sinensis*, the flagship fungus of China: terminology, life strategy and ecology. Mycology 3(1):2–10
62. Zhang YJ, Liu XZ, Wang M (2008) Cloning, expression, and characterization of two novel cuticle-degrading serine proteases from the entomopathogenic fungus *Cordyceps sinensis*. Res Microbiol 159(6):462–469
63. Zhang YJ, Xu LL, Zhang S, Liu XZ, An ZQ, Wang M, Guo YL (2009) Genetic diversity of *Ophiocordyceps sinensis*, a medicinal fungus endemic to the Tibetan Plateau: implications for its evolution and conservation. BMC Evol Biol 9:290
64. Zhang YJ, Zhang S, Liu XZ, Wen HA, Wang M (2010) A simple method of genomic DNA extraction suitable for analysis of bulk fungal strains. Lett Appl Microbiol 51(1):114–118
65. Zhang YJ, Zhang S, Wang M, Bai FY, Liu XZ (2010) High diversity of the fungal community structure in naturally-occurring *Ophiocordyceps sinensis*. PLoS ONE 5(12):e15570
66. Zhong X, Peng QY, Qi LL, Lei W, Liu X (2010) rDNA-targeted PCR primers and FISH probe in the detection of *Ophiocordyceps sinensis* hyphae and conidia. J Microbiol Meth 83(2):188–193
67. Zhu JS, Gao L, Li XH, Yao YS, Zhao JQ (2010) Maturational alteration of oppositely orientated rDNA and differential proliferation of GC- and AT-biased genotypes of *Ophiocordyceps sinensis* and *Paecilomyces hepiali* in natural *Cordyceps sinensis*. Am J Biomed Sci 2(3):217–238
68. Zhu JS, Halpern GM, Jones K (1998) The scientific rediscovery of an ancient Chinese herbal medicine: *Cordyceps sinensis* Part I. J Altern Complement Med 4(3):289–303
69. Zhu JS, Zhao JG, Gao L, Li XH, Zhao JQ, Lu JH (2012) Dynamically altered expressions of at least 6 *Ophiocordyceps sinensis* mutants in the stroma of *Cordyceps sinensis*. J Fungal Res 10(2):100–112