1 **Decoding herbal materials of representative TCM** 2 **preparations with the multi-barcoding approach**

3
4 Qi Yao^{1,#}, Xue Zhu^{1,#}, Maozhen Han¹, Chaoyun Chen¹, Wei Li², Hong Bai^{1,*}, Kang Ning^{1,*}

5 ¹*Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key*
6 *Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics*
7 *and Systems Biology, College of Life Science and Technology, Huazhong University*
8 *of Science and Technology, Wuhan, Hubei 430074, China*

9 ²*Faculty of Pharmaceutical Sciences, Toho University, Tokyo, 1438540, Japan*

10 [#] These authors contributed equally to this work

11 * To whom correspondence should be addressed. Email: baihong@hust.edu.cn,
12 ningkang@hust.edu.cn

14 **Abstract**

15 With the rapid development of high-throughput sequencing (HTS) technology, the
16 techniques for the assessment of biological ingredients in Traditional Chinese
17 Medicine (TCM) preparations have also advanced. By using HTS together with the
18 multi-barcoding approach, all biological ingredients could be identified from TCM
19 preparations in theory, as long as their DNA is present. The biological ingredients of a
20 handful of classical TCM preparations were analyzed successfully based on this
21 approach in previous studies. However, the universality, sensitivity and reliability of
22 this approach used on TCM preparations remain unclear. Here, four representative
23 TCM preparations, namely Bazhen Yimu Wan, Da Huoluo Wan, Niu Huang Jiangya
24 Wan and You Gui Wan, were selected for concrete assessment of this approach. We
25 have successfully detected from 77.8% to 100% prescribed herbal materials based on
26 both ITS2 and *trnL* biomarkers. The results based on ITS2 have also shown a higher
27 level of reliability than those of *trnL* at species level, and the integration of both
28 biomarkers could provide higher sensitivity and reliability. In the omics big-data era,
29 this study has undoubtedly made one step forward for the multi-barcoding approach
30 for prescribed herbal materials analysis of TCM preparation, towards better

31 digitization and modernization of drug quality control.

32 **KEY WORDS:** TCM preparations; multi-barcoding approach; universality;

33 sensitivity; reliability

1. Introduction

Traditional Chinese Medicine (TCM) preparation has been used in clinics in China for at least 3,000 years^{1,2}. It has been utilized to prevent and cure various diseases in China and has become more popular all over the world during the last decades. TCM preparation is composed of numerous plants, animal-derived and mineral materials. According to the guidance of Chinese medicine theory and Chinese Pharmacopeia (ChP)³, different medicinal materials were crushed into powder or boiled, then mixed and molded into pills together with honey or water to get a TCM preparation (also called patented drug). Although TCM preparations have been extensively used in recent years, many problems remain to be resolved, such as quality control (QC), in which particular attention should be focused on its materials and production process to ensure its efficacy and safety. The quality of TCM preparations is the prerequisite for their clinical efficacy, its quality assessment includes the qualitative and quantitative analysis of chemical ingredients and biological ingredients⁴. Current methods for the QC of TCMs have been mainly assessed based on chemical profiling⁴ (e.g. TLC⁵, HPLC-UV^{6,7}, HPLC-MS⁸). Through comparing with reference herbal materials or targeted compounds, TLC and HPLC method can retrieve species information but not precise enough, especially in identifying the hybrid species of genetics, which might occur the incorrect identification, introduce biological pollution and adulteration during the herbal materials collection and preprocessing. However, the utilization of DNA, a fragment that stably exists in all tissues⁹, could identify herbal materials at species level accurately, providing a higher level of sensitivity and reliability, thus complementing the drawback of chemical analysis^{10,11}.

The concept of biological ingredient analysis based on DNA-barcoding was proposed by Hebert¹². Chen *et al.* have first applied a serval candidate DNA barcodes to identify medicinal plants and their closely related species¹³. Coghlan *et al.*, for the first time, have used DNA barcoding to determine whether TCM preparations contain derivatives of endangered, trade-restricted species of plants and animals². In 2014, Cheng *et al.* have first reported the biological ingredients analysis for Liuwei Dihuang Wan (LDW) using the metagenomic-based method (M-TCM) based on ITS2 and *trnL* biomarkers¹⁴. After that, the reports on the herbs of TCM preparations based on DNA biomarkers have been sprung up, such as Yimu Wan¹⁵ (YMW), Longdan Xiegan

Wan¹⁶ (LXW) and Jiuwei Qianghuo Wan¹⁷ (JQW). Interestingly, recent studies have reported several TCM preparations that might be effective in the prevention and treatment for COVID-19^{18,19}, such as Lianhua Qingwen capsule²⁰, Jinhua Qinggan granules²⁰, Yiqi Qingjie herbal compound²¹, etc. Among these, Lianhua Qingwen capsule, is reported to be effective in the prevention or treatment for COVID-19 mainly due to its biological ingredients such as Glycyrrhizae Radix Et Rhizoma and Rhei Radix Et Rhizome³. The same principle applies for Jinhua Qinggan granules and Yiqi Qingjie herbal compound. These findings again emphasized the importance of biological ingredient analysis of TCM preparations.

A TCM preparation can be regarded as a “synthesized mixture of species”, which resembles the analytical target of metagenomic approach. In metagenomics approach, based on suitable DNA biomarkers, the genetic information of all DNA-contained ingredients could be obtained in a most effective and cost-effective way via HTS. Due to the conservation of ITS2²² and its high inter-specific and intra-specific divergence power²³⁻²⁵, and the convenience of amplification DNAs from heavily degraded samples based on a short fragment *trnL*²⁶⁻²⁸, these two fragments are usually chosen as biomarkers for herbal species identification. Such an approach based on multiple barcodes for herbal ingredient analysis is referred to as the "multi-barcoding approach".

In spite of scientific advances of recent studies, the solidity (*i.e.*, universality, sensitivity and reliability) of multi-barcoding approach on identifying a variety of biological ingredients of TCM preparations simultaneously remains unclear and needs to be investigated systematically. Therefore, we selected three TCM preparations with simple compositions and pervasively used named Niuhuang Jiangya Wan (NJW), Bazhen Yimu Wan (BYW), Yougui Wan (YGW), and one TCM preparation named Da Huoluo Wan (DHW) with much more complicated components and widely applied, as targets for herbal materials assessment by using ITS2 and *trnL* biomarkers. Based on the assessment of their prescribed herbal species (PHS) of the prescribed herbal materials (PHMs), the universality, sensitivity and reliability of the multi-barcoding approach have been evaluated, from which the multi-barcoding approach stands out as a superior method for PHMs analysis for TCM preparations.

2. Materials and Methods

100 2.1. Sample collections

101 Four TCM preparations, each purchased from two different manufacturers (marked as
102 A and B) with three batches (I, II and III), were collected (**Supplementary Table 1**).
103 Each batch was implemented with three biological replicates based on ITS2 and *trnL*
104 respectively. Therefore, $4 \times 2 \times 3 \times 3 \times 2 = 144$ samples in total were used for the
105 subsequent experiment. Here, we gave an example to clarify the mean of SampleID:
106 DHW.A.II means the DHW sample was bought from the first batch of manufacturer
107 A, and it was one of the three biological replicates (II) of the first batch (I).

108 2.2. DNA extraction and quantification

109 For DNA extraction, we used an optimized cetyl trimethyl ammonium bromide
110 (CTAB) method (TCM-CTAB)²⁹. Each sample (1.0 g) was completely dissolved with
111 0.1 M Tris-HCl, 20 mM EDTA (pH 8.0, 2 ml). Dissolved solution (0.4 mL) was
112 diluted with extraction buffer (0.8 mL) consisting of 2% CTAB; 0.1 M Tris-HCl (pH
113 8.0); 20 mM EDTA (pH 8.0); 1.4 M NaCl, and then 100 μ L 10% SDS, 10 μ L 10
114 mg/mL Proteinase K (Sigma, MO, USA) and 100 μ L β -Mercaptoethanol (Amresco,
115 OH, USA) were added and incubated at 65 °C for 1 h with occasional swirling.
116 Protein was removed by extracting twice with an equal volume of phenol: chloroform:
117 isoamyl-alcohol (25: 24: 1), and once with chloroform: isoamyl-alcohol (24: 1). The
118 supernatant was incubated at -20 °C with 0.6 folds of cold isopropanol for 30 min to
119 precipitate DNA. The precipitate was washed with 75% ethanol, dissolved and diluted
120 to 10 ng/ μ L with TE buffer, and then used as a template for PCR amplification
121 (**Supplementary Figure 1**). DNA concentration was quantified on Qubit[®] 2.0
122 Fluorometer.

123 2.3. DNA amplification and DNA sequencing

124 The PCR amplification was performed in a 50 μ L reaction mixture that contain 1 μ L
125 of DNA extracted from TCM preparations, 10.0 μ L of 5 $\times$ PrimeSTAR buffer (Mg²⁺
126 plus) (TaKaRa), 2.5 μ L of 10 μ M dNTPs (TaKaRa), 0.5 μ L each of forward and
127 reverse primers (10 μ M), 2.5 μ L dimethylsulfoxide (DMSO) and 0.5 μ L PrimeSTAR[®]
128 HS DNA Polymerase (Takara, 2.5 U/ μ L). For amplification and sequencing of ITS2
129 region, the forward primers S2F¹³ and the reverse primer ITS4³⁰ (**Supplementary**
130 **Table 2**) with seven bp MID tags (**Supplementary Table 3**) were designed for PCR
131 amplification. PCR reactions were implemented as follows: pre-denaturation at 95 °C

for five min, then 10 cycles made up of 95 °C for 30 s and 62 °C for 30 s with
ramping of -1 °C per cycle, followed by 72 °C for 30 s, next followed by 40 cycles of
95 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s; the procedure ended with 72 °C for
10 min. For *trnL* region, the forward primers *trnL*-c and the reverse primer *trnL*-h
with 7 bp MID tags were also designed for PCR amplification. The PCR reactions
were carried out according to the conditions: pre-denaturation at 95 °C for five min,
10 cycles made up of 95 °C for 30 s and 62 °C for 30 s with ramping of -1 °C per
cycle, followed by 72 °C for 30 s; then followed by 40 cycles of 95 °C for 30 s, 58 °C
for 30 s and 72 °C for 30 s; the procedure ended with 72 °C for 10 min. For better
amplification effect, touchdown PCR^{30,31} was carried out. The PCR products were
electrophoresed on 1% agarose gel (**Supplementary Figure 2**) and purified with
QIAquick Gel Extraction kit (QIAGEN). The DNA concentration was quantified on
Qubit[®] 2.0 Fluorometer. After removing one *trnL*-marked BYW specimen that failed
to be amplified, which was potentially caused by severe PCR inhibition, and one
ITS2-marked YGW sample that failed to be built the next-generation sequencing
library preparation, 142 samples (**Supplementary Table 4**) were sent for Illumina
MiSeq PE300 paired-end sequencing. The raw sequencing data for TCM preparation
samples were deposited to the NCBI SRA database with accession number
PRJNA562480.

2.4. Sequencing data analysis procedure and software configuration

We first used the FastQC software (version 0.11.7) with default parameters to evaluate
the quality of the sequencing reads. Reads from the same sample were assembled
together by using QIIME script 'join_paired_end.py'. Then we used the
'extract_barcode.py' to extract the double-end barcodes from all reads, and the
'split_libraries_fastq.py' was used to split the sample according to their barcodes
(**Supplementary Table 3**) from the mixed sequencing data, and we also used its '-q
20 --max_bad_run_length 3 --min_per_read_length_fraction 0.75
--max_barcode_errors 0 --barcode_type 7' parameters to preliminarily filter the
low-quality sequences, then Cutadapt software (version 1.14) was used to remove the
primers (**Supplementary Table 2**) and adapter from all samples.

These reads of all samples were QCed by MOTHUR³² (version 1.41.0). Per reads
of ITS2 whose length is <150 bp or >510 bp and the reads of *trnL* whose length is <75
bp were removed. After that, we discarded the sequence whose average quality score

was below 20 in each five bp-window rolling along with the whole reads. Then the sequences that contained ambiguous base call (N), homopolymers of more than eight bases or primers mismatched, uncorrectable barcodes, were also removed from ITS2 and *trnL* datasets.

To match the target species for each sequence, we used the BLASTN (E-value=1E-10) to search in ITS2 and *trnL* database based on GenBank³³, respectively. Among all results, we first chose the prescribed herbal species with the highest score, else we selected the top-scored species. In addition, we also manually searched all prescribed herbal species of prescribed herbal materials in all samples. Then, we discarded the corresponding species of ITS2 and *trnL* sequences with relative abundance below 0.002 and 0.001, respectively. Rarefaction analysis was performed with R³⁴ (version 3.5.2) using the "vegan" package (<https://cran.r-project.org/web/packages/vegan/index.html>) to evaluate the sequencing depth of TCM preparations samples.

To understand the difference of samples between manufacturers and batches, the distance between any two samples was calculated based on Euclidean distance. By using the sample as the node and the distance of any two samples as the edge, we built a network cluster for each TCM preparation and visualized in Cytoscape³⁵ (version 3.7.1) based on ITS2 and *trnL*, respectively. Principal component analysis (PCA) analysis was also utilized in R package "ade4" (<https://cran.r-project.org/web/packages/ade4/index.html>) to detect the difference between two manufacturers based on clustering result. We also used the LDA Effect Size (LEfSe)³⁶ to select legacy biomarker, and then performed feature selection using minimum Redundancy Maximum Relevance Feature Selection (mRMR)³⁷ to select the most discriminative biomarkers. The receiver operating characteristic (ROC) curve³⁸ analysis was applied to evaluate the classification effectiveness of the biomarker selected from different manufacturers.

2.5. Terminology and abbreviation definitions

The prescribed herbal materials were defined as the herbal materials of a TCM preparation contained and recorded in ChP, abbreviated to PHMs.

The prescribed herbal species (abbreviated to PHS) was the original species of PHMs, any one of them should be considered as the PHS.

The species that has the same genus with PHS was defined as substituted herbal

materials (SHS), the species excluded the two above species was considered as the contaminated herbal species (CHS).

For easier understanding of the abbreviations of used in this study, we took one TCM preparation, YGW, as an example, as shown in **Table 1**, and the information for other TCM preparations was shown in **Supplementary Table 5**. We have also provided detailed information about the animal and mineral materials for the four TCM preparations in **Supplementary Table 6**.

Table 1. The abbreviation of prescribed herbal materials, un-prescribed herbal materials and its corresponding prescribed herbal species of Yougui Wan (YGW).

TCM preparation	Prescribed herbal material (PHM)	Prescribed herbal species (PHS)	Substituted herbal species (SHS)	Contaminated herbal species (CHS)
Yougui Wan (YGW)	Aconitum carmichaelii Debx.	<i>Aconitum carmichaelii</i>	Possible substituted herbal species	Possible contaminated herbal species
	Angelica sinensis (Oliv.) Diels.	<i>Angelica sinensis</i>		
	Cinnamomum cassia Presl.	<i>Cinnamomum cassia</i>		
	Cornus officinalis Sieb. et Zucc.	<i>Cornus officinalis</i>		
	Cuscuta australis R. Br.	<i>Cuscuta australis</i> <i>Cuscuta chinensis</i>		
	Dioscorea opposita Thunb.	<i>Dioscorea opposita</i>		
	Eucommia ulmoides Oliv.	<i>Eucommia ulmoides</i>		
	Lycium barbarum L.	<i>Lycium barbarum</i>		
	Rehmanniae radix praeparata	<i>Rehmannia glutinosa</i>		

The universality was a measurement to evaluate how multi-barcoding approach could be applied on a broad scope of TCM preparations. The four representative TCM preparations were selected for this purpose.

The sensitivity was defined as the ratio of the number of detected PHMs, over the number of PHMs that could be identified in theory, that is,

$$\text{Sensitivity} = (\text{the number of detected PHMs}) / (\text{the number of PHMs could be identified in theory})$$

The reliability was defined as the number of detectable PHMs from the TCM preparations by multi-barcoding approach. The larger number of detectable PHMs, the better reliability.

3. Results

3.1. Profiling the prescribed herbal materials for all TCM preparations in Chinese pharmacopoeia

While several widely-used TCM preparations including LDW¹⁴, YMW¹⁵, LXW¹⁶ and JQW¹⁷, have their herbal materials assessed recently, it is important to choose representative preparations for deeper understanding of multi-barcoding approach for the quality assessment of TCM preparation. We examined the all ingredients (including herbs, animals and minerals) and herbal materials only for the TCM preparations recorded in ChP (2015 version) (**Supplementary Figure 3A and B**). On the basis of **Supplementary Figure 3A and B**, it is obvious that most TCM preparations have less than 25 ingredients (the total number of herbs, animals and minerals) and 20 herbal materials, respectively. Therefore, we selected three TCM preparations (BYW, NJW and YGW) with simple compositions and pervasively applications, and one TCM preparation (DHW) with much more complex ingredients, on the assessment of multi-barcoding approach for the quality assessment of TCM preparation. The detailed ingredients information about these four representative TCM preparations (such as the number of their herbal, animal and mineral materials) was shown in **Supplementary Figure 3C**.

3.2. Overview of the herbal materials from TCM preparations

The four selected TCM preparations were purchased from two manufacturers with three batches each. Each sample was implemented with three biological replicates based on ITS2 and *trnL*, respectively. Thus, 144 samples were obtained for DNA extraction, PCR amplification, library building. In this process, except for two failed samples, 142 samples were subjected to paired-end sequencing for subsequent data analysis.

After preliminary filtering (see more details in **Materials and Methods**), we obtained 25,271,042 ITS2 and 27,599,145 *trnL* sequencing reads, averages of 48,493 (BYW), 87,911 (DHW), 161,025 (NJW) and 58,501 (YGW) ITS2 sequencing reads per sample, respectively, and 57,954 (BYW), 139,512 (DHW), 129,560 (NJW) and 61,685 (YGW) *trnL* sequencing reads per sample, respectively (**Table 2**). Then rarefaction analysis was performed for each sample to detect whether the sequencing depth enough. At around 10,000 sequences per sample, all curves tended to approach the saturation plateau, suggesting that the sequencing depth was enough to capture all species information in all samples for the four TCM preparations (**Supplementary Figure 4**). Considering the smaller *trnL* database comparing to the database of ITS2, we filtered the corresponding species of ITS2 and *trnL* sequences with the relative

abundance below 0.002 and 0.001, respectively. There were 47,533, 86,422, 160,712 and 58,008 remained on average of per sample of BYW, DHW, NJW and YGW based on ITS2, and 56,367 (BYW), 130,330 (DHW), 129,012 (NJW) and 59,709 (YGW) based on *trnL* (**Table 2**).

Table 2. The average number of reads of each sample after preliminary QC and threshold filtration for the four TCM preparations.

	Biomarker	BYW	DHW	NJW	YGW
Preliminary QC	ITS2(150~510 bp)	48,493	87,911	161,025	58,501
	<i>trnL</i> ($\geq$ 75 bp)	57,954	139,521	129,560	61,685
Threshold selection	ITS2($\geq$ 0.002)	47,553	86,642	160,712	58,008
	<i>trnL</i> ($\geq$ 0.001)	56,367	130,330	129,012	59,709

Note that, QC means quality control, the reads were removed that below 150 bp or over 510 bp for ITS2, and the reads less than 75 bp for *trnL*, or the sequences that had an average quality score < 20 in each 5 bp-window rolling along with the whole read. Then we filtered out the sequences whose corresponding species was evidenced by the relative abundance less than 0.002 for ITS2 and 0.001 for *trnL*.

In general, several herbal materials have more than one prescribed herbal species, such as licorice, recorded as *Glycyrrhizae Radix Et Rhizoma* in ChP, includes species of *Glycyrrhiza uralensis*, *Glycyrrhiza inflata* and *Glycyrrhiza glabra*. Consequently, anyone original species of prescribed herbal materials (PHMs) should be regarded as prescribed herbal species (PHS). In this work, BYW contained eight prescribed herbal materials, NJW and YGW contain nine PHMs, and DHW contains 36 PHMs, they include 11, 15, 10 and 57 PHS (listed in **Supplementary Table 5** and **Table 1**), respectively.

The results of the ITS2 audit on 18 BYW samples, average of 8.2 PHS, 1.0 substituted herbal species (SHS, the species has the same genus with PHS), and 13.8 contaminated herbal species (CHS, the other detected species except PHS and SHS) was detected, while 5.0 PHS, 0.3 SHS and 14.9 CHS were found in each *trnL* samples (**Figure 1A and B**). For DHW, each sample has the average of 23.7 PHS, 5.1SHS and

284 21.1 CHS based on ITS2, while average of 17.9 PHS, 6.8 SHS and 27.7 CHS based
 285 on *trnL* (**Figure 1C and D**). For NJW samples, average of 7.2 PHS, 2.8 SHS and 1.8
 286 CHS were detected in individual samples based on ITS2, which was more than *trnL*
 287 (3.0 PHS, 3.0 SHS and 24.0 CHS; **Figure 1E and F**). The mean values of PHS, SHS
 288 and CHS detected in per YGW sample were 4.8, 0.9 and 10.4, and 3.7, 0.5, 17.3 based
 289 on ITS2 and *trnL*, respectively (**Figure 1G and H**). These differences may partially
 290 be due to the completeness of ITS2 and *trnL* database, as well as their intrinsic
 291 resolution properties.

292

293 **Figure 1. The number of detected prescribed herbal species (PHS), substituted herbal species**
 294 **(SHS) and contaminated herbal species (CHS), of all samples for per TCM preparation from**
 295 **two manufacturers (A & B). (A) BYW samples based on ITS2; (B) BYW samples based on *trnL*;**
 296 **(C) DHW samples based on ITS2; (D) DHW samples based on *trnL*; (E) NJW samples based on**

297 ITS2; (F) NJW samples based on *trnL*. (G) YGW samples based on ITS2; (H) YGW samples
298 based on *trnL*.

299

300 The phylogenetic trees for each sample were also built based on the ITS2 and
301 *trnL* datasets (**Figure 2** for DHW samples and **Supplementary Figure 5** for other
302 TCM preparations). Each species whose relative abundance was greater than or equal
303 to 0.1%, was displayed with 100% resolution in this tree (that is, any species existed
304 in a sample could be exactly identified at species level). The genetic relationship and
305 the coverage of the detected species were scattered widely, indicating the high
306 sensitivity of the designed primer, and also confirmed there was no biological bias in
307 our experiment.

308

309

310

311

Figure 2. Phylogenetic analysis of the representative species that had at least 0.1% relative abundance for DHW samples. (A) Based on ITS2; (B) Based on trnL. The branch depicts the taxonomic classification of species. The word marked in red means the prescribed herbal species, and the colorful bar means the average relative abundance of species across the three batches from the two manufacturers (A&B).

316

317 In summary, the multi-barcoding approach could accurately identify the herbal
318 materials, including prescribed, substituted, and contaminated materials, for
319 representative TCM preparations (including BYW, DHW, NJW and YGW). The result
320 has demonstrated that the multi-barcoding approach has good universality for
321 detecting PHMs from TCM preparation samples.

322 **3.3. Sensitivity analysis of on prescribed herbal materials from TCM** 323 **preparations**

324 For more detailed probing of the composition of TCM preparations, we chose one
325 TCM preparation named NJW with a relatively simple composition and pervasively
326 application, and another TCM (DHW) with more complex ingredients, as targets to
327 decode their PHMs through identifying their prescribed herbal species of each TCM
328 preparations based on ITS2 and *trnL* datasets.

329 ***Analysis of herbal materials in the TCM preparations based on ITS2:*** The result of
330 the ITS2 auditing on NJW samples, revealed that it could successfully detect all
331 PHMs (9 herbal materials), including the processed herbal materials (such as the
332 extractive of Huangqin), covering 12 detected PHS (**Table 3**). *Senna obtusifolia* (the
333 average relative abundance was 48.4%) and *Senna tora* (45.4%) were the dominant
334 species in all samples, followed by *Paeonia lactiflora* (3.4%) and *Ligusticum*
335 *chuanxiong* (1.0%), suggesting that the modified CTAB method was suitable to
336 extract their DNA and the primers were more suitable to amplify their sequences.
337 Besides the prescribed herbal species, seven substituted herbal species were also
338 found, belonging to *Codonopsis*, *Ligusticum*, *Mentha*, *Paeonia* and *Senna* (their
339 average relative abundance was 0.035%) and six possible contaminated genera
340 namely *Ipomoea*, *Amaranthus*, *Anemone*, *Cuscuta*, *Pogostemon* and *Zanthoxylum*,
341 which might be introduced during the biological experiment.

342

343 **Table 3. Prescribed herbal species for NJW preparation and their presence in each sample**
 344 **by multi-barcoding approach based on ITS2 biomarker.**

Prescribed herbal species (PHS)	NJW.A									NJW.B								
	I1	I2	I3	II1	II2	II3	III1	III2	III3	I1	I2	I3	II1	II2	II3	III1	III2	III3
<i>Astragalus membranaceus</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Codonopsis pilosula</i>	√	√	√	√	√	√	√	√	√		√	√	√	√	√	√	√	√
<i>Curcuma kwangsiensis</i>				√		√												
<i>Curcuma longa</i>				√														
<i>Curcuma wenyujin</i>								√										
<i>Ligusticum chuanxiong</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Mentha haplocalyx</i>	√		√		√	√	√		√			√	√					
<i>Nardostachys jatamansi</i>	√				√		√	√	√						√			
<i>Paeonia lactiflora</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Scutellaria baicalensis</i>	√		√		√		√		√			√			√			
<i>Senna obtusifolia</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Senna tora</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

345 Note that “NJW.A” and “NJW.B” means the “Niu Huang Jiangya Wan” bought from manufacturer
 346 A and B, respectively. “I1” represents the sample is one of three biological replicates of the first
 347 batch sample, and the “√” means that the prescribed herbal species is detected in this sample.
 348

349
 350 For DHW preparation, we detected 35 PHS covering 25 PHMs, including the
 351 processed herbal materials such as Chao Baishu, the sensitivity of PHMs (define as
 352 the ratio of the detected prescribed herbal materials and the prescribed herbal
 353 materials in theory) was 69.4% based on ITS2 (**Table 4**), which was the largest
 354 number of detected PHS in this work. Among the detected PHS from 18 samples, 15
 355 of the 35 detected PHS were found with an average relative abundance over 0.1%,
 356 where seven PHS were identified with an average relative abundance over 1%,
 357 including *Angelica sinensis* (2.0%), *Asarum sieboldii* (1.2%), *Notopterygium*
 358 *franchetii* (1.9%), *Notopterygium incisum* (1.8%), *Paeonia lactiflora* (5.3%), *Paeonia*
 359 *veitchii* (2.0%) and *Pogostemon cablin* (3.7%). Three PHS (*Clematis hexapetala*,
 360 *Coptis teeta*, *Paeonia lactiflora*) were found in all samples, but highly enriched in
 361 DHW.A samples. Average of the relative abundance of *Glycyrrhiza uralensis* (1.56%)
 362 and *Osmunda japonica* (1.64%) detected in samples from DHW.A was 1.6 times more
 363 than DHW.B samples (the relative abundance of these species in DHW.B was 0.94%
 364 and 0.98%, respectively), while *Coptis deltoidei* (one reads in DHW.A.III3), *Ephedra*
 365 *intermedia* (three reads in DHW.A.III2), *Gastrodia elata* (one reads in DHW.A.III3)
 366 and *Rheum tanguticum* (three reads in DHW.B.III1) were only detected from one
 367 sample. Noticeably, the substituted herbal species *Anemone nemorosa* (0.31%) with
 368 the same genus with PHS, was found with high relative abundance in most samples,
 369 especially in DHW.A.II and DHW.A.III.

370

371 **Table 4. Prescribed herbal species for DHW preparation and their presence in each sample**

372 **by multi-barcoding approach based on ITS2 biomarker.**

Prescribed herbal species (PHS)	DHW.A									DHW.B								
	I1	I2	I3	II1	II2	II3	III1	III2	III3	I1	I2	I3	II1	II2	II3	III1	III2	III3
<i>Aconitum kusnezoffii</i>				√	√				√					√	√			
<i>Anomum compactum</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Anemone raddeana</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Angelica sinensis</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Aquilaria sinensis</i>	√	√		√	√		√		√	√	√							√
<i>Asarum heterotropoides</i>					√	√	√	√	√									√
<i>Asarum sieboldii</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Atractylodes macrocephala</i>				√	√	√	√	√	√		√	√	√	√	√	√		√
<i>Clematis hexapetala</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Commiphora myrrha</i>				√	√	√	√	√	√	√	√	√		√	√	√	√	√
<i>Coptis chinensis</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Coptis deltoidea</i>									√									
<i>Coptis teeta</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Cyperus rotundus</i>						√	√	√	√					√				
<i>Ephedra equisetina</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Ephedra intermedia</i>								√										
<i>Ephedra sinica</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Gastrodia elata</i>									√									
<i>Glycyrrhiza glabra</i>				√	√	√	√	√	√	√	√	√		√	√	√	√	√
<i>Glycyrrhiza uralensis</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Lindera aggregata</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Notopterygium franchetii</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Notopterygium incisum</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Osmunda japonica</i>	√			√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Paeonia lactiflora</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Paeonia veitchii</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Panax ginseng</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Pogostemon cablin</i>		√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Rheum officinale</i>				√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Rheum palmatum</i>		√		√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
<i>Rheum tanguticum</i>																		
<i>Saposhnikovia divaricata</i>				√	√	√	√	√	√	√	√	√		√	√	√	√	√
<i>Scrophularia ningpoensis</i>					√		√			√			√			√	√	
<i>Scutellaria baicalensis</i>				√	√	√	√	√	√		√	√		√		√	√	√
<i>Styrax tonkinensis</i>				√	√	√	√	√	√	√	√	√	√	√	√			√

373

374 **Analysis of herbal materials in the TCM preparations based on *trnL*:** For NJW,

375 seven PHS belonged to four genera were detected with low abundance, including

376 *Codonopsis pilosula*, *Curcuma kwangsiensis*, *Curcuma longa*, *Curcuma phaeocaulis*,

377 *Nardostachys chinensis*, *Nardostachys jatamansi*, *Scutellaria baicalensis* (**Table 5**),

378 among them, *Nardostachys chinensis* was captured in all samples, while *Codonopsis*

379 *pilosula* and *Nardostachys jatamansi* were only identified in one sample with one

380 reads, which suggested that DNA of these low relative abundance species was hard to

381 be extracted or the *trnL* c/h primers were not suitable enough for the determination of

382 *Codonopsis pilosula* and *Nardostachys jatamansi*. The substituted *Astragalus* (3.9%)

383 and *Mentha* (8.1%) were captured with high relative abundance. As for possible

384 contaminated herbal species, they were dispersedly distributed in 52 genera.

385

386 **Table 5. Prescribed herbal species for NJW preparation and their presence in each sample**
387 **by multi-barcoding approach based on *trnL* biomarker.**

Prescribed herbal species (PHS)	NJW.A									NJW.B								
	I1	I2	I3	II1	II2	II3	III1	III2	III3	I1	I2	I3	II1	II2	II3	III1	III3	
<i>Nardostachys chinensis</i>	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	
<i>Nardostachys jatamansi</i>								√										
<i>Scutellaria baicalensis</i>									√	√		√	√		√		√	
<i>Curcuma kwangsiensis</i>	√			√						√						√		
<i>Curcuma longa</i>	√			√	√					√	√	√				√		
<i>Curcuma phaeocaulis</i>	√			√	√			√	√	√	√	√			√			
<i>Codonopsis pilosula</i>												√						

388

389

390 Because of complex biological ingredients of DHW, the sensitivity of PHMs (18
391 PHMs, 22 PHS) was only 50% based on *trnL*. Among 22 detected PHS, 12 (**Table 6**)
392 were detected in all samples with an average relative abundance greater than 0.1%,
393 except *Coptis chinensis* (0.05%), in which six of them exceeded 1%, 10 of 22 PHS
394 were below 0.05%. Moreover, the relative abundance of 22 PHS detected from
395 DHW.A was higher than DHW.B. *Boswellia neglecta* (6.4%) was the dominate
396 species, followed by *Glycyrrhiza uralensis* (4.2%), and then *Coptis deltoidei* (2.6%).
397 Nevertheless, *Ephedra equisetina* (12 reads in DHW.A.II3 and 8 reads in DHW.A.III3)
398 and *Scrophularia ningpoensis* (only one reads in both DHW.A.I2 and DHW.A.III1)
399 were only found in two samples. The reason for this low abundant PHS might be due
400 to the manufacturing process: several herbal materials (such as Chao Baishu,
401 vinegar-process Xiangfu) need to be boiled or fried before adding into a TCM
402 preparation.

403

Table 6. Prescribed herbal species for DHW preparation and their presence in each sample by multi-barcoding approach based on *trnL* biomarker.

Prescribed herbal species (PHS)	DHW.A									DHW.B								
	I1	I2	I3	II1	II2	II3	III1	III2	III3	I1	I2	I3	II1	II2	II3	III1	III2	III3
<i>Anemone raddeana</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Angelica sinensis</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Aquilaria sinensis</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Asarum heterotropoides</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Asarum sieboldii</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Atractylodes macrocephala</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Boswellia neglecta</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Cinnamomum cassia</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Clematis hexapetala</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Coptis chinensis</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Coptis deltoidea</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Cyperus rotundus</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Ephedra equisetina</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Ephedra sinica</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Glycyrrhiza uralensis</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Panax ginseng</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Pogostemon cablin</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Rehmannia glutinosa</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Rheum officinale</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Rheum tanguticum</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Scrophularia ningpoensis</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<i>Scutellaria baicalensis</i>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

The analysis of the sensitivity on BYW and YGW samples based on ITS2 and *trnL* biomarker was shown in **Supplementary Table 7-10**. Comparing the analysis result of DHW and NJW, NJW only contains one preprocessed PHM (the extractive of Huangqin), while DHW has seven preprocessed PHMs. This might be due to the more complex preprocessing procedure of DHW. Comparing the detecting result with ITS2 biomarker, much fewer species were identified using *trnL* biomarker, which might be due to DNA extraction, primer specification and the limitation of *trnL* database of Genbank. The three biological replicates from these batches have shown different prescribed herbal species (PHS) compositions based on both ITS2 or for *trnL* (Table 3-6 and **Supplementary Table 7-10**), which might be potentially caused by DNA extraction, PCR amplification, high-through sequencing technology, and the previous researches of LDW¹⁴, YMW¹⁵, LXW¹⁶ and JQW¹⁷ have also shown this phenomenon.

All detected species including PHS, SHS and CHS of these four TCM preparations were also provided in **Supplementary Table 11** (provided as a separate attachment in .xlsx format). Based on ITS2 biomarker, we detected 8, 25, 9 and 6 prescribed herbal materials of BYW, DHW, NJW and YGW, respectively. The detected proportion of prescribed herbal materials was 100% for BYW and NJW,

426 followed by DHW (69.4%) and YGW (66.7%). As for *trnL*, 5, 18, 4 and 4 prescribed
427 herbal materials of BYW, DHW, NJW and YGW were respectively detected, and the
428 maximum sensitivity of prescribed herbal materials was 62.5% among the four TCM
429 preparations in this experiment. The analysis strongly suggested the multi-barcoding
430 approach has a high sensitivity in identifying prescribed herbal materials of TCM
431 preparations, especially based on ITS2 dataset.

432 **3.4. Prediction model to predict the identity and quality of TCM preparations**

433 By enabling a model to differentiate the sample from a different group, we can also
434 identify the manufacturer, the batch from where samples were collected. Various
435 distance measures can be used to evaluate the inter/intra-manufacturers difference.
436 Here, we calculated the Euclidean distances of any two samples based on the
437 existence of all detected species and then clustered the samples according to their
438 similarity. We took DHW as a case study. The results showed that most samples from
439 DHW.A and DHW.B clustered together respectively based on both ITS2 (**Figure 3A**
440 and **B**) and *trnL* (**Figure 3C** and **D**) biomarkers, suggesting that the high similarity of
441 intra-manufacturer samples. It is obvious that DHW.A.II and DHW.A.III is clustered
442 with DHW.B samples, whereas three samples of DHW.A.I were gathered and distant
443 from the other samples (**Figure 3A and B**). The reason for such separation might be
444 the existence of substituted herbal species such as *Senna*, *Amaranthus*, *Glycine* and
445 contaminated herbal species such as *Arachis*, *Brassica*, *Solanum* and *Oryza*. As for
446 NJW (**Supplementary Figure 6**), the samples from two manufacturers (A&B) were
447 scattered, based on either ITS2 or *trnL*, while clustered tighter within batches, which
448 depicted the high consistency between batches of NJW samples. The cluster analysis
449 of BYW and YGW samples was shown in **Supplementary Figure 7-8** respectively,
450 which showed the clear difference between manufacturers, as well as high similarity
451 within the same manufacturer.

Figure 3. Comparison of the similarity of all DHW samples from intra-/inter-manufacturers based on prescribed herbal materials using Euclidean distances. Heatmap clusters displayed the distance of all samples based on the existence of prescribed herbal species using hierarchical clustering, and network clusters illustrated these differences in Cytoscape based on ITS2 (A and B) and *trnL* (C and D) sequencing results. For heatmap (A & C), the gradient color bars mean the distance between any two samples, while the red and the blue color depicts the two extreme distances between samples. For network (B & D), each edge represents the distance of any two samples with distance less than or equal to 5.0 for ITS2 and 4.2 for *trnL*.

PCA analysis was also performed to explore the consistency of samples from two manufacturers. The samples from DHW.B were clustered more closely than DHW.A based on ITS2 and *trnL* biomarker. Based on ITS2, the samples of DHW from intra-batch were clustered together, and the inter-batches distributed sparsely, whereas based on *trnL*, the samples of DHW.A were dispersed far apart (**Supplementary Figure 9C and D**), which suggested that the consistency of DHW.B samples was better than DHW.A. The cluster degree of samples from NJW (**Supplementary Figure 9E and F**) was more dispersive than DHW. The result of BYW and YGW

(Supplementary Figure 9 A&B and G&H) was also showed similar results.

To explore which species drove the difference of intra-/inter-manufacturers samples, LEfSe analysis was conducted. 13 prescribed herbal species from DHW.A and four of DHW.B (Figure 4A) were identified as tentative biomarkers. Through mRMR, five PHS from DHW.A and two PHS of DHW.B were selected and visualized in ROC curves (Figure 4B) to evaluate their classification ability. As the curve of *Glycyrrhiza glabra* was below the model score curve, we removed this biomarker from DHW.A. Thus, *Coptis chinensis*, *Ephedra equisetina*, *Lindera aggregate* and *Panax ginseng* were chosen as unique biomarkers of DHW.A, whereas *Rheum palmatum* and *Clematis hexapetala* were selected as representative biomarkers of DHW.B. All of them are of high discrimination power, which could be used separately or in combination to differentiate the samples from the two manufacturers.

Figure 4. The difference of samples from the two manufacturers (A & B) could be driven by a few discriminative prescribed herbal species of prescribed herbal materials of DHW using ITS2 biomarker. (A) The legacy biomarkers selected by LEfSe; (B) ROC curves to evaluate the effect of the legacy biomarkers after removing redundant markers from the two manufacturers.

3.5. Comparison of ITS2 and *trnL* on resolutions and sensitivities

Through detecting their prescribed herbal species, the detected proportion of prescribed herbal materials was 100% for BYW and NJW, followed by DHW (69.4%) and YGW (66.7%) based on ITS2, while 62.5%, 50%, 44.4% and 44.4% for BYW, DHW, NJW and YGW based on *trnL* datasets respectively (Table 7). The sensitivities of ITS2 is better than *trnL* in all TCM preparations, but *trnL* biomarker could also detect the PHS of PHMs that ITS2 couldn't, and the union of both increases the sensitivity of the lower limit to 77.8%, providing a more reliable (as for positive

496 detections) detected result.

497

498 **Table 7. The sensitivity of prescribed herbal materials for four TCM preparations based on**

499 **ITS2 and *trnL* biomarker.**

	ITS2 (%)	<i>trnL</i> (%)	Union (%)
BYW	100	62.5	100
DHW	69.4	50	77.8
NJW	100	44.4	100
YGW	66.7	44.4	77.8

500

501 Note that the sensitivity was defined as the ratio of the detected prescribed herbal materials and
502 the prescribed herbal materials could be detected in theory.

503

504 As can be observed from the Venn diagram (**Figure 5**), the detection result of
505 BYW, all its prescribed herbal materials were detected. As for DHW, the union
506 detection result of these two regions was 38 PHS, covering 28 prescribed herbal
507 materials, which increased the identification efficiency to 77.8%. Similarly, the
508 detection result of *trnL* from NJW preparation was a subset of ITS2, with 100%
509 sensitivity. For YGW samples, the union of these two regions increased the sensitivity
510 to 77.8%, because of two undetected PHMs named *Myristicae semen* (Rougui) and
511 *Dioscoreae rhizome* (Shanyao). This result has also confirmed the high reliability of
512 the multi-barcoding approach. We then compared our result with the previous studies,
513 including JQW, LXW, YMW and the YYW (**Table 8**), which indicated the reliability
514 of the multi-barcoding approach, this was also suggested that the complexity of
515 biological ingredients of TCM preparation has also negatively affected the detected
516 results.

517

Figure 5. The identified specific and shared prescribed herbal species of TCM preparations based on ITS2 and *trnL*. Results on (A) BYW; (B) DHW; (C) NJW; (D) YGW were shown. The numbers below the Venn diagram mean the number of prescribed herbal species detected based on ITS2 only, *trnL* only and the intersection of the two.

Table 8. Comparison of the sensitivity of prescribed herbal materials through detected prescribed herbal species of TCM preparations.

TCM preparations	All materials	PHMs	The number of detected PHMs			Sensitivity (%)			Undetected PHMs	Reference
			Biomarker 1	Biomarker 2	Union	Biomarker 1	Biomarker 2	Union		
BYW	9	8	8 (ITS2)	5 (<i>trnL</i>)	8	100	62.5	100	Fulin	this work
DHW	48	36	25 (ITS2)	18 (<i>trnL</i>)	28	69.4	50	77.8	Six (a)	this work
JQW	9	9	6 (ITS2)	6 (<i>psbA-trnH</i>)	8	66.7	66.7	88.9	Baizhi	17
LDW	6	5	5 (ITS2)	4 (<i>trnL</i>)	5	100	80	100	—	14
LXW	10	10	8 (ITS2)	—	8	80	—	80.0	Zexie Dihuang	16
NJW	14	9	9 (ITS2)	4 (<i>trnL</i>)	9	100	44.4	100	—	this work
YGW	10	9	6 (ITS2)	4 (<i>trnL</i>)	7	66.7	44.4	77.8	Fuzi Shanyao	this work
YMW	4	4	4 (ITS2)	3 (<i>psbA-trnH</i>)	4	100	75	100	—	15
YYW	4	1	—	1 (<i>trnL</i>)	100	—	100	100	—	2

Note that we only calculated the sensitivity of prescribed herbal materials of TCM preparation samples bought from manufacturers. The word in red is the research targets of this work, while those in black are from the previous studies.

(a) The six undetected PHMs of DHW were *Arisaematis rhizome* (Tiannanxing), *Aucklandiae radix* (Muxiang), *Olibanum* (Ruxiang), *Citri reticulatae pericarpium viride* (Qingpi), *Draconis sanguis* (Xuejie), *Drynariae rhizome* (Gusuibu), *Caryophylli flos* (Dingxiang), *Polygoni multiflori radix* (Heshouwu), *Puerariae lobatae radix* (Gegen).

534

535 Though the sensitivity and reliability of multi-barcoding approach have been
536 clearly demonstrated, the results on ITS2 and *trnL* are clearly different. It is obvious
537 that ITS2 showed a higher sensitivity than that of *trnL* for PHMs detection. The
538 reason might be due to the longer conserved region of ITS2 which can capture more
539 information. Nevertheless, the role of *trnL* is irreplaceable, as it could complement
540 ITS2 for more reliable identification of the prescribed herbal materials of TCM
541 preparations such as for the biological ingredient analysis of DHW and YGW in this
542 work.

543

544 **4. Discussions and Conclusion**

545 As already know to us, herbal materials are the most important elements in different
546 traditional medicines. An increasing number of papers on DNA-based authentication
547 of single herbs have been published^{27,39-44}, while a few applications of the
548 multi-barcoding approach for TCM preparations were reported^{14,45-47}.

549 **4.1. The multi-barcoding approach decoded the prescribed herbal materials of** 550 **the four TCM preparations**

551 In this work, we have systematically examined the universality, sensitivity and
552 reliability of multi-barcoding approach for four representative TCM preparations. This
553 method has successfully detected the species (including prescribed, substituted and
554 contaminated species) contained in a sample with high sensitivity, indicating the good
555 universality of the method and its potential value for daily TCM supervision. As we
556 could determine the existence of all species contained in one sample at species level,
557 these results have indicated an adequate sensitivity of this method in decoding herbal
558 materials of TCM preparations through authenticating their corresponding species.
559 The combined results of ITS2 and *trnL* have increased the sensitivity from 77.8% to
560 100% that highlights the practical application value and high reliability of this
561 approach. Particularly, the ITS2 exhibited an excellent ability and sensitivity for
562 identifying herbal materials. Although the resolution of *trnL* was lower than that of
563 ITS2, it could also be used to reinforce or complement ITS2 for more reliable results.
564 These results have demonstrated that multi-barcoding was an efficient tool for
565 decoding the herbal materials of various kinds of TCM preparations.

For example for BYW and NJW, all prescribed herbal materials were detected through authenticating its corresponding prescribed herbal species. The detected prescribed herbal species of DHW were 35 (covered 25 prescribed herbal materials), 22 (covered 18 prescribed herbal materials) based on ITS2 and *trnL*, respectively. The union dataset of ITS2 and *trnL* has boosted the sensitivity increasing from 69.4% to 77.8% for DHW samples. However, six prescribed herbal materials were not detected in all DHW samples based on either ITS2 or *trnL*. These phenomena might be due to various preprocessing procedures, such as decocted or stir-fried herbal materials, who's DNA was damaged or degraded. We also note that due to several influencing factors, such as geological location, cultivation conditions, climate and other conditions, the sensitivity of PHMs of each TCM preparation sample is different.

This multi-barcoding approach has successfully analyzed the herbal materials of four TCM preparations, which could not be realized through traditional methods, such as morphological and biochemical means. In the future, more diverse sets of TCM preparations could be assessed by this method, which not only making the identification of TCM preparation automatically, but also accelerating the digitization and modernization of TCM management process.

4.2. Outlook and future plans

However, a deeper and more comprehensive improvement of this multi-barcoding approach still needs to be carried out. A more comprehensive species database was necessary, since the reliability of the biological ingredient analysis method for TCM preparation were largely dependent on the coverage of the reference database². In our future study, we can utilize multiple databases, including the GenBank database, as well as tcmbarcode database⁴⁸, EMBL, DDBJ and PDB² to obtain more complete results. Additionally, more biomarker candidates can be considered for assessing the quality of TCM preparation.

Firstly, the multi-barcoding approach could be an attempt to use in identifying the animal materials, because the animal materials still are an important component of TCM, are often combined with medical herbs to exert its pharmacological effects⁴⁹.

Secondly, chemical ingredients and biological ingredients are indivisible yet both important for quality assessments of TCM preparations. Therefore, combining the chemical methods with DNA barcoding approach, the detection of TCM ingredients will outperform than the results of any one of them. Although this thought was

initially tested by our group¹¹, there is still room for further improvement.

Thirdly, the network pharmacology approach has provided us with a more direct view about the drug-target interactions⁵⁰, which gives us an insight into how to optimize the existing drugs and to discover the new medicine for satisfying the requirements of overcoming complex diseases. Thus, the pharmacological usage should be considered in the QC of TCM preparations, especially for the specific usages of TCM, such as the mechanism-based QC of YIV-906⁵¹. This theory has also inspired us to explore the potential treatments of COVID-19 from biological ingredients of TCM preparations⁵². In fact, the ingredients such as Glycyrrhizae Radix Et Rhizoma could frequently interact with the target of COVID-19: ACE2^{20,52}. Through data-mining, the characteristic of eight biological ingredients of DHW is corresponding to the classic Warm disease's symptoms of syndrome differentiation of COVID-19, which might prove effective for the treatment of COVID-19⁵². These biological ingredient information, if combine with public health data, might shed more lights on the susceptibility of patient who has taken these TCM preparations, especially those elderly people.

Finally, many herbal medicines are taken orally⁵³, undoubtedly exposed to the whole gastrointestinal tract microbiota, which provides sufficiently spatiotemporal opportunities for their direct or indirect interactions. For example, berberine, the major pharmacological ingredients of *Coptidis rhizome* (Huanlian)⁵⁴, it promotes the production of short-chain fatty acid to shift the gut microbiota structure, while the poorly solubilized berberine⁵⁵ was converted into dihydroberberine through a reduction reaction mediated by bacterial nitroreductase, then recovered to the original form after penetrating into the intestinal wall tissues⁵⁶, through interactions, the microbial diversity in high-fat diet mice intestines was profoundly decreased⁵⁷.

We believe that all of these efforts on QC of TCM preparations could and would joint-force and provide much better and optimized approaches for the next-generation TCM preparation quality control system. Through reshaping the symbiotic microbiome composition, we could provide novel therapeutic strategies to accelerate the realization of personalized therapeutics.

Acknowledgments

This work was partially supported by National Science Foundation of China grant

81573702, 81774008, 31871334 and 31671374, and National Key Research and Development Program of China grant 2018YFC0910502.

Authors' contributions

KN and HB designed the whole study. HB, MZH, CYC and QY collected the samples and conducted the DNA extraction and sequencing. XZ analyzed the sequencing data. XZ, HB, MZH and KN wrote, revised and proof-read the manuscript. All authors read and approved the final manuscript.

Competing financial interests

The authors declare no competing financial interest.

Data availability

The raw sequencing data used in this work was deposited to NCBI SRA database with accession number PRJNA562480. The ITS2 sequences of the sequenced single herbs were also deposited to NCBI SRA database with NCBI SRA database with accession number PRJNA600815.

Reference

- 1 Lindsay, P., Ross, M. E., Carvalho, G. R. & Rob, O. A DNA-based approach for the forensic identification of Asiatic black bear (*Ursus thibetanus*) in a traditional Asian medicine. *J Forensic Sci* 2010,**53**:1358-1362.
- 2 Coghlan, M. L., Haile, J., Houston, J., Murray, D. C., White, N. E., Moolhuijzen, P. *et al.* Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. *PLoS Genet* 2012,**8**:e1002657.
- 3 Pharmacopoeia, C. C. Pharmacopoeia of the People's Republic of China. *China Medical Science Press* 2015,**Vol. I**:478-479.
- 4 Bai, H., Ning, K. & Wang, C. Y. Biological ingredient analysis of traditional Chinese medicines utilizing metagenomic approach based on high-throughput-sequencing and big-data-mining. *Acta Pharm Sin B* 2015,**50**:272-277.
- 5 Kim, H. J., Jee, E. H., Ahn, K. S., Choi, H. S. & Jang, Y. P. Identification of marker compounds in herbal drugs on TLC with DART-MS. *Arch Pharm Res* 2010,**33**:1355-1359.
- 6 Ciesla, Ł., Hajnos, M., Staszek, D., Wojtal, Ł., Kowalska, T. & Waksmundzka-Hajnos, M. Validated binary high-performance thin-layer chromatographic fingerprints of polyphenolics for

- 666 distinguishing different *Salvia* species. *J Chromatogr Sci* 2010,**48**:421-427.
- 667 7 Chen, S.-B., Liu, H.-P., Tian, R.-T., Yang, D.-J., Chen, S.-L., Xu, H.-X. *et al.* High-performance
668 thin-layer chromatographic fingerprints of isoflavonoids for distinguishing between *Radix*
669 *Puerariae Lobate* and *Radix Puerariae Thomsonii*. *J Chromatogr A* 2006,**1121**:114-119.
- 670 8 Zhang, J. M., Li, L., Gao, F., Li, Y., He, Y. & Fu, C. M. Chemical ingredient analysis of sediments
671 from both single *Radix Aconiti Lateralis* decoction and *Radix Aconiti Lateralis* - *Radix*
672 *Glycyrrhizae* decoction by HPLC-MS. *Acta Pharm Sin B* 2012,**47**:1527-1533.
- 673 9 Miller, S. E. DNA barcoding and the renaissance of taxonomy. *Proc Natl Acad Sci U S A*
674 2007,**104**:4775-4776.
- 675 10 Jiang, Y., David, B., Tu, P. & Barbin, Y. Recent analytical approaches in quality control of
676 traditional Chinese medicines-A review. *Anal Chim Acta*. 2010,**657**:9-18.
- 677 11 Bai, H., Li, X., Li, H., Yang, J. & Ning, K. Biological ingredient complement chemical ingredient
678 in the assessment of the quality of TCM preparations. *Sci Rep* 2019,**9**:5853-5853.
- 679 12 Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through
680 DNA barcodes. *Proc Biol Sci* 2003,**270**:313-321.
- 681 13 Shilin, C., Hui, Y., Jianping, H., Chang, L., Jingyuan, S., Linchun, S. *et al.* Validation of the ITS2
682 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 2010,**5**:e8613.
- 683 14 Cheng, X., Su, X., Chen, X., Zhao, H., Bo, C., Xu, J. *et al.* Biological ingredient analysis of
684 traditional Chinese medicine preparation based on high-throughput sequencing: the story for
685 Liuwei Dihuang Wan. *Sci Rep* 2014,**4**:5147.
- 686 15 Jia, J., Xu, Z., Xin, T., Shi, L. & Song, J. Quality Control of the Traditional Patent Medicine Yimu
687 Wan Based on SMRT Sequencing and DNA Barcoding. *Front Plant Sci* 2017,**8**:926.
- 688 16 Xin, T., Su, C., Lin, Y., Wang, S., Xu, Z. & Song, J. Precise species detection of traditional Chinese
689 patent medicine by shotgun metagenomic sequencing. *Phytomedicine* 2018,**47**:40-47.
- 690 17 Xin, T., Xu, Z., Jia, J., Leon, C., Hu, S., Lin, Y. *et al.* Biomonitoring for traditional herbal medicinal
691 products using DNA metabarcoding and single molecule, real-time sequencing. *Acta Pharm Sin B*
692 2018,**8**:488-497.
- 693 18 Ren, J. L., Zhang, A. H. & Wang, X. J. Traditional Chinese medicine for COVID-19 treatment.
694 *Pharmacol Res* 2020,**155**:104743.
- 695 19 Du, H. Z., Hou, X. Y., Miao, Y. H., Huang, B. S. & Liu, D. H. Traditional Chinese Medicine: an
696 effective treatment for 2019 novel coronavirus pneumonia (NCP). *Chin J Nat Med*
697 2020,**18**:206-210.
- 698 20 Zhang, D., Zhang, B., Lv, J.-T., Sa, R.-N., Zhang, X.-M. & Lin, Z.-J. The clinical benefits of
699 Chinese patent medicines against COVID-19 based on current evidence. *Pharmacol Res*
700 2020,**157**:104882.
- 701 21 Li, S. & Li, J. Treatment effects of Chinese medicine (Yi-Qi-Qing-Jie herbal compound) combined
702 with immunosuppression therapies in IgA nephropathy patients with high-risk of end-stage renal
703 disease (TCM-WINE): study protocol for a randomized controlled trial. *Trials* 2020,**21**:31.
- 704 22 Keller, A., Schleicher, T., Schultz, J., Muller, T., Dandekar, T. & Wolf, M. 5.8S-28S rRNA
705 interaction and HMM-based ITS2 annotation. *Gene* 2009,**430**:50-57.
- 706 23 Chen, S.-L., Yao, H., Han, J.-P., Xin, T.-Y., Pang, X.-H., Shi, L.-C. *et al.* Principles for molecular
707 identification of traditional Chinese materia medica using DNA barcoding. *Chin J Chin Mater*
708 *Med* 2013,**38**:141-148.
- 709 24 Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., Liu, J. Q. *et al.* Comparative analysis of a large

dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc Natl Acad Sci U S A* 2011,**108**:19641-19646.

25 Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y. *et al.* Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One* 2010,**5**:e13102.

26 Ward, J., Peakall, R., Gilmore, S. R. & Robertson, J. A molecular identification system for grasses: a novel technology for forensic botany. *Forensic Sci Int* 2005,**152**:121-131.

27 Wang, G. P., Fan, C. Z., Zhu, J. & Li, X. J. Identification of original plants of uyghur medicinal materials fructus elaeagni using morphological characteristics and DNA barcode. *Chin J Chin Mater Med* 2014,**39**:2216-2221.

28 Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A. *et al.* Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Res* 2007,**35**:e14.

29 Cheng, X., Chen, X., Su, X., Zhao, H., Han, M., Bo, C. *et al.* DNA Extraction Protocol for Biological Ingredient Analysis of Liuwei Dihuang Wan. *GPB* 2014,**12**:137-143.

30 Chen, J., Dai, L., Wang, B., Liu, L. & Peng, D. Cloning of expansin genes in ramie (*Boehmeria nivea* L.) based on universal fast walking. *Gene* 2015,**569**:27-33.

31 Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K. & Mattick, J. S. 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res* 1991,**19**:4008.

32 Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microb* 2009,**75**:7537-7541.

33 Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* 2000,**28**:15-18.

34 Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J Comput Graph Stat* 1996,**5**:299-314.

35 Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003,**13**:2498-2504.

36 Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol* 2011,**12**:R60.

37 Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 2005,**27**:1226-1238.

38 Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982,**143**:29-36.

39 Osathanunkul, M., Suwannapoom, C., Ounjai, S., Rora, J. A., Madesis, P. & de Boer, H. Refining DNA Barcoding Coupled High Resolution Melting for Discrimination of 12 Closely Related Croton Species. *PLoS One* 2015,**10**:e0138888.

40 Carles, M., Cheung, M. K., Moganti, S., Dong, T. T., Tsim, K. W., Ip, N. Y. *et al.* A DNA microarray for the authentication of toxic traditional Chinese medicinal plants. *Planta Med* 2005,**71**:580.

41 Zhou, J., Wang, W., Liu, M. & Liu, Z. Molecular authentication of the traditional medicinal plant *Peucedanum praeruptorum* and its substitutes and adulterants by dna-barcoding technique. *Pharmacogn Mag* 2014,**10**:385.

754 42 Yip, P. Y., Chau, C. F., Mak, C. Y. & Kwan, H. S. DNA methods for identification of Chinese
755 medicinal materials. *Chin Med* 2007,**2**:9.

756 43 Khan, S., Al-Qurainy, F. & Nadeem, M. Biotechnological approaches for conservation and
757 improvement of rare and endangered plants of Saudi Arabia. *Saudi J Biol Sci* 2012,**19**:1-11.

758 44 Ma, X.-X., Sun, W., Ren, W.-C., Xiang, I., Zhao, B., Zhang, Y.-Q. *et al.* Identification of cattail
759 pollen (puhuang), pine pollen (songhuafen) and its adulterants by ITS2 sequence. *Chin J Chin*
760 *Mater Med* 2014,**39**:2189-2193.

761 45 Newmaster, S. G., Grguric, M., Shanmughanandhan, D., Ramalingam, S. & Ragupathy, S. DNA
762 barcoding detects contamination and substitution in North American herbal products. *BMC Med*
763 2013,**11**:222.

764 46 Li, M., Cao, H., BUT, P. P. H. & SHAW, P. C. Identification of herbal medicinal materials using
765 DNA barcodes. *J Syst Evol* 2011,**49**:271-283.

766 47 Chiou, S.-J., Yen, J.-H., Fang, C.-L., Chen, H.-L. & Lin, T.-Y. Authentication of medicinal herbs
767 using PCR-amplified ITS2 with specific primers. *Planta Med* 2007,**73**:1421-1426.

768 48 Chen, S., Pang, X., Song, J., Shi, L., Yao, H., Han, J. *et al.* A renaissance in herbal medicine
769 identification: From morphology to DNA. *Biotechnol Adv* 2014,**32**:1237-1244.

770 49 Still, J. Use of animal products in traditional Chinese medicine: environmental impact and health
771 hazards. *Complement Ther Med* 2003,**11**:118-122.

772 50 Hopkins, A. L. Network pharmacology. *Nat Biotechnol* 2007,**25**:1110.

773 51 Lam, W., Ren, Y., Guan, F., Jiang, Z., Cheng, W., Xu, C.-H. *et al.* Mechanism Based Quality
774 Control (MBQC) of Herbal Products: A Case Study YIV-906 (PHY906). *Front Pharmacol*
775 2018,**9**:1324-1324.

776 52 Ren, X., Shao, X. X., Li, X. X., Jia, X. H., Song, T., Zhou, W. Y. *et al.* Identifying potential
777 treatments of COVID-19 from Traditional Chinese Medicine (TCM) by using a data-driven
778 approach. *J Ethnopharmacol* 2020,**258**:112932.

779 53 Qiu, J. 'Back to the future' for Chinese herbal medicines. *Nat Rev Drug Discov* 2007,**6**:506-507.

780 54 Kamada, N., Chen, G. Y., Inohara, N. & Núñez, G. Control of pathogens and pathobionts by the gut
781 microbiota. *Nat Immunol* 2013,**14**:685.

782 55 Zhaojie, M., Ming, Z., Shengnan, W., Xiaojia, B., Hatch, G. M., Jingkai, G. *et al.* Amorphous solid
783 dispersion of berberine with absorption enhancer demonstrates a remarkable hypoglycemic effect
784 via improving its bioavailability. *Int J Pharm* 2014,**467**:50-59.

785 56 Feng, R., Shou, J.-W., Zhao, Z.-X., He, C.-Y., Ma, C., Huang, M. *et al.* Transforming berberine into
786 its intestine-absorbable form by the gut microbiota. *Sci Rep* 2015,**5**:12155.

787 57 Chen, F., Wen, Q., Jiang, J., Li, H.-L., Tan, Y.-F., Li, Y.-H. *et al.* Could the gut microbiota reconcile
788 the oral bioavailability conundrum of traditional herbs? *J Ethnopharmacol* 2016,**179**:253-264.

789