Supplemental Information: The Chinese giant salamander exemplifies the hidden extinction of cryptic species

Fang Yan, Jingcai Lü, Baolin Zhang, Zhiyong Yuan, Haipeng Zhao, Song Huang, Gang Wei, Xue Mi, Dahu Zou, Wei Xu, Shu Chen, Jie Wang, Feng Xie, Minyao Wu, Hanbin Xiao, Zhiqiang Liang, Jieqiong Jin, Shifang Wu, CunShuan Xu, Benjamin Tapley, Samuel T. Turvey, Theodore J. Papenfuss, Andrew A. Cunningham, Robert W. Murphy, Yaping Zhang and Jing Che

SUPPLEMENTAL METHODS

Sample collection. A total of 1104 samples of Chinese giant salamanders (CGS), 70 wild-caught and 1034 farm-bred, were collected. Wild-caught individuals were collected before 2010 from 15 localities at which no captive CGS had been previously released (Fig. 1B; Table S1). Most tissue samples consisted of exfoliating skin. Some samples consisted of skeletal muscle or liver collected from dead individuals. From the farm-bred CGSs, we collected buccal swab samples from 35 farms between 2014 and 2016 (Fig. 1C). The Kunming Institute of Zoology Animal Care and Ethics Committee approved all protocols (SYDW–2008019). The field work and buccal swabbing of farmed salamanders was approved by the Zoological Society of London ethics committee (WLE569).

Laboratory protocols. Total genomic DNA was isolated from tissue samples using the standard three-step phenol-chloroform extraction method. For buccal swab samples, a TIANamp Swab DNA Kit (Tiangen, Beijing, China) was used for total

DNA extraction. Mitochondrial DNA (mtDNA) fragments of genes encoding cytochrome oxidase subunit I (*COI*), cytochrome *b* (*Cytb*) and the displacement-loop (D-loop) were amplified for all 70 wild-caught individuals. For all 1034 buccal swab samples, mostly only *COI* was amplified for quick assignment to mtDNA haplotype clades A-E, or U1-U2, which correspond to cryptic species.

Given the huge genome size of Chinese giant salamander (50Gb), the reduced-representation genome sequencing (RRGS) method was used to assess genome-level data [S1]. Twenty-one samples from wild-caught individuals with genome-quality DNA were sequenced for nuDNA assessment via Specific Locus Amplified Fragment sequencing (SLAF-seq). Laboratory work was performed using the protocol previously described [S1]. Reduced complexity libraries were created with genomic DNA using EcoRV-HF restriction enzyme digestions.

One hundred farm-bred individuals from Guizhou were genotyped for 12 microsatellite DNA loci. Genotyping was conducted using an ABI 3730 DNA Analyzer. Allele sizes were determined using GeneMapper ID v3.2 (Applied Biosystems).

Phylogenetic analyses. We proofread and assembled each mitochondrial sequence using DNAStar v5.0, then aligned, edited and trimmed nucleotide sequences using MEGA 5 [S2]. Unique haplotypes of *COI*, *Cytb* and D-loop were deposited in GenBank (accession numbers: MH051336–MH051555, Table S1). The sequences of outgroup species were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/).

After concatenating the three mtDNA fragments, we explored the phylogeny of all wild-caught CGSs via Bayesian inference (BI) using MRBAYES v3.1.2 [S3]. The best nucleotide substitution model for D-loop and each codon position of *COI* and *Cytb* was determined based on the Akaike information criterion (AIC). Using four independent runs for BI, the Markov chains were estimated for 10 million generations. We sampled every 1000th generation and deleted the first 25% of samples as burn-in.

The mitochondrial phylogeny of farm bred CGSs was hypothesized using BI based on the inclusion of the wild-caught salamanders and all three mtDNA genes for U1 and U2. The other farmed samples were sequenced for *COI* only to assign them to their hypothesized cryptic species and determine the proportion of those species in farms. Tree construction was as before.

IBD analysis and timing. The relationship between genetic distance and geographic distance was calculated based on concatenated *COI*, *Cytb* and D-loop sequences. Two analyzes, separately for wild-caught salamanders from Yellow River and Yangtze River, were conducted by IBDWS v3.23 [S4].

Time since divergence of the cryptic species was estimated based on the concatenated mtDNA sequences with BEAST v1.8.0 [S5]. The strict clock model was selected by performing the marginal likelihood estimation (MLE). A Yule process was employed for speciation events. A biogeographic event, the isolation of Japan from mainland continental Eurasia at 16 Ma [S6], was imposed on the split between Chinese and Japan giant salamanders. Analyses were undertaken with 10,000,000

generations while sampling every 1000th tree, and the first 25% sampled trees were treated as burn-in.

Genetic structure of native Chinese giant salamanders. The quality of the raw SLAF sequences was checked by using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). After removing the restriction enzyme cutting sites, we used the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to discard sequences containing one or more bases with a Phred quality score below 10, or when more than 50% of the positions had a Phred quality score below 30. All sequences were end-trimmed to 100bp because of low sequence qualities at position 101 to 125. To discover SNP markers for phylogenetic analysis, we utilised the STACKS pipeline [S7] and applied it to the forward-sequence data of each sample. Each stack was required to contain at least five sequences, and the maximum distance between stacks within a locus was set to the default value (–m 5 and –M 2). The removal and deleveraging algorithms of USTACKS were invoked to delete highly repetitive and over-merged stacks, respectively. Homologous loci were identified using CSTACKS.

The consensus loci constructed in CSTACKS were used to map quality-filtered reads from each individual to these loci sequences using BWA [S8] with default parameters. Raw SNPs were determined using SAMTOOLS [S9]. To obtain high quality genotype calls for downstream analyses, we kept the SNPs that met the following criteria: 1) all analysed sites were at least 6bp away from a predicted insertion/deletion; 2) depth ranged from 2.5% to 97.5% in depth-quartile; 3) the

consensus quality was ≥ 40; and 4) SNPs that occurred at least 80% of the individuals. We excluded 1) triallelic alleles and indels, 2) SNPs with a minor allele frequency ≥ 0.01, and 3) variants that showed highly significant deviation from Hardy-Weinberg Equilibrium ($P<0.001$).

Genotypic clustering was performed using STRUCTURE v2.2 [S10]. Analyses were applied to wild-caught salamanders. To exclude potential linkage effects, we randomly used one SNP for each locus for the analyses. The number of genotypic groups (K) was set from 1 to 10. A total of 10000 MCMC repetitions of burn-in and 50000 MCMC subsequent replicates were employed using the default settings. The optimum value of K was determined by examining the change in *Ln P(D)* and using the Delta K approach [S11].

We used a Principal Coordinate Analysis (PCoA) to visualize population structure based on the genomic SNP dataset using GCAT v1.24.7 [S12].

Bayesian species delimitation. To test if the CGS consisted of more than one species according to population structure analyses, we used the Bayesian species delimitation (BSD) method [S13]. BSD was conducted using the program BPP v3.1 [S14] based on 500 randomly selected SNP loci. Five analyses used five randomly compiled datasets to confirm consistency. Each analysis consisted of 100000 MCMC generations sampled every 5^{th} generation, and the first 20% of samplings were discarded as burn-in.

Admixture analyses for farm-bred CGSs. Genetically admixed individuals were identified using STRUCTURE v2.2. Simulations were repeated 10 times. We used

500,000 Markov chain Monte Carlo (MCMC) generations and 50,000 generations of burn-in for each value of K between 1 and 10. The mixed ancestry model was assumed and allele frequencies among populations were correlated. The best K value was defined using Delta K. In the event that the best value was K = 1, we chose to display the second best value to depict the distribution of variation.

SUPPLEMENTAL RESULTS

Genetic analyses suggest that wild-caught CGSs consist of multiple species. Bayesian inference analyses of concatenated mtDNA sequences identify seven major haplotype clades, five of which relate unambiguously to wild-caught populations (A–E, Fig. 1A, B) whose distributions associate with geography. Haplotype clade A is from the Pearl River of Guangxi (Maoershan). Widespread haplotype clades B and C mainly associate with the Yellow River. Haplotype clade D from Chongqing and Guizhou occurs in the Yangtze River drainage. In the Qian Tang River drainage, haplotype clade E is from Anhui, which corresponds to the Huangshan population [S12]. In addition, haplotype clades U1 and U2 are known only from farms, as well as one sequence in GenBank from a CGS in Japan. Based on 23,159 SNPs for the 21 wild-caught individuals, both Bayesian clustering analysis and principal coordinate analysis (PCoA) identify five pure distinctive groups (Fig. S1; best K = 5). These groups largely correspond with haplotype clades A–E. Bayesian species delimitation of the SNPs resolves these groups as five species (BPPs = 1.00 for all species). Genetic differentiation within species often correlates with geographic distance,

otherwise known as isolation-by-distance (IBD). Our analyses do not obtain significant IBD within either the Yangtze River (P = 0.0737) or Yellow River (P = 0.1084) drainages, suggesting that gene flow is not ongoing. Therefore, either alone or taken together, these analyses reject the null hypothesis of conspecificity.

Populations grouped to A-E, U1, and U2 represent evolutionary species, or at least historically they did so. This is consistent with their divergence times, which range from 4.71 to 10.25 Ma. Small sample sizes in museums and their conserved morphology challenge morphological assessments; even the diagnosis of Chinese and Japanese giant salamanders is often problematic. Considering the native populations (A–E), the two groups with no specific localities (U1, U2), and the Tibetan Plateau population [S14], Chinese giant salamanders appear to have once consisted of up to eight species, and possibly more.

Genetic analyses based on microsatellite data for farm bred individuals of these salamanders reveal a disturbing vision that contrasts with nature (Fig. 1C). Most of the 1034 farm-bred individuals tested (78.82%) share mitochondrial hayplotypes of species B from the Yellow River. This species occurs ubiquitously in farms across China and is the only one detected in 12 of 34 farms surveyed. Species A occurs in a single farm near Maoershan; all other matrilines exist in multiple, widespread venues. STRUCTURE analyses of microsatellites for farms in Guizhou only, which had sufficient sample sizes for analyses, show broad genetic mixing. Suboptimal K = 3 (Fig. 1C) fails to obtain a genetically pure individual among the 100 samples tested and displays the extent of genetic mixing.

AUTHOR CONTRIBUTIONS

Designed the research, JC, YPZ, RWM and AAC; performed the molecular work, FY, JCL, ZYY, XM, and WX; analyzed the data, FY and BLZ; performed the field work and collected samples, FY, JCL, HPZ, SH, GW, DHZ, SC, MYW, JW, FX, MYW, HBX, ZQL, JQJ, SFW, CSX, BT, STT, TJP, RWM and JC; prepared the paper FY, RWM, BLZ, JC, YPZ, and AAC. All authors reviewed the manuscript.

DECLATION OF INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENTS

Shaoneng Wang, Peng Guo, Zhigang Qiao, Xuefu He, Xiaoming Wang, Bin Li, Jingcheng Xu, Zuogang Peng, and staffs of local reserves/Forestry Bureaus helped with sampling. Todd W. Pierson, Weiwei Zhou, Minsheng Peng kindly provided thoughtful comments. Thanks for the suggestions from the two anonymous reviewers. JC is in debt to the continuous encouragement and discussions from David B. Wake on study of the Chinese salamanders. This work is supported by the programs of the Strategic Priority Research Program, CAS (XDPB020406), State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, CAS (GREKF15-10), National Natural Science Foundation of China (NSFC 31090250, 31360144), Darwin Initiative (Project No. 19-003) and EDGE programme from Zoological Society of London, Ocean Park Conservation Foundation of Hong Kong,

China's Biodiversity Observation Network CAS, Animal Branch of the Germplasm Bank of Wild Species, CAS (Large Research Infrastructure Funding), and Natural Science Foundation of Anhui Provincial Bureau of Education (KJ2014A244). JC is supported by the NSFC (31622052) and the Youth Innovation Promotion Association CAS. FY is supported by the NSFC (31401958) and the West Light Project of CAS. RWM is supported by the CAS President's International Fellowship Initiative (PIFI) and the ROM foundation.

SUPPLEMENTAL REFERENCES

S1. Sun, X. *et al*. (2013) SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. PLoS One *8*, e58700.

S2. Tamura, K. *et al*. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739.

S3. Ronquist, F., and Huelsenbeck, J.P. (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572–1574.

S4. Jensen, J.L., Bohonak, A.J., and Kelley, S.T. (2005) Isolation by distance, web service. BMC Genet. *6*, 13.

S5. Drummond, A.J., and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*, 214.

S6. Isozaki, Y., Aoki, K., Nakama, T., and Yanai, S. (2010) New insight into a subduction-related orogen: a reappraisal of the geotectonic framework and evolution of the Japanese Islands. Gondwana Res. *18(1)*, 82–105.

S7. Catchen, J.M. *et al*. (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3 Genes, Genomes, Genetics *1(3)*, 17.

S8. Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Oxford University Press.

S9. Li, H. *et al*. (2009) The sequence alignment/map format and SAMtools. Bioinformatics *25(16)*, 2078–2079.

S10. Pritchard, J.K., Stephens, P., and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

S11. Evanno, G., Regnaut, S., and Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. *14*, 2611–2620.

S12. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011) Gcta: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88(1)*, 76.

S13. Yang, Z., and Rannala, B. (2010) Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA *107*, 9264–9269.

S14. Yang, Z. (2015) The BPP program for species tree estimation and species delimitation. Curr. Zool. *61*, 854–865.

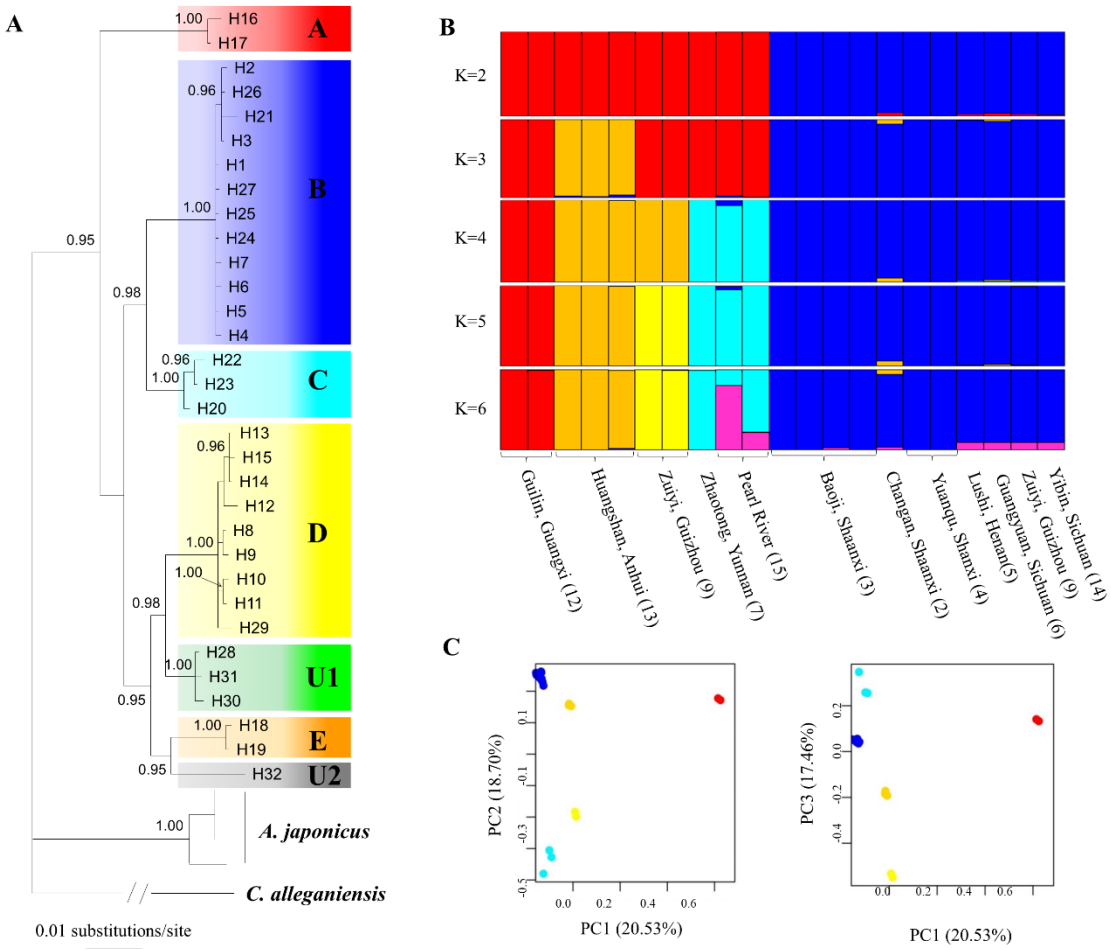

Figure S1 Chinese giant salamander phylogeny based on mitochondrial DNA haplotype**s and the genetic structure of wild-caught Chinese giant salamanders. A)** Bayesian inference tree based on concatenated *Cytb*, *COI* and D-loop for all wild-caught and four farm-bred CGSs. Numbers near branches are posterior probabilities (BPP≥0.90). Haplotype information is detailed in Table S1. **B**) Clusters obtained using STRUCTURE for K = 2 to K = 6 based on SNPs for wild-caught CGSs. The colours represent different genetic clusters. At optimal K=5, the genetic clusters corresponded to mtDNA haplotype clades A-E in part A. Population localities and numbers are shown below the lowest chart. **C)** PCoA plot based on SNPs for wild-caught CGSs. Colours correspond with the STRUCTURE analysis.

Table S1 Summary of sample localities and genetic information details for wild-caught Chinese giant salamanders, four farm-bred individuals and outgroups. Locality numbers correspond to Figure 1B. Asterisks (*) denote individuals successfully genotyped by Reduced-Representation Genome Sequencing.

Species	Localitie of native CGS	Tissue No.	Haplotype No.	Mitochondrial lineage	GenBank Accession No.			Remarks
					Cytb	D-loop	*COI*	
A. davidianus								
	Fengxian, Shannxi (1)	KIZ020235	Hap1	B	MH051410	MH051482	MH051336	
	Changan, Shannxi (2)	11041*	Hap2	B	-	MH051483	MH051337	
	Baoji, Shannxi (3)	KIZYPX14528	Hap3	B	MH051411	MH051484	MH051338	
		KIZYPX14529	Hap4	B	MH051412	MH051485	MH051339	
		KIZYPX14530*	Hap5	B	MH051413	MH051486	MH051340	
		KIZYPX14531*	Hap3	B	MH051414	MH051487	MH051341	
		KIZYPX14532*	Hap6	B	MH051415	MH051488	MH051342	
		KIZYPX14533	Hap3	B	MH051416	MH051489	MH051343	
		KIZYPX14534	Hap5	B	MH051417	MH051490	MH051344	
		KIZYPX14535	Hap4	B	MH051418	MH051491	MH051345	
		KIZYPX14537*	Hap3	B	MH051419	MH051492	MH051346	
		KIZYPX14538	Hap6	B	MH051420	MH051493	MH051347	

Yuanqu, Shanxi (4)	11051*	Hap7	B	MH051421	MH051494	MH051348	
	11052	Hap7	B	MH051422	MH051495	MH051349	
	11053*	Hap7	B	MH051423	MH051496	MH051350	
Lushi, Henan (5)	KIZYPX44113*	Hap7	B	MH051424	MH051497	MH051351	
Qingchuan, Guangyuan, Sichuan (6)	KIZ020236*	Hap5	B	MH051425	MH051498	MH051352	
	KIZYPX25999	Hap24	B	MH051426	MH051499	MH051353	
	KIZYPX25990	Hap22	C	MH051427	MH051500	MH051354	
	KIZYPX25991	Hap23	C	MH051428	MH051501	MH051355	
Zhaotong, Yunnan (7)	KIZYPX26845*	Hap5	B	MH051429	MH051502	MH051356	mitochondrial introgression
Xinglong, Chongqing (8)	KIZYPX2505	Hap8	D	MH051430	MH051503	MH051357	
	KIZYPX2506	Hap8	D	MH051431	MH051504	MH051358	
	KIZYPX2507	Hap8	D	MH051432	MH051505	MH051359	
	KIZYPX2508	Hap8	D	MH051433	MH051506	MH051360	
	KIZYPX2509	Hap8	D	MH051434	MH051507	MH051361	
	KIZYPX2513	Hap9	D	MH051435	MH051508	MH051362	
	KIZYPX2514	Hap9	D	MH051436	MH051509	MH051363	
	KIZYPX2515	Hap9	D	MH051437	MH051510	MH051364	
	KIZYPX2516	Hap9	D	MH051438	MH051511	MH051365	
	KIZYPX2517	Hap8	D	MH051439	MH051512	MH051366	
	KIZYPX2518	Hap9	D	MH051440	MH051513	MH051367	
	KIZYPX2519	Hap9	D	MH051441	MH051514	MH051368	
Zhengan, Zunyi,	KIZZA2*	Hap10	D	MH051442	MH051515	MH051369	

Guizhou (9)							
	KIZZA9*	Hap11	D	MH051443	MH051516	MH051370	
	KIZZA4*	Hap7	B	MH051444	MH051517	MH051371	translocation from localities 4 or 5
Leishan, Guizhou (10)	KIZYPX10535	Hap12	D	MH051445	MH051518	MH051372	
Guiding, Guizhou (11)	KIZYPX10518	Hap13	D	MH051446	MH051519	MH051373	
	KIZYPX10519	Hap13	D	MH051447	MH051520	MH051374	
	KIZYPX10521	Hap13	D	MH051448	MH051521	MH051375	
	KIZYPX10522	Hap14	D	MH051449	MH051522	MH051376	
	KIZYPX10523	Hap13	D	MH051450	MH051523	MH051377	
	KIZYPX10524	Hap13	D	MH051451	MH051524	MH051378	
	KIZYPX10525	Hap15	D	MH051452	MH051525	MH051379	
	KIZYPX10526	Hap13	D	MH051453	MH051526	MH051380	
	KIZYPX10527	Hap13	D	MH051454	MH051527	MH051381	
	KIZYPX10528	Hap13	D	MH051455	MH051528	MH051382	
	KIZYPX10529	Hap13	D	MH051456	MH051529	MH051383	
	KIZYPX10530	Hap13	D	MH051457	MH051530	MH051384	
	KIZYPX10531	Hap13	D	MH051458	MH051531	MH051385	
	KIZYPX10532	Hap13	D	MH051459	MH051532	MH051386	
	KIZYPX10533	Hap13	D	MH051460	MH051533	MH051387	
Maoershan, Guilin, Guangxi (12)	KIZYPX10536	Hap16	A	MH051461	MH051534	MH051388	
	KIZGXDN3	Hap17	A	MH051462	MH051535	MH051389	
	KIZGXDN4	Hap17	A	MH051463	MH051536	MH051390	
	KIZGXDN5	Hap16	A	MH051464	MH051537	MH051391	
	KIZ020273*	Hap16	A	MH051465	MH051538	MH051392	

	KIZ022435*	Hap17	A	MH051466	MH051539	MH051393	
	KIZ020272	Hap17	A	MH051467	MH051540	MH051394	
Huangshan, Anhui (13)	11036*	Hap18	E	MH051468	MH051541	MH051395	
	11037*	Hap18	E	MH051469	MH051542	MH051396	
	11038	Hap18	E	MH051470	MH051543	MH051397	
	11039*	Hap18	E	MH051471	MH051544	MH051398	
	11040*	Hap18	E	MH051472	MH051545	MH051399	
	KIZYPX6151	Hap19	E	MH051473	MH051546	MH051400	
	KIZYPX6152	Hap18	E	MH051474	MH051547	MH051401	
	KIZYPX6153	Hap18	E	MH051475	MH051548	MH051402	
Yibin, Sichuan (14)	KIZ014338*	Hap25	B	MH051476	MH051549	MH051403	translocation from locality 1
Pearl River (15)	11046*	Hap20	C	-	MH051550	MH051404	
	11047*	Hap21	B	MH051477	MH051551	MH051405	mitochondrial introgression
Unknown (GenBank)		Hap26	B	AB445782	AB445800	-	
		Hap27	B	AB445783	AB445801	-	
		Hap28	U1	AB445784	AB445802	-	
		Hap29	D	NC004926	NC004926	-	
Farm-bred (Guangxi, 17)	CGS1009	Hap30	U1	MH051478	MH051552	MH051406	
Farm-bred (Guangxi, 18)	CGS947	Hap32	U2	MH051479	MH051553	MH051407	
Farm-bred (Guizhou, 24)	CGS725	Hap31	U1	MH051480	MH051554	MH051408	
Farm-bred (Jiangxi, 34)	CGS291	Hap32	U2	MH051481	MH051555	MH051409	

outgroup

A. japonicus	AB445781	AB445799	-
	AB445776	AB445794	-
	AB445780	AB445798	-
	AB208679	AB208679	AB208679
C. alleganiensis	AB445785	AB445803	KU985766