Genome size estimation of Chinese cultured artemisia *annua* L.

Zhixiang Liu[1]; Shuai Guo[1,2]; Jiang Xu[1]*; Yujun Zhang[1]; Linlin Dong[1]; Shuiming Xiao[1]; Rui Bai[3]; Baosheng Liao[1]; He Su[1,4]; Ruiyang Cheng[1]; Shilin Chen[1]

[1]Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, 100700, China

[2]College of Agriculture/College of Peony, Henan University of Science and Technology, Luoyang, Henan 471023, China

[3]College of Pharmacy and Chemistry, Dali University, Dali, Yunan 671000, China

[4]Guangdong Provincial Hospital of Chinese Medicine, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, and China Academy of Chinese Medical Sciences Guangdong Branch, China Academy of Chinese Medical Sciences, Guangzhou, 510006, China

*Corresponding Author(s): Jiang Xu,

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, 100700, China

Email: jxu@icmm.ac.cn

Abbreviations: FCM: flow cytometry; CV: coefficient of variation; HPCV: half peak coefficient of variation; NJ tree: neighbor joining tree; PI: propidium iodide; FSC: forward scatter; SSC: side scatter

Abstract

Almost all of antimalarial artemisinin is extracted from the traditional Chinese medicinal plant Artemisia *annua* L. However, under the condition of insufficient genomic information and unresolved genetic backgrounds, regulatory mechanism of artemisinin biosynthetic pathway has not yet been clear. The genome size of genuine *A. annua* plants is an especially important and fundamental parameter, which helpful for further insight into genomic studies of artemisinin biosynthesis and improvement. In current study, all those genome sizes of *A. annua* samples collected with Barcoding identification were evaluated to be 1.38-1.49 Gb by Flow Cytometry (FCM) with Nipponbare as the benchmark calibration standard and soybean and maize as two internal standards individually and simultaneously. The genome estimation of seven *A. annua* strains came from five China provinces (Shandong, Hunan, Chongqing, Sichuan, and Hainan) with a low coefficient of variation (CV, ≤ 2.96%) wasrelative accurate, 12.87% (220 Mb) less than previous reports about a foreign A. annuaspecies with a single control. It facilitated the schedule of *A. annua* whole genome sequencing project, optimization of assembly methods and insight into its subsequent genetics and evolution.

Introduction

Artemisiae annuae herba, the dried aerial part of the annual herbaceous plant Artemisia *annua* L. [1], characteristically synthesizes and accumulates the unique sesquiterpene endo peroxide lactone the antimalarial drug artemisinin. Artemisi-

Cite this article: Liu Z, Guo S, Xu J, Zhang Y, Dong L, et al. Genome size estimation of Chinese cultured artemisia *annua* L. J Plant Biol Crop Res. 2018; 1: 1002.

nin-based combination therapies (ACTs) is recommended to be the best choice for quick and reliable treating acute malaria by WHO [2-4]. What's more, the artemisinin isolation enabled the inventor Professor Tu to receive the 2011 Lasker DeBakey Clinical Medical Research Award and the 2015 Nobel Prize in Physiology or Medicine [5]. Moreover, antimalarial artemisinin was confirmed to have other multifunctions, such as anticancer [6,7], antiviral [8,9], and antischistosomal activities [10].

As still the main source of artemisinin *A. annua* plant is cosmopolitan species in the world (such as in Viet Nam and India), but most widespread in each province of China with the artemisinin content ranging from 0.1%~1.5% dried leaf weight, which affected by ecological environment and varietal difference [11]. However, the *A. annua* strains that contain less than 0.5% artemisinin content could not be used as raw material for artemisinin, especially the strains in Northern China, Viet Nam and India (less than 0.1% artemisinin content) [11]. It's urgent to increase the artemisinin yield by numerous attempts, focusing on genetic modification and bioengineering of artemisinin biosynthesis in plants during the last two decades [5,12,13]. Unfortunately, under the condition of insufficient genomic information and genetic backgrounds, regulatory mechanism of artemisinin biosynthetic pathway has not yet been clarified. Also, artemisinin can be semi synthesized via artemisinic acid or dihydroartemisinic acid feasibly obtained from genetically modified yeast [14-17], but it cannot far reach high commercial values. Fundamentally, it seems be especially imperative to research accurate evaluation of Chinese *A. annua* genome size for its subsequent genetics study.

What's more, it is well known that diversity lines have significantly different genetics and genome sizes in many plant species [18-22], especially just as rapid genome size change in *Arabidopsis* (*A. thaliana*) [23-27], rice (*Oryza sativa* L.) [28-33], maize (*Zea mays*) [34] and soybean (*Glycine max*) [35]. The size flexibility of the Zeagenome varies from 4.92 to 6.87 pg/2C [34] in wild and cultivated maize, which related with geography and altitudinal gradients [34]. The maize luxurians genome was ~1.5-fold larger than that of B73 [36], while there were very different karyotypes between the domesticated subspecies maize (*Z. mays ssp. mays*) and two prominent wild subspecies (*Z. mays ssp. Parviglumis,* and *Z. mays ssp. Mexicana*) [34,36,37]. Moreover, the estimated Soybean genome sizes ranges from 889.33 Mbp to 1118.34 Mbp [38-42]. The flow cytometric data of different maize varieties were slightly inconsistent using the same internal standard, while flow cytometric results of the same soybean and maize varieties with different internal references were still diverse [33,43-48]. Therefore, it is necessary to analysis genome size variation in Chinese A. annua populations with different internal references.

There upon, seven wild *A. annua* sample strains identified by DNA barcoding were chosen in China five different areas (Shandong, Hunan, Chongqing, Sichuan, and Hainan). Then those genome sizes were estimated by Flow Cytometry (FCM) with Nipponbareas a benchmark calibration standard, having a relatively distinct genome. And considering that both of soybean and maize showed appropriately closed genome sizes with *A. annua* (Soybean was 889.33 – 1118.34 Mbp [38-42], maize was 2300–3360 Mbp [34,49], *A. annua* maybe 1710 Mbp [50]), they were adopted as two typical internal standards in FCM. The accurate genome size could facilitate the schedule of *A. annua* whole genome sequencing project and may be helpful to give further insight into artemisinin improvement genomic studies.

Results

Identification of sample strains with *A. annua* and other species

Seven *ITS2* and seven *psbA-trnH* sequences were obtained from seven collected wild samples. Based on *ITS2* and *psbA-trnH* sequences of seven wild samples, *A. annua*, its closely related species and counterfeits, two Neighbor Joining (NJ) trees were constructed separately. All the *ITS2* and *psbA-trnH* control sequences of *A. annua* and its adulterants were generated from our previous studies [51,52] (**Supplementary Table S1**).

The sequence length, GC content, and K2P genetic distance of the *ITS2* and *psbA-trnH* regions of samples, *A. annua* and its adulterants were analyzed and summarized (**Supplementary Table S2**). The *ITS2* sequence length of sample strains gathered from five provinces was 225 bp, while the *psbA-trnH* sequence length was 353 bp. No variable site existed both in those *ITS2* (**Figure 1b**) and *psbA-trnH* sequences of *A.annua* in 5 provinces. The average GC content of *ITS2* sequences of *A.annua* was 56.40%, and that of *psbA-trnH* sequences was 25.20%. On the basis of the *ITS2* and *psbA-trnH* sequences, the intraspecific divergence of *A.annua* calculated using the K2P model was zero, which was far lower than the minimum interspecific distance of *A.annua* and 23 other closely related species and counterfeits (**Supplementary Table S2**).

All our *ITS2* sequences were in accord with *A. annua* (**Figure 1b**), separating from other closely related species in neighbor joining (NJ) NJ trees (**Figure 1a**). In agreement with our previous studies [51,52], *A. annua* and the closely related species and counterfeits could be distinguished from each other on the basis of the *ITS2* sequences (not *psbA-trnH*). Thus, all strains were identified as genuine *A. annua* with no variation of their *ITS2* and *psbA-trnH* sequences

A. annua genome size analysis by FCM

In our study, the peaks of Nipponbare, soybean, maize, and seven *A. annua* strains were alone determined. With no peak overlap, it made sure that the peaks of *A. annua* strains were well separated from internal standards. Using high-quality sequenced Nipponbare [53] as a benchmark calibration standard, the mixed Nipponbare with soybean, and the mixed Nipponbare with maize were detected three times, separately. And according to the formula: sample 2C DNA content = sample peak mean/standard peak mean standard 2C DNA content, the genome sizes of soybean and maize were measured as 0.92 ± 0.00 Gb and 2.17 ± 0.02 Gb, respectively. Then, the mixed internal standards and per *A. annua* strain (such as *A. annua* mixed with soybean, *A. annua* with maize, *A. annua* with soybean and maize) were measured with three technical replicates, parallelly twice. The flow cytometric results of seven *A. annua* strains collected from five provinces were analysed and accounted in **Figure 2** and **Table 1**. The values of *A. annua* genome sizes were assessed ranging from 1.31 Gb to 1.54 Gb, referred to Nipponbare as primitively control, and soybean and maize as control individually and simultaneously.

Through the flow cytometric data analysis of two parallel per strain detected three repeatedly with three different controls in individual and simultaneous ways, the maximum coefficient of variation values (CV) was detected to be 2.96% in the group data of HK strain with *Z. Mays* as a control. Using the same control, the differences among genome sizes of 7 *A. annua* strains were 45 Mb (Nipponbare), 88 Mb (*G. max*) and 41 Mb (*Z. Mays*),

respectively. And the CV of seven strains with Nipponbare as control ranged from 0.29% to 2.94%, that of *G. max* was 0.67-2.66%, and that of *Z. Mays* was 0.86-2.96%. This suggested that there was a quite stability of the flow cytometric instrument (BD AccuriTM C6).

However, *G. max* and *Z. Mays* genome sizes assessed by FCM using Nipponbare as primitively control were slightly larger than that in previous reports [49,54]. Various varieties and their different growth conditions may be the key points [55]. From those high-quality data with low CV ≤ 2.96%, seven strains detections had little differences using Nipponbare, *G. max* and *Z. Mays* as control, respectively. It also illustrated that diversity lines in *A. annua* species had little influence on their genome sizes.

The genome sizes variation of the same *A. annua* strain using the three different control ranged from 38 Mb (the HK strain) to 96 Mb (the SC strain). And the difference sizes among 5 other strains genome were close (57 Mb for LQ strain, 55 Mb for YJ strain, 64 Mb for YY-1 strain, 67 Mb for YY-2 strain, and 67 Mb for JX strain). Those flow cytometric results indicated that the estimation for the same *A. annua* species with different standards had little diversity.

The biggest genome size of *A. annua* YY-2 strain was 1.49 ± 0.07 Gb, whereas the smallest of YY-1 strain was 1.38 ± 0.06 Gb by a margin of 110Mb (**Table 1**). The result of all flow cytometric data of *A. annua* strains and their mean value (1.44 Gb) differ by 73 Mb, showing that there was slight genome size variation in *A. annua* species.

Discussion

Estimation of A. annuagenome size

Originally, the genome size of *A. annua* estimated by Nagl and Ehrendorfer (in 1974) [56], Geber and Hasibeder (in 1980) [57] was respectively 4.1 pg and 3.8 pg, with microdensitometry measurements based on Feulgen staining. Using *Pisumsativum* cv. Express long (2C = 8.37 pg) [58] as only one internal standard, the DNA content per haploid genome of *A. annua* (1.75 pg, 1.71 Gb) was assessed using flow cytometry (FCM) by Torrell and Valles [50]. In previous studies, *A. annua* samples had been deposited in the Herbarium of the Laboratory of Botany, Faculty of Pharmacy, university of Barcelona (BCF) since 1997, lacking of study on molecular identification. The half peak coefficient of variation (HPCV) of *A. annua* reached to 3.01 and its DNA amount value was very different from other taxa.

In this paper, seven *A. annua* strains collected from five provinces were identified on the basis of the *ITS2* sequences. And their flow cytometric analysis with a low CV (≤2.96%) indicated that their genome sizes range from 1.38 Gb to1.49 Gb, using Nipponbare as the benchmark calibration standard and *G. max* and *Z. Mays* as two internal standards individually and simultaneously. However, the biggest size (1.49 Gb) of the YY-2 strain genome was 12.87% (220 Mb) less than the estimation result (1.71 Gb) of *A. annua* in Spain by Torrell and Valles with *Pisumsativum* as an internal standard [50]. It was resulted from that there was absence of significant genome size variation in *Pisumsativum* [59,60], which was unfitness for an internal standard without proofreading by a benchmark calibration standard in FCM. So it does require a benchmark calibration standard in the method of FCM.

Considering the same area, the genome size of YY-1 stain was less about 110Mb than YY-2 while JX was less 54 Mb than

HK. Both of YY-1 and YY-2 had little difference with other area strains. Moreover, the variation between all seven strains' genome sizes and their mean value (1.44 Gb) was merely 73 Mb, showing no significant discrepancy. It indicated that there was minor genome size variation in *A. annua* species, which may predominantly result from transposable element accumulation, expansion/contraction of tandem repeats, variation in intron length and so on [61].

In addition, we have attempted to carry out genome-wide survey with low-depth (<30X) high-throughput sequencing data. The estimation value was a little larger than flow cytometric data (unreported) and *A. annua* genome was rich in high repeat content sequences. It was assessed in conjunction with *Dendrobium officinale* (a traditional Chinese Orchid herb), whose genome size is about 1.27 Gb based on flow cytometric data and 1.35 Gb assembled by combining the second-generation and third-generation PacBio sequencing technologies [62]. And it also accords with the size of *Eriobotrya* Lindl. 'Jiefangzhong' genome (654.40 Mb estimated by FCM and 773.00 Mb by 17-mer spectrum) [63]. These inflated genome sizes attribute to their high repeat content and heterozygosity [36]. However, flow cytomeric data with internal standards *Caenorhabditis elegans* (~100 Mb) and *Drosophila melanogaster* (~175 Mb) showed the genome size of the first sequenced *A. thaliana* (157 Mb) was 25% larger than that initiative estimate of 125 Mb, partially was set down to genes mission in centromeric and ribosomal DNA regions [64]. Though the discrepancy between those two sets of data also exists in many sequenced plant genomes, both can determine the same order of magnitude of plant genome size and flow cytometric results were quite credible.

Hence, the unreported k-mer analysis results confirmed that *A. annua* genome was closed to 1.38-1.49 Gb, having a complex genome with high repeat content.

Performance of FCM for *A. annua* strains

At present, Feulgen Spectrophotometry and FCM are commonly used methods for estimating genome size. FCM is a powerful method in qualitative and quantitative analysis of animal, botanical, microbial monoplast, and other microscopic particles in liquid suspension [65,66]. Served as a traditional standard technology for estimating genome size, FCM can confirm nuclear DNA content exactly [67,68]. Bennett et al appealed to provide a precise angiosperm C-value served as a benchmark calibration standard for plant genome estimations [64]. So far, *Arabidopis* (*A. thaliana*) [69] and rice (*O. sativa* L.) [28-32] are two high-quality sequenced plants. The sequences of rice genome is considered as a "gold standard" in plants. Moreover, the Nipponbare RefSeq has the best quality information, compared to known crop genome sequences [28-32]. It is quite important for sequence comparison of herbaceous plants, and Nipponbare can be used as a benchmark calibration standard. Besides,the genomes of soybean (*G. max*) and maize (*Z. mays*) whose genome sizes are both close to that of *A. annua* L., had been estimated by FCM [43-46,48,70,71] and also sequenced [38,49]. Therefore, soybean and maize remeasured by Nipponbare are competent for well accurate determination of the *A. annua* genome size.

The method and technique of FCM are simple and convenient, extending the spectrum of its application. FCM has the advantages of great flux per a batch of operations, but it is limited to internal reference and easy to be affected by endogenous DNA and secondary metabolites. In general, high content of cy-

tosolic compounds in medicinal plant leaves containing proteins and secondary metabolites is considerably liable to bias nuclear DNA content estimations by FCM, which cannot be completely overcame [72,73]. Recently researches were mainly focused on the most appropriate buffers and procedures for sample preparation. Otto I buffer can precipitate nucleus, as well wipe off cytosolic compounds in officinal plant leaves and nucleus debris in a certain degree [74]. From our flow cytometric data, it can be found that the Otto buffer was also applicable to *A. annua* samples.

Estimates for the same species with different standards were sometimes surprisingly divergent, but not in accordance with our *A. annua* flow cytometric results. Comparisons with two results of various species genome estimations by FCM and genome survey, we discovered some sequenced model plants characteristically stability in DNA content and ease of preparation were inadequate to serve as internal standards. On condition that the flow cytomeric estimations in conformity with genome assembly results, the genome size discrepancies between Mosobamboo (*P. pubescens*) and its internal standard soybean (*G. max*) [75-77], between barley (*Hordeum vulgare*) and standard *P. sativum* [43,78] are 1.70-fold and 1.49-fold, respectively. It indicated that the optimum genome discrepancies of the uknown and internal standard should be approximately 0.5- or 2-fold [79]. However, in our study (**Table 2**), under Nipponbare as a benchmark calibration standard of *G. max* and *Z. mays*, the size of *A. annua* genome using Nipponbare (3.76-fold discrepancy), *G. max* (1.56-fold discrepancy) and *Z. mays* (0.66-fold discrepancy) respectively as control were all suitable.

All flow cytometric data of *A. annua* species (CV≤2.96%) were essentially stable, however, the estimations of different strains with the three different control were disproportionate. The results showed that the different contents of inclusion in different strains would influence the intensity of fluorescent staining in *A. annua* plants. It may be a valuable reference for quality evaluation of their different compounds in *A. annua* plants.

In virtue of high-quality sequenced Nipponbareas a benchmark calibration standard, the genome size of genuine *A. annua* identified by *ITS2* was estimated to be approx. 1.38-1.49 Gb by FCM with Nipponbare, *G. max* and *Z. mays* as different internal standards. Further more, genome size did not show significant variation in seven wild *A. annua* strains coming from five provinces. It showed that no rapid expansion and contraction in *A. annua* genome was found. So, it is necessary to conduct a further study on the relationship among environment factors, genetic information and artemisinin content variation. The assessment of *A. annua* genome size would provide a deeper understanding of its genome. It facilitate the suitable schedule of its whole genome sequencing project, and provide references for insight into its subsequent genetics and evolution.

Materials and Methods

Plant materials

In this essay (**Table 1**), seven wild sample plants or seeds were gathered from five provinces (Shandong, Hunan, Chongqing, Sichuan, and Hainan). Some wild seedings were transplanted into our greenhouse, and some were seedlings germinated from wild seeds in soil or Murashige and Skoog (MS) medium.

Rice (Nipponbare) seeds, whose genome size is quite definite, came from the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences. And the seeds of soy-

bean and maize were purchased from the market, uncertain varieties. Growth and development of Nipponbare, soybean and maize were carried out by the water culture.

DNA extraction and cloning DNA barcoding sequences

We adopted *ITS2* and *psbA-trnH* sequences as barcodes to identify seven samples, *A. annua*, and the others. Samples genomic DNA isolation and their *ITS2* and *psbA-trnH* sequences were obtained, assembled and analyzed according to the protocol of "Standard DNA Barcodes of Chinese Material Medica in Chinese Pharmacopoeia" [51,80].

Flow cytometric measurement

80mg fully developed fresh leaves of *A. annua* strains, Nipponbare, soybean and maize or 200mg callus and adventitious bud of cultured strains were collected in clean Petri dishes. Those samples were rapidly chopped in 2 mL cold Nuclei suspension extractions (Ottobuffer I [55,81]) with a sharp razor blade, and filtered through a 40 μm nylon cell strainer, keeping on ice. The extraction liquid was transferred in a new 2 mL tube, and centrifuged coldly for 3min at low speed (1844 g, for 5000 rpm) to remove the supernatant particles. Then the precipitate in the bottom was suspended again with 600ul fresh ice-cold Otto I solution, and centrifuged coldly for 30s at 500g, twice repeatly. Before flow cytometric analysis, staining with propidium iodide (PI, with RNase, BD Biosciences PharmingenTM, San Diego, US) [43,45,82] was performed equivalently in a mixture of Otto I and Otto II buffers (1:2) for 15min.

The nuclear DNA content measurements and analysis were carried out by a FCM (BD AccuriTM C6, USA) at a low flow rate (14μl /min) with more than 100000 cells. Forward scatter (FSC), side scatter (SSC), blue (488nm) and red (640 nm) fluorescence for PI were acquired. Two parallel with three technical replicates per sample were detected for the stability of the instrument. And mean values and standard deviations of all flow cytometric data were calculated. The formula and storage of Otto I and II solution were referred to the Dolezel's protocol [55].

Tables

Table 1: *A. annua* plants

Number	Collecting locations	Material	Geographical position
LQ	Liaocheng, Shangdong province	callus and adventitious bud	115.49°E, 36.71°N
YJ	Yuanjiang, Hunan province	tender leaves of wild seedling in soil	112.36°E,28.85°N
YY-1	Youyang, Chongqing city	tender leaves of wild seedling in soil	108.77°E,28.84°N
YY-2	Youyang, Chongqing city	tender leaves of wild seedlings in MS	108.77°E,28.84°N
SC	Yibin, Sichuan province	tender leaves of wild seedling in soil	104.64°E,28.75°N
JX	JiaCha farm, Hainan province	tender leaves of wild seedling in soil	110.35°E,20.03°N
HK	Wuzhishang, Hainan province	tender leaves of wild seedlings in MS	110.35°E,20.03°N

Table 2: Flow cytometric data of seven *A. annua* strains.

Number	A. annua genome size (Gb)			
	Nipponbare as control	Soybean as control	Maize as control	Mean value ± s[b]
	mean ± s[a]	mean ± s[a]	mean ± s[a]	
LQ	1.434±0.022	1.518±0.038	1.408±0.027	1.454±0.057
YJ	1.543±0.045	1.458±0.020	1.419±0.040	1.460±0.055
YY-1	1.486±0.008	1.331±0.036	1.375±0.020	1.380±0.064
YY-2	1.475±0.012	1.564±0.010	1.422±0.012	1.490±0.067
SC	1.416±0.004	1.311±0.021	1.508±0.044	1.411±0.096
JX	1.483±0.041	1.337±0.035	1.427±0.028	1.402±0.067
HK	1.422±0.031	1.479±0.024	1.427±0.043	1.456±0.038

Note: a: mean value ± standard deviation of two samples with three technical replicates

b: mean value ± standard deviation of all flow cytometric data per A. annuastrain

Supplementary Table S1: Sample informations of *Artemisia annua* and counterfeits

Species	Collecting locality	Sample number	GenBank accession number	
			ITS2	psbA
Artemisia annua	Hefei, Anhui province	HHH1	KU555591	KU555718
Artemisia argyi	Liuan, Anhui province	A1	KU555640	KU555764
Artemisia vexans	Kemu, Yunnan province	ZDH1	KU555710	KU555834
Artemisia sieversiana	Shengnongjia, Hubei province	A140013	KU855169	KU855309
Artemisia maximowicziana	Haerbing, Heilongjiang province	A08001	KU855160	KU855300
Artemisia carvifolia	Ningbo, Zhejiang province	QQH1	KU555619	—
Artemisia hedinii	Qinghai, Gansu province	CH1	KU555622	KU555747
Artemisia vestita	Dangchang, Gansu province	MLH1	KU555624	KU555749
Artemisia eriopoda	Wutaishan, Shanxi province	NMH2	KU555628	—
Artemisia tangutica	Guangzhou, Guangdong province	GQH1	KU555713	KU555837
Artemisia scoparia	Daofu, Sichuan province	ZMH1	KU555632	KU555756
Artemisia aurata	Kunming, Yunnan province	HJH1	KU555707	KU555831
Artemisia stolonifera	Wenshan, Yunnan province	KSH1	KU555696	KU555820
Artemisia princeps	Taiwan	A11001	KU855165	KU855305
Artemisia anomala	Nanchuan, Chongqing city	QH1	KU555703	KU555827

Artemisia igniaria	Nanyang, Henan province	QJH1	KU555715	KU555839
Artemisia indica	Guillin, Guangxi province	WYA2	KU555687	KU555811
Artemisia gmelinii	Lasa, Xizang province	XYLH1	KU555629	KU555753
Artemisia lavandulifolia	Nanning,Guangxi province	YAH1	KU555698	KU555822
Artemisia capillaris	Beijing,Beijing city	YCH1	KU555700	KU555824
Artemisia argyi var. gracilis		CXA4	KU555683	KU555807
Artemisia mongolica			FJ643026	JX073853
Artemisia lactiflora			JQ173384	JX073846
Artemisia vulgaris			KM58983	JX073849

Supplementary Table S2: Sequence characters and K2P genetic distances of samples, *A. annua* and counterfeits

Species	ITS2				psbA-trnH			
	Length	G+C content (%)	Intraspecific distance (mean)	Interspecific distance (mean)	Length	G+C content (%)	Intraspecific distance (mean)	Interspecific distance (mean)
LQ	225	54.7	0.000	0.000-0.018(0.011)	353	25.2	0.000	0.000-0.007(0.004)
YJ	225	54.7	0.000	0.000-0.018(0.011)	353	25.2	0.000	0.000-0.007(0.004)
YY-1	225	54.7	0.000	0.000-0.018(0.011)	353	25.2	0.000	0.000-0.007(0.004)
YY-2	225	54.7	0.000	0.000-0.018(0.011)	353	25.2	0.000	0.000-0.007(0.004)
SC	225	54.7	0.000	0.000-0.018(0.011)	353	25.2	0.000	0.000-0.007(0.004)
JX	225	54.7	0.000	0.000-0.018(0.011)	353	25.2	0.000	0.000-0.007(0.004)
HK	225	54.7	0.000	0.000-0.018(0.011)	353	25.2	0.000	0.000-0.007(0.004)
A. annua	225	54.7	0.000	0.000-0.075(0.023)	353	25.2	0.000	0.000-0.007(0.004)
A. argyi	225	56.4	—	0.000-0.046(0.020)	483	29.2	—	0.000-0.011(0.006)
A. lactiflora	225	55.1	—	0.000-0.023(0.014)	383	25.8	—	0.004-0.014(0.008)
A. vulgaris	221	55.2	—	0.004-0.052(0.023)	383	25.6	—	0.002-0.008(0.005)
A. vexans	225	57.3	—	0.004-0.046(0.020)	362	24.9	—	0.000-0.009(0.005)
A. argyivargracilis	225	56.4	—	0.000-0.046(0.021)	489	29.2	—	0.003-0.011(0.006)
A. sieversiana	227	55.5	—	0.000-0.042(0.022)	362	25.1	—	0.003-0.014(0.006)
A. maximovicziana	228	55.7	—	0.000-0.042(0.022)	362	24.9	—	0.003-0.017(0.008)
A. carvifolia	224	54.9	—	0.013-0.048(0.029)	—	—	—	—
A. hedinii	226	53.5	—	0.009-0.047(0.024)	517	29.2	—	0.004-0.016(0.009)
A. vestita	225	54.2	—	0.007-0.062(0.026)	487	27.7	—	0.003-0.017(0.010)
A. eriopoda	226	58.8	—	0.004-0.094(0.051)	—	—	—	—
A. tangutica	225	55.1	—	0.000-0.066(0.019)	366	24.6	—	0.000-0.014(0.006)
A. scoparia	226	58.8	—	0.004-0.089(0.050)	457	28.7	—	0.003-0.027(0.016)
A. aurata	225	56.9	—	0.004-0.071(0.028)	362	24.6	—	0.003-0.020(0.006)
A. stolonifera	225	55.1	—	0.008-0.102(0.038)	362	24.9	—	0.000-0.022(0.006)
A. princeps	225	55.6	—	0.007-0.092(0.034)	362	24.9	—	0.000-0.022(0.006)
A. mongolica	219	54.3	—	0.005-0.099(0.040)	383	25.1	—	0.003-0.024(0.009)
A. anomala	225	56.4	—	0.013-0.071(0.041)	362	24.6	—	0.006-0.025(0.012)
A. igniaria	225	54.2	—	0.010-0.113(0.047)	362	25.1	—	0.000-0.025(0.008)
A. indica	225	56.9	—	0.004-0.081(0.030)	414	25.8	—	0.002-0.022(0.010)

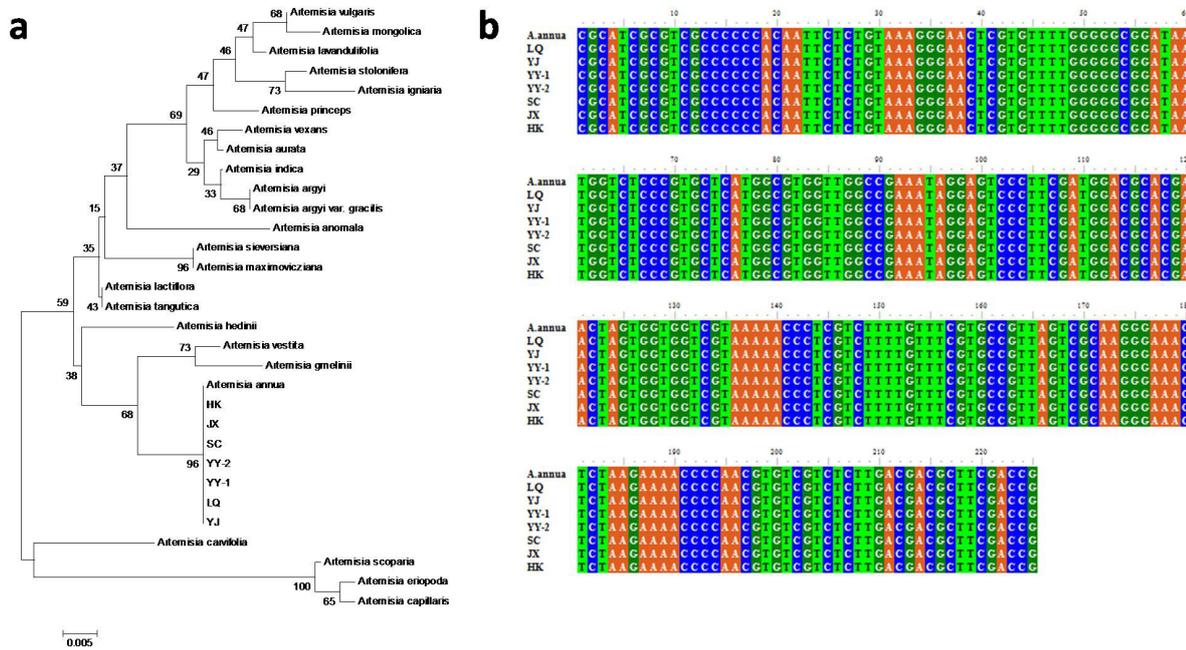| *A. gmelinii* | 219 | 53.0 | — | 0.014-0.076(0.046) | 446 | 25.3 | — | 0.006-0.024(0.015) |
| *A. lavandulifolia* | 225 | 54.7 | — | 0.009-0.091(0.037) | 362 | 24.9 | — | 0.000-0.022(0.007) |
| *A. capillaris* | 226 | 59.3 | — | 0.004-0.113(0.075) | 363 | 24.5 | — | 0.003-0.028(0.019) |

Figures:

Figure 1A: annua ITS2 sequences anylsis

a): the NJ phylogenetic tree based on ITS2 sequences of A. annua and counterfeits.
b): multiple Aligenment of samples and A. annua ITS2 sequences
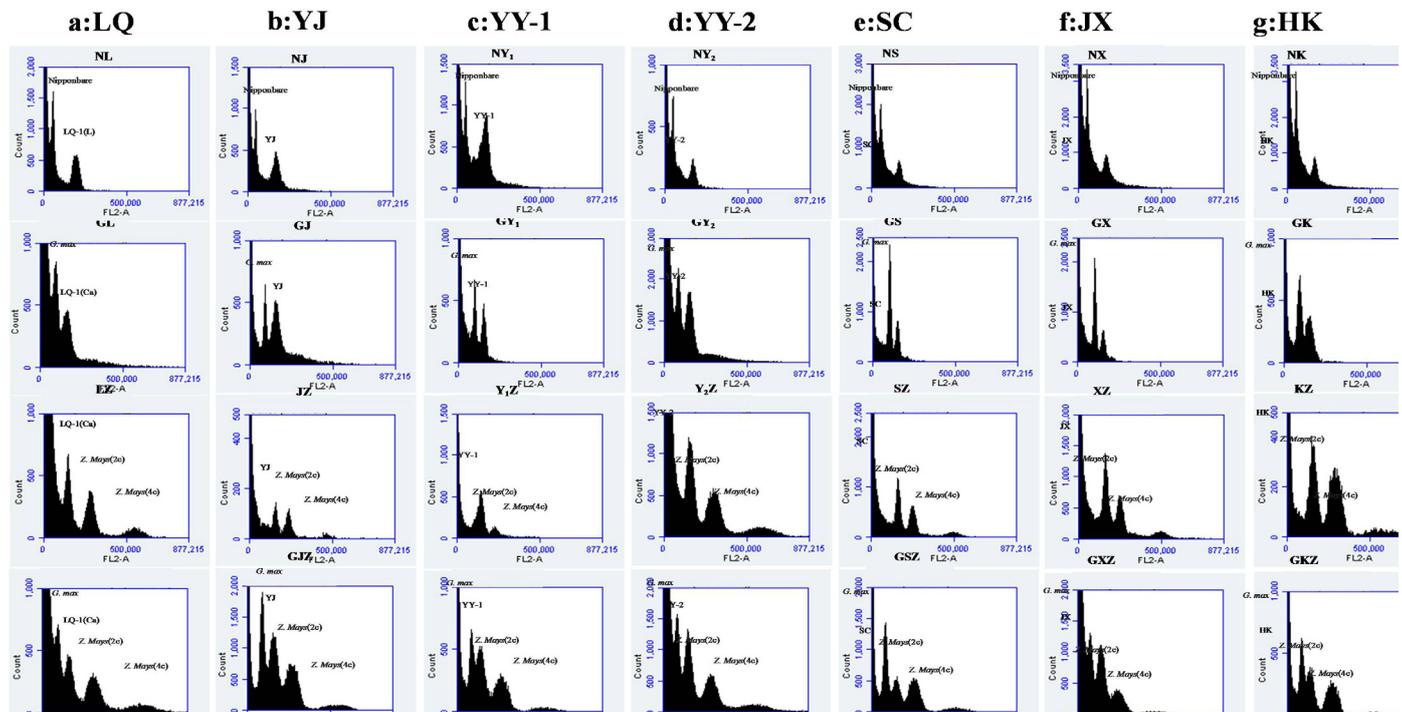

Figure 2: Schematic views of *A. annua* flow cytometric data

N: Nipponbare, G: *G. max*, Z: *Z. mays*. The genome size of *A. annua* was estimated with four groups of standards (Nipponbare, *G. max*, *Z. Mays*, both *G. max* and *Z. Mays*) in strains LQ(a), YJ(b), YY-1(c), YY-2(d), SC(e), JX(f) and HK(g), respectively.

Competing interests: The authors declare that they have no competing interests.

References

1. Committee Tnp. Chinese Pharmacopoeia: The Pharmacopoeia of the People's Republic of China. Bejing: China Medical Science Press. 2015.

2. Mutabingwa TK. Artemisinin-based combination therapies (ACTS): Best hope for malaria treatment but inaccessible to the needy! Acta Trop. 2005; 95: 305-15.

3. Organization WH. World Malaria Report 2014. Accessed 20 Nov 2015.

4. Duffy PE, Mutabingwa TK. Artemisinin combination therapies. Lancet. 2006; 367: 2037-2039.

5. Shen Q, Yan TX, Fu XQ, et al. Transcriptional regulation of artemisinin biosynthesis in *Artemisia annua* L. Sci Bull. 2016; 61: 18-25.

6. Nam W, Tak J, Ryu JK, et al. Effects of artemisinin and its derivatives on growth inhibition and apoptosis of oral cancer cells. Head Neck-J Sci Spec. 2007; 29: 335-340.

7. Efferth T. Molecular pharmacology and pharmacogenomics of artemisinin and its derivatives in cancer cells. Curr Drug Targets. 2006; 7: 407-421.

8. Romero MR, Serrano MA, Vallejo M, et al. Additive antiviral effect of artemisinin and ribavirin on an "in vitro" model of hepatitis C infection. J Hepatol. 2005; 42: 166.

9. Romero MR, Serrano MA, Vallejo M, et al. Antiviral effect of artemisinin from *Artemisia annua* against a model member of the Flaviviridae family, the bovine viral diarrhoea virus (BVDV). Planta Med. 2006; 72: 1169-1174.

10. Utzinger J, Xiao SH, Tanner M, et al. Artemisinins for schistosomiasis and beyond. Curr Opin Investig D. 2007; 8: 105-116.

11. Wei HG. Study on evaluation of germplasm resources and analysis of objective traits of *Artemisia annua* L. Guangxi University. 2007.

12. Abdin MZ, Alam P. Genetic engineering of artemisinin biosynthesis: prospects to improve its production. Acta Physiol Plant. 2015; 37: 33.

13. Tang KX, Shen Q, Yan TX, et al. Transgenic approach to increase artemisinin content in *Artemisia annua* L. Plant cell reports. 2014; 33: 605-615.

14. Paddon CJ, Westfall PJ, Pitera DJ, et al. High-level semi-synthetic production of the potent antimalarial artemisinin. Nature. 2013; 496: 528-532.

15. Vranova E, Coman D, Gruissem W. Network Analysis of the MVA and MEP Pathways for Isoprenoid Synthesis. Annu Rev Plant Biol. 2013; 64: 665-700.

16. Qiu Y, Gu WM. Chinese scientists using conventional chemical methods for the first time efficient synthetic of artemisinin. Xin Hua News Agency. 2012; 7: 4.

17. Zhang WB, Liu DL, Yuan QJ. A method of preparation of artemisinin from arteannuic acid: China. 2012.

18. Bennett MD, Leitch IJ. Nuclear DNA amounts in angiosperms: Progress, problems and prospects. Ann Bot-London. 2005; 95: 45-90.

19. Leitch IJ, Soltis DE, Soltis PS, et al. Evolution of DNA amounts across land plants (embryophyta). Ann Bot-London. 2005; 95:207-217.

20. Gregory TR, Nicol JA, Tamm H, et al. Eukaryotic genome size databases. Nucleic acids research. 2007; 35: 332-338.

21. Leitch IJ, Kahandawala I, Suda J, et al. Genome size diversity in orchids: consequences and evolution. Ann Bot-London. 2009; 104: 469-481.

22. Poggio L, Realini MF, Fourastie MF, et al. Genome downsizing and karyotype constancy in diploid and polyploid congeners: a model of genome size variation. Aob Plants. 2014; 6.

23. Schmuths H, Meister A, Horres R, et al. Genome size variation among accessions of Arabidopsis thaliana. Ann Bot-London. 2004; 93: 317-321.

24. Hu TT, Pattyn P, Bakker EG, et al. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet. 2011; 43: 476.

25. Wolf DE, Steets JA, Houliston GJ, et al. Genome size variation and evolution in allotetraploid Arabidopsis kamchatica and its parents, A-lyrata and A-halleri. Aob Plants. 2014; 6.

26. Lu PL, Han XW, Qi J, et al. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome Res. 2012; 22: 508-518.

27. DeBolt S. Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family Generational Scales. Genome Biol Evol. 2010; 2: 441-453.

28. Goff SA, Ricke D, Lan TH, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science. 2002; 296: 92-100.

29. Yu J, Hu SN, Wang J, et al. A draft sequence of the rice (*Oryza sativa* ssp *indica*) genome. Chinese Sci Bull. 2001; 46: 1937-1942.

30. Yu J, Hu SN, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science. 2002; 296:79-92.

31. Goff SA. A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science. 2005; 309: 879.

32. Ammiraju JSS, Yu YS, Luo MZ, et al. Random sheared fosmid library as a new genomic tool to accelerate complete finishing of rice (*Oryza sativa* spp. Nipponbare) genome sequence: sequencing of gap-specific fosmid clones uncovers new euchromatic portions of the genome. Theoretical and Applied Genetics. 2005; 111: 1596-1607.

33. Piegu B, Guyot R, Picault N, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 2006; 16: 1262-1269.

34. Diez CM, Gaut BS, Meca E, et al. Genome size variation in wild and cultivated maize along altitudinal gradients. New Phytologist. 2013; 199: 264-276.

35. Rayburn AL, Biradar DP, Nelson RL, et al. Documenting intraspecric genome size variation in soybean. Crop Sci. 2004; 44: 261-264.

36. Tenaillon MI, Hufford MB, Gaut BS, et al. Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and Zea luxurians. Genome Biol Evol. 2011; 3: 219-229.

37. Fourastie MF, Poggio L, Hernandez JC, et al. Chromosomes and Karyotype Parameter Variations of Andean Argentinian Maize Populations. Cytogenet Genome Res. 2016; 148: 109.

38. Schmutz J, Cannon SB, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010; 463: 178-183.

39. Lam HM, Xu X, Liu X, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010; 42: 1053.

40. Kim MY, Lee S, Van K, et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107: 22032-22037.

41. Li YH, Zhou GY, Ma JX, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature biotechnology. 2014; 32: 1045.

42. Qi XP, Li MW, Xie M, et al. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Commun. 2014; 5: 4340.

43. Doležel J. Plant genome size estimation by flow cytometry: interlaboratory comparison. Ann Bot (Lond) 1998; 82: 17-26.

44. Doležel J, Doleželová M, Novák FJ. Flow cytometric estimation of nuclear DNA amount in diploid bananas (Musa acuminata and M. balbisiana). Biol Plantarum. 1994; 36: 351-357.

45. Johnston JS, Bennett MD, Rayburn AL, et la. Reference standards for determination of DNA content of plant nuclei. Am J Bot. 1999; 86: 609.

46. Doležel J, Sgorbati S, Lucretti S. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. Physiologia Plantarum. 1992; 85: 625-631.

47. Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Report. 1991; 9:208-218.

48. Biradar DP, Rayburn A. Heterosis and nuclear DNA content in maize. Heredity. 1993; 71: 300-304.

49. Schnable PS, Ware D, Fulton RS, et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science. 2009; 326: 1112-1115.

50. Torrell M, Valles J. Genome size in 21 Artemisia L. species (Asteraceae, Anthemideae): Systematic, evolutionary, and ecological implications. Genome. 2001; 44: 231-238.

51. Xiang L, Zhang W, Chen SL. Literature study and DNA barcoding of traditional Chinese medicine Qinghao (Artemisia annua L.). Acta Pharmaceutica Sinica. 2016; 51: 486-495.

52. Mei QX, Chen XL, Xiang L, et al. DNA Barcode for Identifying Folium Artemisiae Argyi from Counterfeits. Biol Pharm Bull. 2016; 39: 1531-1537.

53. Matsumoto T, Wu JZ, Kanamori H, et al. The map-based sequence of the rice genome. Nature. 2005; 436: 793-800.

54. Lysak MA, Doležel J. Estimation of nuclear DNA content in Sesleria (Poaceae). Caryologia 1998; 52: 123-132.

55. Doležel J, Greilhuber J, Suda J. Estimation of nuclear DNA content in plants using flow cytometry. Nat Protoc. 2007; 2: 2233-2244.

56. Nagl W, Ehrendorfer F. DNA content, heterochromatin, mitotic index, and grown in perennial and annual Anthemidea (Asteraceae). Plant Syst Evol. 1974; 123: 737-740.

57. Geber G, Hasibeder G. Cytophotometrische Bestimmung von DNA-Mengen: Vergleich einer neuen DAPI-Fluoreszenz Methode mit Feulgen-Absortionsphotometrie. Microsc Acta. 1980; 4: 31-35.

58. Godelle B, Cartier D, Marie D, et al. Heterochromatin study demonstrating the non-linearity of fluorometry useful for calculating genomic base composition. Cytometry. 1993; 14: 618-626.

59. Greilhuber J, Ebert I. Genome size variation in Pisum sativum. Genome. 1994; 37: 646-655.

60. Baranyi M, Greilhuber, J. Flow cytometric analysis of genome size variation in cultivated and wild Pisum sativum (Fabaceae). Plant Systematics and Evolution. 1995; 194: 231-239.

61. Petrov DAaW JF. Genome evolution in eukaryotes: The genome size perspective. Oxford, UK: Oxford: University Press. 2006.

62. Yan L, Wang X, Liu H, et al. The Genome of Dendrobium officinale Illuminates the Biology of the Important Traditional Chinese Orchid Herb. Molecular plant. 2015; 8: 922-934.

63. Zhang ZK, Wang YQ, Lin SQ, et al. Ploidy idenitifcation of loquats for genome sequencing project by flow cytometry. Joumal of Fruit Science. 2012; 29: 498-504.

64. Bennett MD, Leitch IJ, Price HJ, et al. Comparisons with Caenorhabditis (similar to 100 Mb) and Drosophila (similar to 175 Mb) using flow cytometry show genome size in Arabidopsis to be similar to 157 Mb and thus similar to 25 % larger than the Arabidopsis genome initiative estimate of similar to 125 Mb. Ann Bot-London. 2003; 91: 547-557.

65. Ornellas AA, Campos MM, Ornellas MH, et al. Flow cytometric analysis of DNA cellular content in 90 consecutive cases of squamous cell carcinoma of penis. Prog Urol. 2000; 10: 72-77.

66. Hedley DW, Friedlander ML, Taylor IW, et al. Method for analysis of cellular DNA content of paraffin-embedded pathological material using flow cytometry. Journal of Histochemistry & Cytochemistry. 1983; 31: 1333-1335.

67. Galbrain DW, Harkinsk R, Maddox JM, et al. Rapid flow cytometric analysis of the cell cycle in intact plnat tissues. Science. 1983; 220: 1049-1051.

68. Arumuganathan K, Earlee D. Estimation of nuclear DNA contents of plants by flow cytometry. Plant Mol Biol Report. 1991; 9: 229-241.

69. Kaul S, Koo HL, Jenkins J, et al. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408: 796-815.

70. Rayburn AL, Biradar DP, Bullock DG, et al. Nuclear DNA content in F"-hybrids of maize. Heredity. 1993; 70: 294-300.

71. Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Report. 1991; 9:208-218.

72. Greilhuber J. "Self-tanning"—a new and important source of stoichiometric error in cytophotometric determination of nuclear DNA content in plants. Plant Syst Evol. 1988; 158: 87-96.

73. Greilhuber J. Intraspecific variation in genome size: a critical reassessment. Ann Bot. 1998; 82: 27-35.

74. Zhang LL, Cao B, Bai CK. The detection of genome size of medicinal plant Scutellaria baicalensis Georgi by flow cytometry. Zhongguo Nongxue Tongbao (Chinese Agricultural Science Bulletin). 2013; 29: 130-135.

75. Li LB, Wu JY, Hu T, et al. Estimation of genome size of Moso bamboo (Phyllostachys edulis). Chin Bull Bot. 2008; 25:574-578.

76. Gui YJ, Wang S, Quan LY, et al. Genome size and sequence composition of moso bamboo: A comparative study. Sci China Ser C. 2007; 50: 700-705.

77. Peng ZH, Lu Y, Li LB, et al. The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla). Nat Genet. 2013; 45:456-461.

78. Mayer KFX, Waugh R, Langridge P, et al. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012; 49: 711.

79. Li QS, Xu J, Zhu YJ, et al. Estimation of the genome size of *Ganoderma lucidum* based on a flow cytometric analysis. Mycosystema. 2013; 32: 899-906.

80. Chen SL. Standard DNA Barcodes of Chinese Material Medica in Chinese Pharmacopoeia. Beijing: Science Press. 2014.

81. J. De. Flow cytometric analysis of nuclear DNA content in higher plants. Phytochemical Analysis. 1991; 2: 143-154.

82. Doležel J. Sgorbati S, Lucretti S. Comparison of three DNA fluorochromes for flow-cytometric estimation of nuclear DNA content in plants. Physiol Plantarum. 1992; 85: 625-631.