SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
MOLECULAR BIOLOGY
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Received
28 January 2014

Accepted
6 May 2014

Published
3 June 2014

Correspondence and requests for materials should be addressed to H.B. (baihong@gmail.com) or K.N. (ningkang@qibebt.ac.cn)

* These authors contributed equally to this work.

Biological ingredient analysis of traditional Chinese medicine preparation based on high-throughput sequencing: the story for Liuwei Dihuang Wan

Xinwei Cheng[1,3]*, Xiaoquan Su[2]*, Xiaohua Chen[2], Huanxin Zhao[1], Cunpei Bo[2], Jian Xu[2], Hong Bai[1] & Kang Ning[2]

[1]Institute of Materia Medica, Shandong Academy of Medical Sciences, Jinan, Shandong, 250062, China, [2]Shandong Key Laboratory of Energy Genetics, CAS Key Laboratory of Biofuels and Bioinformatics Group of SingleCell Center, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao266101, China, [3]School of Medicine and Life Sciences, University of Jinan-Shandong Academy of Medical Sciences, Jinan 250062, China.

Although Traditional Chinese Medicine (TCM) preparations have long history with successful applications, the scientific and systematic quality assessment of TCM preparations mainly focuses on chemical constituents and is far from comprehensive. There are currently only few primitive studies on assessment of biological ingredients in TCM preparations. Here, we have proposed a method, M-TCM, for biological assessment of the quality of TCM preparations based on high-throughput sequencing and metagenomic analysis. We have tested this method on Liuwei Dihuang Wan (LDW), a TCM whose ingredients have been well-defined. Our results have shown that firstly, this method could determine the biological ingredients of LDW preparations. Secondly, the quality and stability of LDW varies significantly among different manufacturers. Thirdly, the overall quality of LDW samples is significantly affected by their biological contaminations. This novel strategy has the potential to achieve comprehensive ingredient profiling of TCM preparations.

Traditional Chinese Medicines (TCMs) have long been utilized to prevent and treat various diseases in China. It has also been gradually accepted and widely used in many other countries. A "TCM preparation" (or "patented drug") is characterized by the utilization of multi-herbal materials (including medicinal plants, animal materials and mineral) with their respective dosages by the guidance of Chinese medicine theory and the rule of "King, Vassal, Assistant and Delivery servant"[1], which is more convenient for administration. In recent years, the Chinese export of herbal materials and TCM extracts is larger than that of TCM preparations. For example, in year 2012, China exported 2.02 billion dollars worth of herbal materials and TCM extracts, yet that of TCM preparations was only 270 million dollars[2]. One possible reason for the scarcity of TCM preparation export is the lack of a standardized method to assess the quality, efficacy and safety of TCM preparations. TCM preparations have different types, including pills, capsules, powders, tablets, and so on. Generally, they contain both plant and animal ingredients, which often include misidentified herbal materials, or adulterants, or even biological contaminants during the complex manufacture procedures. Therefore, there is an urgent need to develop a thorough and standardized method for TCM preparation assessment.

To date, the most frequently used methods for the chemical constituent analysis of TCMs are various chromatographic and spectroscopic methods[3–6]. However, these targeted approaches could only measure the chemicals of interests, yet could not assess the contaminated ingredients. Understanding the biological ingredients of TCMs is a prerequisite for ensuring their quality and therapeutic effects, especially for TCM preparations with heterogeneous origins. However, the biological ingredients of TCM preparation are seldom analyzed thoroughly by conventional methods. The biological ingredients of TCM preparation include both prescribed species and contaminated species. "Prescribed species" refer to ingredients listed on the package, which are the integral parts

of the formula to achieve desirable efficacy; while "contaminated species" refer to ingredients not listed on the package, which are usually considered useless for efficacy, reduce the efficacy or even cause the side effects. Current methods for biological ingredient analysis (or species identification) of TCMs include Thin Layer Chromatography (TLC) identification by comparison with reference herbal materials or targeted compounds[7–10] as well as DNA analysis[11,12]. Among these methods, TLC is simple, low-cost and easy to operate. But it is neither accurate nor specific. DNA analysis has recently been introduced for quality evaluation of TCMs. Many reports have been released regarding DNA-based authentication of herbal materials[13–17], while few[18,19] applied this method to TCM preparations. *Coghlan, et al.* (2012)[18] reported a biological ingredient analysis of TCM preparations based on the second-generation DNA sequencing and metagenomic analysis, in which certain plant and animal ingredients in TCM preparations could be identified.

The second-generation DNA sequencing technology, also referred to as the high-throughput sequencing (HTS) technology, provides a variety of genomic sequencing applications to many researchers in last decades due to its advantages of high-throughput and low-cost. Based on HTS, metagenomic analysis is one of the popular methods for the assessment of taxonomic diversity of biological communities. The main procedure for a typical metagenomic research is based on sequencing data from biomarker amplicon, or shotgun whole-genome metagenomic sequencing (metagenomic WGS). Molecular phylogenetic markers could provide biomarker-based identification and quantification of species in the community, while WGS techniques could theoretically provide all genetic information for the community[20].

TCM preparations are usually prepared based on the combination of medicinal plant and animal materials, which could be considered as "synthesized communities" from the taxonomical constitution point of view. Therefore, the HTS-based metagenomic method could be applied to analyze the unknown biological ingredients in TCM preparations. For example, Coghlan's research[18] identified TCM products based on deep sequencing, in which the plastid *trnL* gene was considered as the biomarker for plant medicinal preparations. The advantage of quality evaluation of TCM preparations based on metagenomic approach via HTS is that it can determine not only the prescribed species but also contaminated species.

A proper biomarker is important for the quality assessment of TCM preparations via HTS. The ribosomal internal transcribed spacer 2 (ITS2) has been used as a standard molecular marker to identify medicinal plants for its high inter-specific and intra-specific divergence power[21–23]. Besides, the 5.8S and 28S regions, which are located at the two ends of ITS2 sequence, are conserved enough for primer design[24]. The chloroplast genome *trnL* (UAA) intron, which has been widely used for identifying plant species[25], might also serve as a biomarker target region. In addition, *trnL* is a short fragment that can be easily amplified in heavily degraded DNA samples[26], such as processed TCMs.

In this study, we have selected Liuwei Dihuang Wan (LDW) as the target for biological ingredient analysis via HTS. LDW is a classical TCM preparation based on a six-herb formula, which has been widely used in China[27]. The traditional six herbal materials of LDW are reported as *Rehmannia glutinosa* Libosch., *Cornus officinalis* Sieb. et Zucc., *Paeonia suffruticosa* Andr., *Dioscorea opposita* Thunb., *Poria cocos* (Schw.) Wolf and *Alisma orientalis* (Sam.) Juzep. Among them, *R. glutinosa* and *C. officinalis* are processed products under steaming, while the others are raw materials. To prepare LDW, these six herbal materials are crushed to powder, then mixed and molded into pills together with either honey or water[28]. We have developed a metagenomic approach, which is referred to as M-TCM for the identification of biological ingredients of TCM preparations. We have used the M-TCM method (in which ITS2 and *trnL* (p-loop) were chosen as biomarkers) to analyze nine commercial LDW specimens from three manufacturers in three batches, together with one reference specimen prepared with six herbal materials. We have also evaluated the possible technical and biological biases of this method based on the sequences for prescribed and contaminated species in each sample, as well as the phylogenetic analysis of all biological ingredients. We also assessed the biological ingredients for both reference and commercial LDW samples, based on which the detectabilities and sensitivities of biomarkers have been assessed. The stability of different batches of samples and the effects of biological contaminations on the overall quality of LDW samples were also evaluated by comparison among reference and commercial samples using PCA and clustering analyses.

Results and Discussion

Sanger sequencing of six herbal materials in LDW. The genomic DNA was extracted from six prescribed herbal materials of LDW bought from the drugstore, *trnL* and ITS2 regions were amplified, followed by Sanger DNA sequencing. The sequencing data was added to our reference database (see "Methods" for details) for biological ingredient analysis of LDW preparations.

Results based on Sanger sequencing indicated that ITS2 and *trnL* sequences of *P. suffruticosa* and *A. orientalis* were consistent with those references in NCBI database. As the fungal ingredient, *P. cocos* does not contain *trnL* DNA sequence, yet contains ITS2 sequence. We used universal ITS2 primers to amply DNA extracted from *P. cocos* sample and then perform TA cloning. However, all PCR fragments were identified as *Vigna* species (with 98% similarity and E-value of 0.0 based on ITS2 sequencing results) and no *P. cocos* related ITS2 sequence was detected from 24 clones. Therefore, we designed a new pair of primers (Poria-F and Poria-R, see Supplementary Table S1 for details) according to 11 ITS2 sequences of *P. cocos* from GenBank[29], and the corresponding PCR product matched *P. cocos* (with 99% similarity). These results also indicated that the herbal material of *P. cocos* was probably contaminated by *Vigna* species. The sequenced *trnL* gene of *D. opposita* was consistent with those in NCBI database. As for the *R. glutinosa* and *C. officinalis* PCR products, their sequences could not be obtained due to lack of PCR amplification. Both *R. glutinosa* and *C. officinalis* are processed TCM products. The process of *R. glutinosa* materials involved steaming, drying and stewing, while stewing was used for the preparation of *C. officinalis*. Hence their DNA could have been damaged to different degrees, making it difficult for biomarker detection.

From the above results, we inferred that ITS2 and *trnL* could be used as biomarkers for plant ingredient identification. However, *P. cocos* could not be detected by the universal ITS2 primers. Therefore, when fungus species were present in TCM, additional molecular markers and primers would be needed. Furthermore, a vital limitation for all DNA marker based approaches is that they might face limitation for evaluation of processed materials with damaged DNA.

General overview of high-throughput sequencing results. We have randomly purchased 9 commercial LDW specimens from 3 manufacturers (MH, MS and MT) each with 3 batches (referred to as A, B and C, respectively) (see Supplementary Table S2). A reference LDW specimen (referred to as RE) was prepared using six prescribed herbal materials according to the Chinese Pharmacopeia (version 2010)[28]. We have used 3 biological replicates from each of these 10 specimens, and generated HTS data for all of these 30 LDW samples, based on 454 GS FLX Titanium sequencing.

After HTS data quality control, the 454 GS FLX Titanium sequencer generated ~80,000 trimmed and filtered reads for all samples (see Supplementary Table S3), included 46,987 and 29,651 reads for ITS2 and *trnL* regions, respectively. Among all these reads, there were 6,402 ITS2 and 2,151 *trnL* reads for 3 RE samples; 20,069 ITS2 and 7,431 *trnL* reads for 9 MH samples; 8,697 ITS2 and 10,849 *trnL* reads for 9 MS samples; 11,819 ITS2 and 9,288 *trnL* reads for 9 MT

samples. We also set a cutoff parameter for the assessment of possible adulterations and contaminations. Considering the incomplete *trnL* database comparing to ITS2 database, we filtered the *trnL* sequences for which the corresponding possible species was evidenced by only 1 read, and ITS2 sequences for which the corresponding possible species was evidenced by 3 or less reads.

In total, we obtained 1,566 ITS2 sequences and 988 *trnL* sequences per sample (all 30 samples for RE, MH, MS and MT). On averages, 3 and 2.4 prescribed species could be detected from a sample based on ITS2 and *trnL*, respectively. In addition, up to 7 contaminated species were found from these LDW samples with a mean of 1.8 through ITS2 sequencing, and up to 4 contaminated species with a mean of 0.4 through *trnL* sequencing (Figure 1). Such differences in the number of prescribed and contaminated species based on different biomarkers might be ascribed to the database coverage and resolution of these two biomarkers.

Assessment of possible technical and biological biases. Before comparison of various LDW samples, we have assessed the possible technical and biological biases that might be introduced during experiments. First, the possible biases introduced from sequencing were evaluated, namely whether more sequences or identifiable ingredients for a sample lead to more contaminated species that could be identified. The correlations between number of detected contaminated species in all samples and the corresponding sequence number, as well as the correlation between number of contaminated species and prescribed species were studied, and then Pearson correlation coefficients (R^2) and related significance values (p-value) were calculated (Figure 2). Dots in Figure 2 were all discrete, and failed to present linear dependence relation. R^2 and p-value further denied their correlativity, suggesting no significant technical biases. In addition, we have generated rarefaction curves for both ITS2 and *trnL* datasets (Figure 3) to illustrate the number of sequences that were sufficient to detect all components in LDW. It was observed that the number of predicted species have not changed with the increasing number of reads, which indicated that for most samples, the current sequencing depth was sufficient for detecting the majority of species in LDW samples in this experiment.

The possible biological contaminations introduced during the process of DNA extraction and PCR amplification were checked by the phylogenetic analysis of species detected in LDW samples. Figure 4 (a) showed that besides the prescribed species, DNA from 17 organisms were detected based on ITS2 sequencing results in LDW from different manufacturers or batches, which scattered widely in 10 orders, 12 families and 17 genera. We also detected 7 other organisms apart from the prescribed species based on *trnL* sequencing results, which belonged to different orders and families (Figure 4 (b)). So it appeared that the contaminated species occurred randomly. In addition, Figure 4 illustrated that the types and relative abundances of contaminated species varied between different batches and manufacturers, based on which we inferred that there was no observable biological contamination resulted from the experiment procedures. Based on both technical and biological bias analyses, we concluded that the HTS data for LDW samples were relatively clean and was suitable for biological ingredient analysis.

Biological ingredient analysis of reference LDW samples. We have obtained over 8,000 sequence reads for three reference LDW samples (RE.1, RE.2 and RE.3) using HTS approach with both ITS2 and *trnL* biomarkers. Using ITS2 as biomarker, two prescribed herbal materials, *A. orientalis* and *P. suffruticosa* were detected in all three samples (Table 1). In addition, *Vigna* genus, a possible contaminated species, was detected in all three samples, which was in concordance with the results from the single herbal material *P. cocos*. As *trnL* was used as biomarker, three prescribed herbal materials, *A. orientalis*, *D. opposita* and *P. suffruticosa* were detected in all three samples, and no other species was detected (Table 2).

The results of this metagenomic study and previous Sanger sequencing results on six herbal materials were very consistent. Three prescribed species and one contaminated species could be identified based on ITS2 sequencing, while three prescribed species could be identified based on *trnL* sequencing from RE samples. It confirmed the feasibility of the untargeted M-TCM method for biological ingredient analysis of TCM preparations.

Biological ingredient analysis of commercial LDW samples. An in-depth analysis of identified species in 27 commercial LDW samples was performed using the same method as that for RE (reference LDW specimen) (Table 1 and Table 2). A total of 16 and 11 plant families were identified in this study with 40,585 ITS2 and 27,535 *trnL* reads (on average 1,503 ITS2 and 1,019 *trnL* sequence reads per sample) for 27 samples (Table 1 and Table 2). For prescribed species, *P. suffruticosa* was detected from all samples based on both biomarkers with relatively high abundance (Figure 5 and Figure 6), which suggested that DNA of *P. suffruticosa* was of high quality and easy to be extracted. *A. orientalis* was identified from 70% (19) and 30% (8) of LDW samples based on ITS2 and *trnL* respectively, and 70% (19) in combination (union). For *D.*

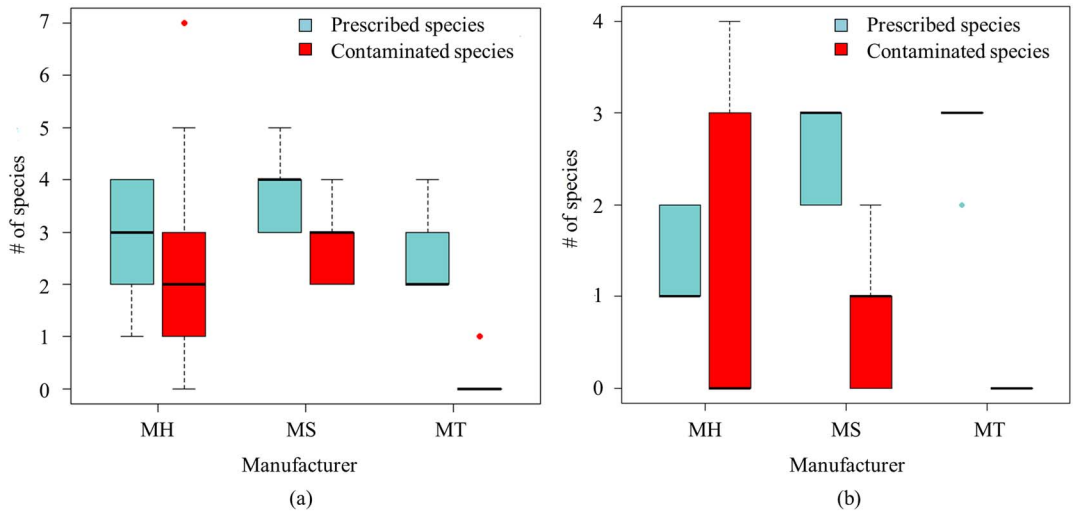


Figure 1 | Number (#) of prescribed species and contaminated species detected in LDW samples from different manufacturers. (a) Species composition analysis based on ITS2 sequencing results. (b) Species composition analysis based on *trnL* sequencing results.

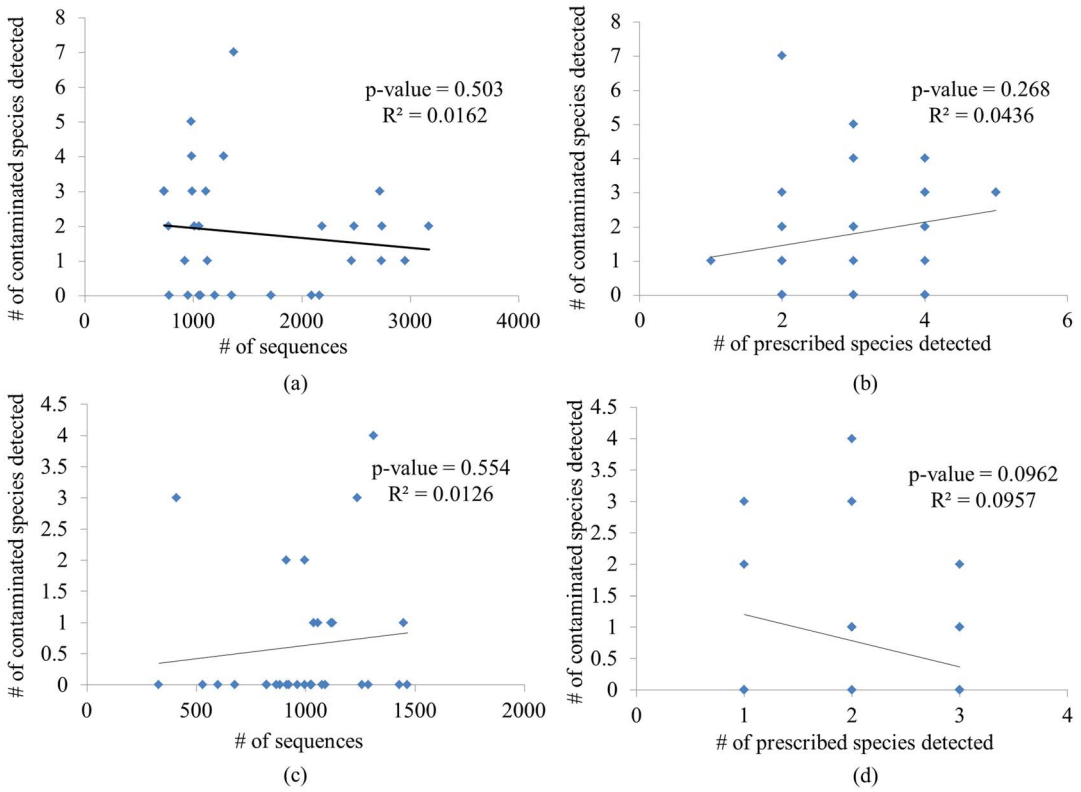


Figure 2 | Correlation among the number (#) of contaminated species, number of sequencing reads and number of prescribed species. (a) Correlation between number of contaminated species and sequence number based on ITS2. (b) Correlation between number of contaminated species and prescribed species based on ITS2. (c) Correlation between number of contaminated species and sequence number based on *trnL*. (d) Correlation between number of contaminated species and prescribed species based on *trnL*.

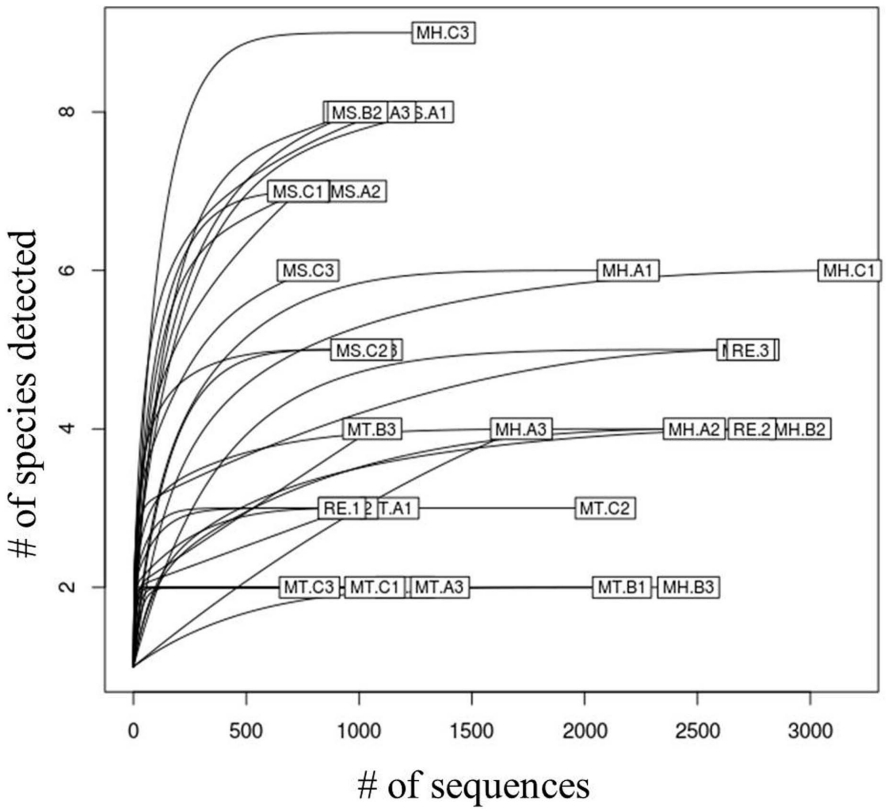
opposita, sequencing success rate for *trnL* (78%) was higher than ITS2 (45%), and 93% (25) by combination of the two regions. *C. officinalis* was detected from all MS samples by ITS2 region, but none of MH samples. *R. glutinosa* was detected from 50% of all samples by ITS2 region, mainly from MH and MS samples. A possible reason for the different identification profiles of the processed herbal materials, namely *R. glutinosa* and *C. officinalis*, might be that the genomic DNA of herbal materials used in three manufacturers were damaged or lost due to various pre-processing procedures and storage. It was also possible that some manufacturers have replaced processed herbal materials with raw ones. In addition, we could not detect *P. cocos* in any LDW samples using the original ITS2 primers, while it could be detected by applying Sanger sequencing with specific primers.

As for the contaminated species (Table 1 and Table 2), 3 of the most common plant families were identified as Convolvulaceae, Fabaceae and Plantaginaceae, based on ITS2 and *trnL* results in combination. Among them, the *Ipomoea* genus of Convolvulaceae were detected from all MH samples and 33% of MS samples; Fabaceae were detected from 40% of samples, including *Vigna*, *Robinia* and *Glycine* genera; *Plantago* genus of Plantaginaceae were detected from 67% of MS samples.

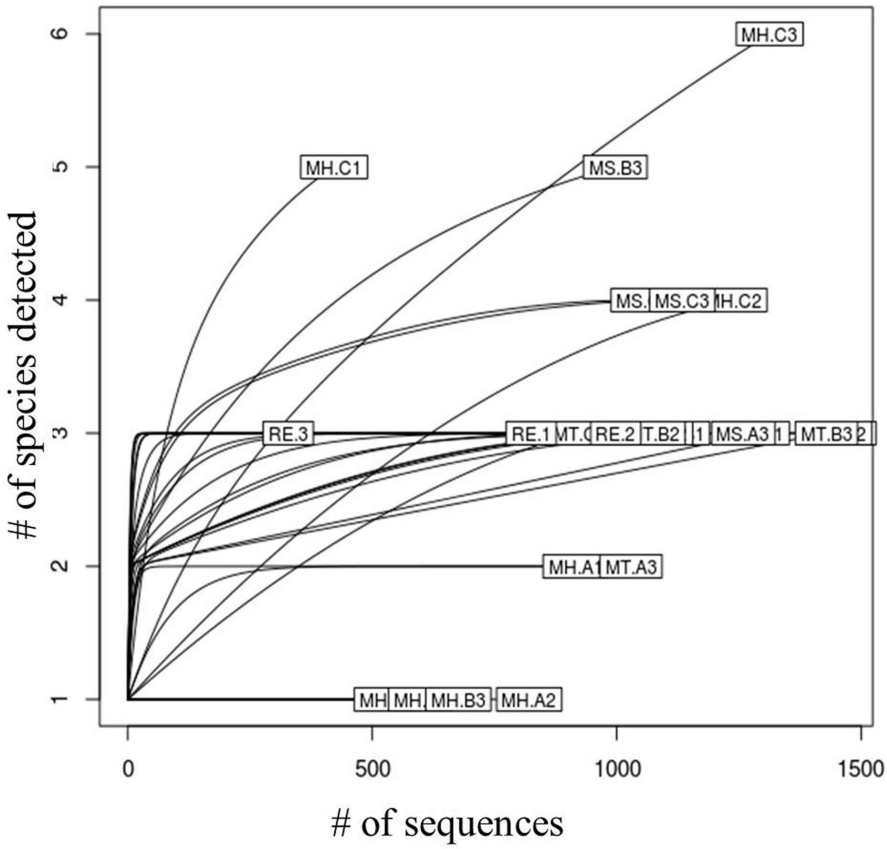
Hence, we concluded that due to the diverse properties of herbal plants and medicinal parts, different pre-processing procedures of Chinese medicinal materials, as well as various production processes of TCM preparations, DNA isolated from different TCM preparation could be highly variable in terms of quality and concentration. DNA of some species could fail to be amplified due to no DNA or severe PCR inhibition, and the preference of methodologies in extraction and amplification also made the relative abundance of different species hard to be an evidence for quantitative measure. It was worthwhile mentioning that DNA from *R. glutinosa* and *C. officinalis* were found in some batches of MH and MS based on HTS while they were undetected by Sanger sequencing for respective herbal materials. One possible reason was that the sequencing depth of HTS experiment was significantly greater than that of the Sanger method (Figure 3), thus the better sensitivities. Another possible reason was that the manufacturers might have followed different herbal material selection standard, and some processed herbal materials might have been unintentionally replaced by unprocessed or incompletely processed ones for the production of TCM preparation. The results would be significant as DNA of unprocessed or incompletely processed herbs might be detected by sequencing, while that of the processed herbs could not be detected. To test this, we have also conducted DNA extraction experiment and ITS2 or *trnL* amplification experiment for both processed and unprocessed *R. glutinosa*. Results have shown that the DNA for unprocessed *R. glutinosa* could always be amplified, while that of processed *R. glutinosa* could not be amplified, which confirmed above inference about different detectabilities of unprocessed and processed herbal materials.

As seen in Table 1 and Table 2, the results based on ITS2 and *trnL* biomarkers were not consistent. ITS2 is a 500 bp DNA that is used as a standard molecular marker to identify medicinal plants, while *trnL* is 200 bp that could be easily amplified in several highly degraded templates such as those in processed TCM preparations. This suggests that the choice of biomarkers could influence the result. The intention for choosing these two biomarkers was that by using both of these two biomarkers, more species could be detected, and their intersection might indicate more reliable identifications.

Additionally, we have observed that these biomarkers have different resolutions during the identification of biological contaminations from LDW samples. When we performed database search of biomarkers (see Methods for details), biological ingredients in the samples could be assigned to the species level based on unique top hit.



(a)



(b)

Figure 3 | Rarefaction curves between the number (#) of sequence reads and the number of detected species. (a) Rarefaction curves based on ITS2. (b) Rarefaction curves based on *trnL*.

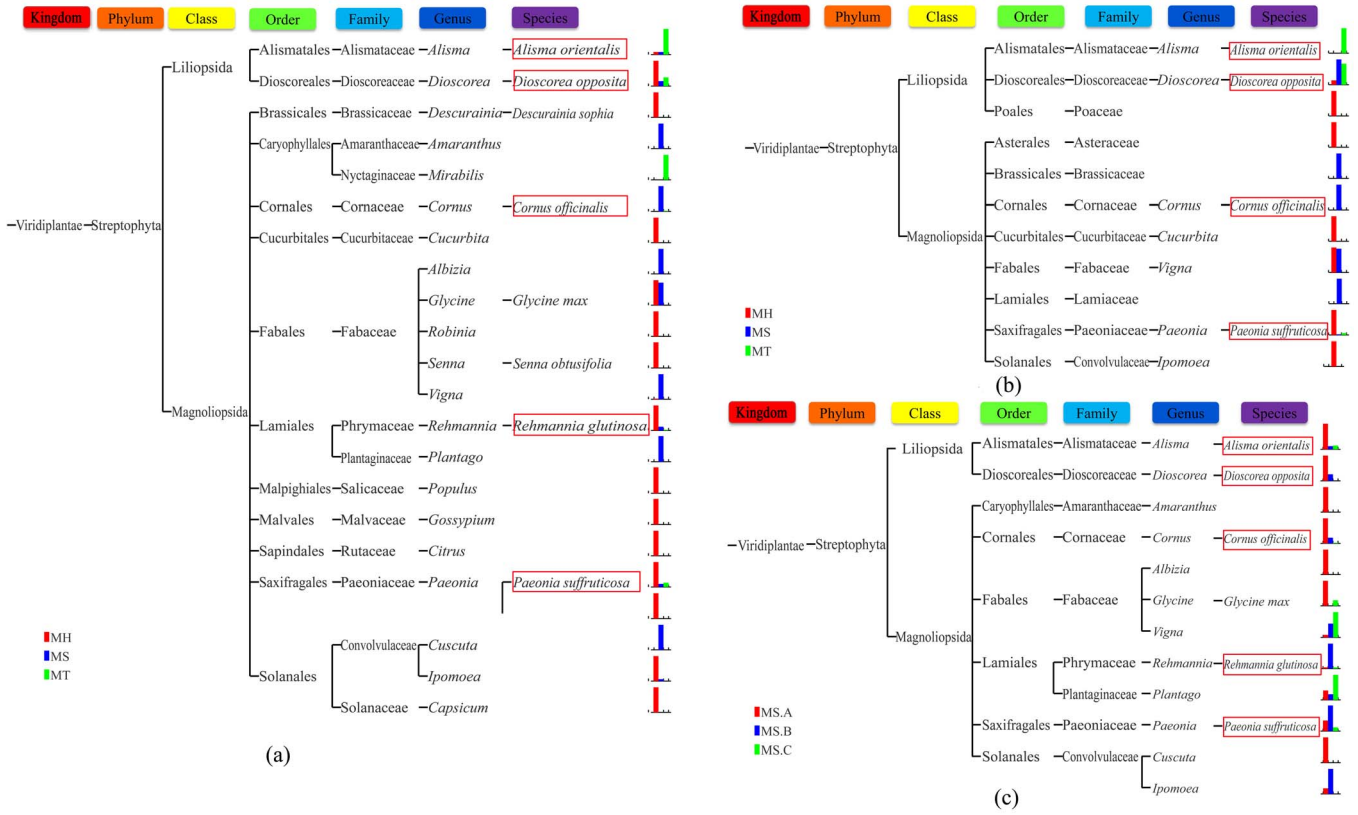


Figure 4 | Phylogeny and relative abundance of species detected in LDW samples from different manufacturers and batches. (a) Phylogeny and relative abundance of species detected in LDW from 3 manufacturers (MH, MS and MT) based on ITS2. (b) Phylogeny and relative abundance of species detected in LDW from 3 manufacturers (MH, MS and MT) based on *trnL*. (c) Phylogeny and relative abundance of species detected in LDW from 3 batches (MS) based on ITS2. Species marked in red boxes are prescribed species, while others are contaminated species. Notice that the relative abundance was not only related to the amount of biological ingredients, but also the quality and concentration of DNA during experiment.

Multiple top hits to different species indicated that it was not possible to distinguish some species. We successfully identified the prescribed species. However, for the contaminated species, all ITS2 sequences and 75.6% *trnL* sequences were located to genus level. The remaining *trnL* sequences were only specific to family level, confirming that longer ITS2 biomarker (~500 bp) could provide better resolution than short *trnL* biomarkers.

The *trnL* marker also has a relatively narrower field of application for species identification compared to ITS2. For example, *trnL* cannot be used for *P. cocos*, as mentioned previously. In addition, the ITS2 data from two MS samples suggested that they contained materials from *Cuscuta* genus, while there was no corresponding *trnL* sequencing reads to support this. This might be due to that as holoparasites, *Cuscuta* species exhibited a tendency towards the extreme reduction of intron sequences and non-coding intergenic regions in plastome. In addition, plastid genome is difficult to produce and enrich[30], leading to an insufficient *trnL* sequence comparing to ITS2 sequence. Therefore, as a biomarker for TCM preparations, ITS2 marker might have a broader field of application than the plastid gene-based *trnL* marker.

Hence we suggest that *trnL* data could be used to reinforce or complement ITS2 for the identification of plant species in TCM preparations, and optimization was needed to improve the sensitivity and accuracy for DNA sequence-based biological ingredients analysis of TCM preparations. Since both the sensitivity and accuracy depend on the reference sequences, a comprehensive database with sufficient species was needed. On the other hand, biomarkers themselves need to be optimized in term of DNA length for the highly degraded DNA samples isolated from TCM preparations.

Comparison among commercial and reference LDW samples. Although the M-TCM method was unable to provide a quantitative measurement for each ingredient, the presence of some major prescribed species could still be examined, which would contribute significantly for the effectiveness of TCM preparations. Besides *R. glutinosa* and *C. officinalis*, whose DNA was destroyed during preprocessing procedures, and *P. cocos* that lacks information regarding its ITS2 sequence, *P. suffruticosa* and *D.opposita* were detected in all samples, whereas *A. orientalis* was not detected from most MH samples based on both ITS2 and *trnL* sequencing.

Additionally, the detection of contaminated species is of crucial importance for the safety of TCM preparations[18]. Comparison among commercial and reference LDW samples have shown that there were significant differences in contaminated species profiles for samples obtained from different manufacturers. It was observed that based on ITS2 and *trnL* sequencing results, *Ipomoea* genus and many other organisms have been identified from MH samples, *Ipomoea*, *Plantago, Vigna* genera and so on have been identified from most MS samples (Table 1 and Table 2). Therefore, different commercial samples might have non-negligible differences in term of quality and safety, as contaminated species might contain toxic compounds. For example, *Senna obtusifolia*, which could potentially induce liver and kidneys damage[31], was detected from MH.C3 sample as a contaminated species.

Finally, consistency is an important consideration in quality evaluation of TCM preparations. From the principle component analysis (PCA) based on ITS2 and *trnL* sequencing results (Figure 7), 3 RE samples were clustered together, indicating the stability of self-made reference samples. 9 MT LDW samples were also closely clustered,

Table 1 | Plant genera identified in 30 LDW samples using HTS based on ITS2. Notice that *P. cocos* was not discovered in any LDW samples based on HTS results with universal ITS2 primers, but its existence was confirmed by specific primers on selected samples

	MH									MS									MT									RE		
Sample ID	A1	A2	A3	B1	B2	B3	C1	C2	C3	A1	A2	A3	B1	B2	B3	C1	C2	C3	A1	A2	A3	B1	B2	B3	C1	C2	C3	1	2	3
Alisma orientalis	✓		✓				✓			✓		✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cornus officinalis										✓	✓	✓	✓	✓	✓	✓	✓	✓					✓							
Dioscorea opposita	✓		✓	✓	✓		✓	✓		✓	✓	✓		✓										✓		✓				
Paeonia suffruticosa	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Rehmannia glutinosa	✓	✓	✓		✓		✓	✓	✓			✓	✓	✓	✓	✓		✓					✓							
Fabaceae																														
- *Albizia*											✓																			
- *Glycine*									✓	✓	✓					✓														
- *Robinia*	✓							✓																						
- *Senna*									✓																					
- *Vigna*								✓			✓	✓	✓	✓	✓	✓	✓	✓										✓	✓	✓
Amaranthaceae																														
- *Amaranthus*										✓																				
Solanaceae																														
- *Capsicum*									✓																					
Rutaceae																														
- *Citrus*				✓																										
Cucurbitaceae																														
- *Cucurbita*								✓	✓																					
Convolvulaceae																														
- *Cuscuta*										✓	✓																			
Brassicaceae																														
- *Descurainia*		✓																												
Malvaceae																														
- *Gossypium*							✓	✓	✓																					
Convolvulaceae																														
- *Ipomoea*	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓															
Nyctaginaceae																														
- *Mirabilis*																			✓											
Paeoniaceae																														
- *Paeonia*				✓																										
Plantaginaceae																														
- *Plantago*												✓	✓	✓		✓	✓	✓												
Salicaceae																														
- *Populus*									✓																					

Table 2 | Plant families or genera identified in 30 LDW samples using HTS based on *trnL*

	MH									MS									MT									RE		
Sample ID	A1	A2	A3	B1	B2	B3	C1	C2	C3	A1	A2	A3	B1	B2	B3	C1	C2	C3	A1	A2	A3	B1	B2	B3	C1	C2	C3	1	2	3
Alisma orientalis																			✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Cornus officinalis										✓	✓																			
Dioscorea opposita	✓						✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Paeonia suffruticosa	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Asteraceae							✓	✓	✓																					
Brassicaceae															✓															
Convolvulaceae																														
- Ipomoea			✓				✓	✓	✓																					
Cucurbitaceae																														
- Cucurbita									✓																					
Fabaceae																	✓													
- Vigna			✓											✓	✓	✓	✓	✓												
Lamiaceae													✓																	
Poaceae							✓	✓	✓																					
- Gleditsia		✓																												

whereas MH and MS samples were relatively far apart. We could also observe that there are large variations among MH samples, because of the existence of contaminated species such as *Senna*, *Robina*, *Paeonia*, *Cucurbita* in some samples based on ITS2 (Figure 7 (a) and Table 1) and *Asteraceae*, *Convolvulaceae*, *Poaceae* based on *trnL* (Figure 7 (b)). The diversity of MS samples was mainly derived from the existence of *Cuscuta*, *Plantago*, *Vigna*, *Glycine* from 9 samples based on ITS2.

To further explore the differences among commercial and reference LDW samples, we performed clustering analysis using average-linkage hierarchical clustering algorithm based on the Euclidean distances of the species in LDW samples (refer to "Methods" section for details). Figure 8 showed that all MT samples clustered together based on both ITS2 and *trnL* biomarkers, which further indicated their intra-group similarity. Results based on ITS2 sequencing (Figure 8 (a) and Table 1) showed that MS samples had a bigger intra-group variation mainly resulted from *Plantago* and *Vigna* genera and obvious inter-group difference originated from *Glycine* and *Cuscuta* genera. In addition, for MH samples, batch A and B clustered together with other commercial samples and reference LDW

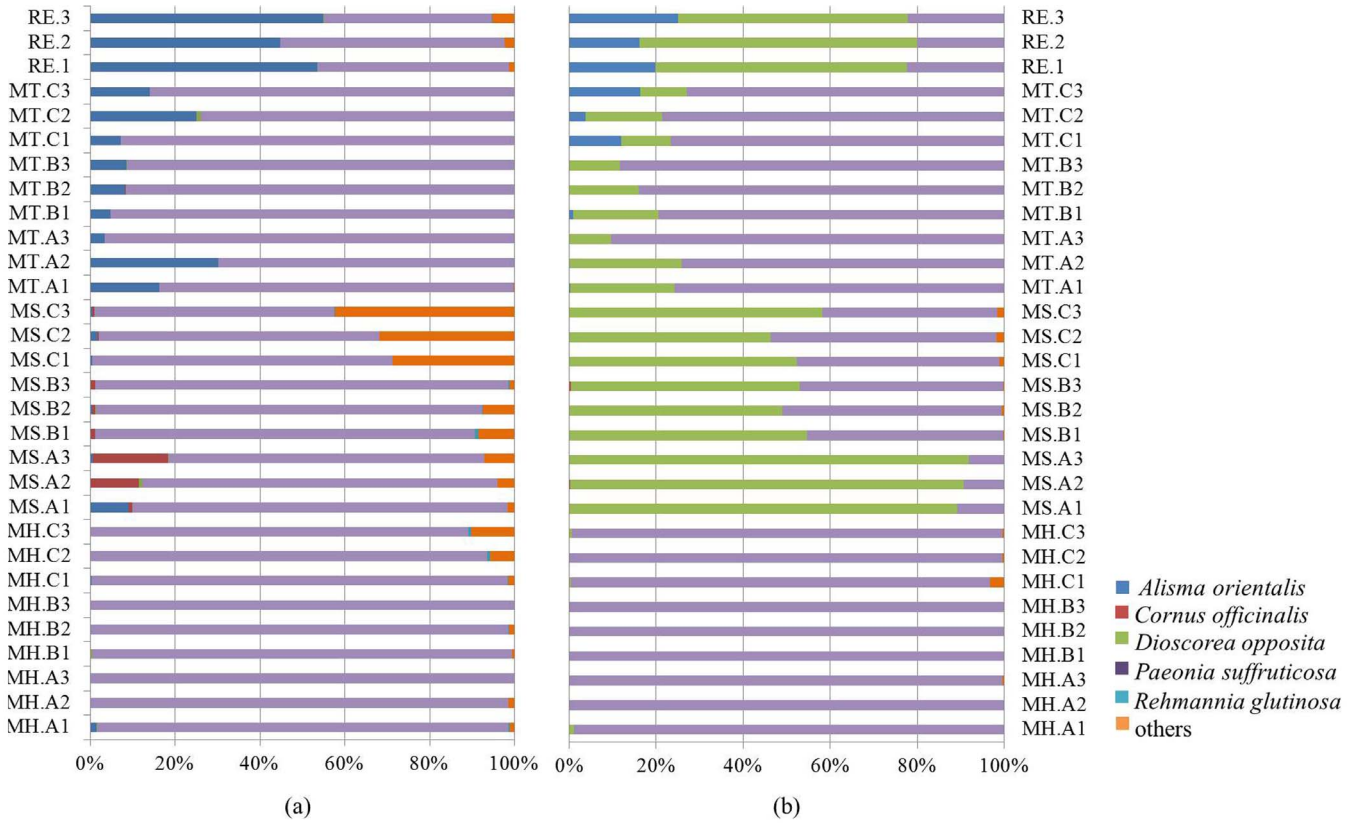


Figure 5 | Prescribed species analysis results for LDW samples. (a) Identification results based on ITS2. (b) Identification results based on *trnL*. Notice that *P. cocos* was not discovered in any LDW samples based on HTS results with universal ITS2 primers, but its existence was confirmed by specific primers on selected samples. In addition, the relative abundance was not only related to the amount of biological ingredients, but also the quality and concentration of DNA during experiment.

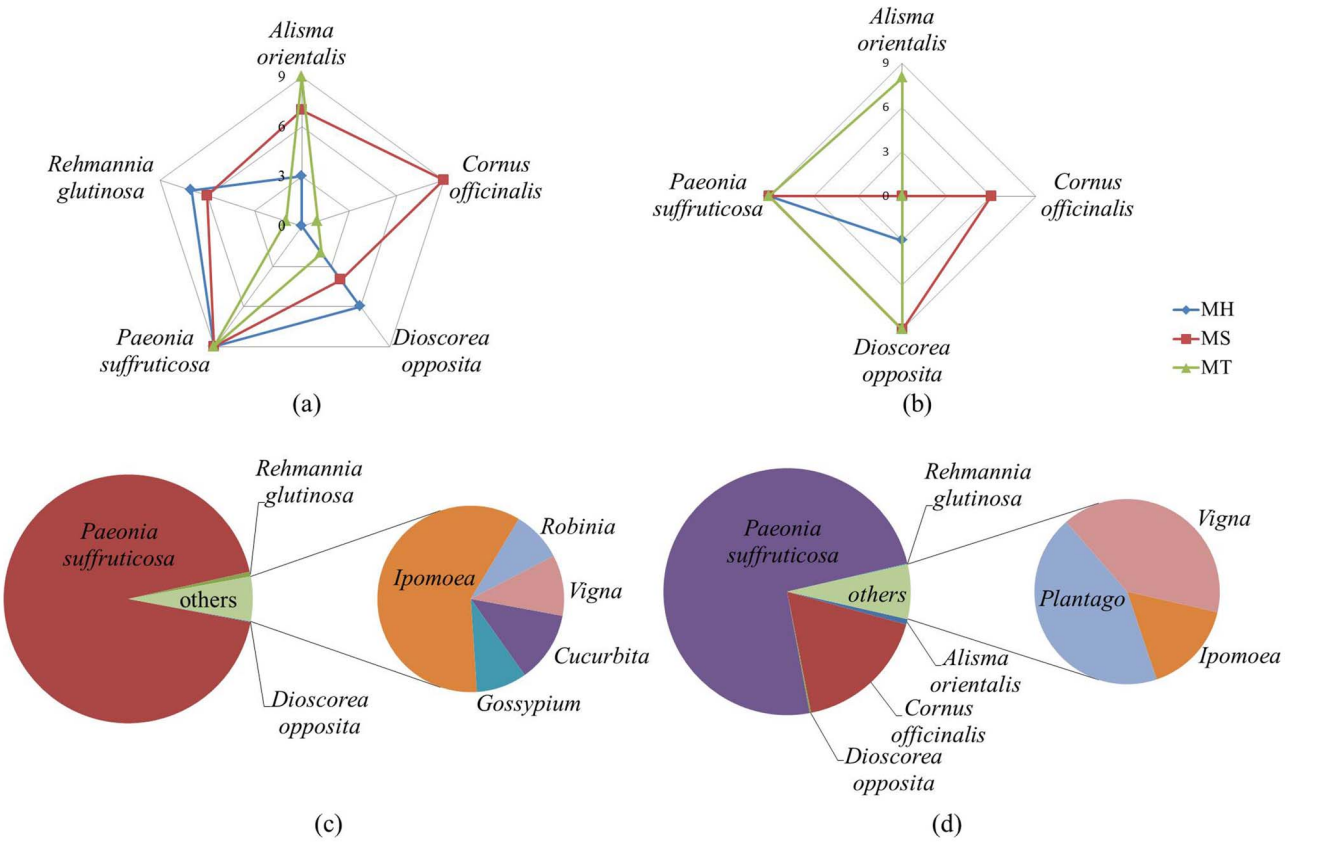


Figure 6 | Radar plot and pie-chart showing the component organism information for LDW samples. There were five prescribed species detected by ITS2 and four prescribed species detected by *trnL* from 27 commercial samples, radar analysis was performed based on the number of samples (within nine samples of different manufacturers) in which each prescribed species was detected. (a) Radar plot based on ITS2 identification results. (b) Radar plot based on *trnL* identification results. (c) ITS2 identification result of prescribed species and contaminated species in MH.C2. (d) ITS2 identification result of prescribed species and contaminated species in MS.A3. Notice that *P. cocos* was not discovered in any LDW samples based on HTS results with universal ITS2 primers, but its existence was confirmed by specific primers on selected samples.

samples, while samples of batch C have so much differences that they presented a separate set of profiles. This might be caused by some contaminated species detected in MH.C samples but not in MH.A and MH.B samples, such as *Cucurbita* and *Gossypium* genera based on ITS2 results (Figure 8 (a) and Table 1) and Asteraceae, Convolvulaceae and Poaceae family based on *trnL* results (Figure 8 (b) and Table 2). These clustering results were also in accordance with PCA analysis results in Figure 7.

We concluded that samples from manufacturer MT had higher consistency, and the biological contaminations could significantly affect the overall quality of LDW samples. As some contaminations might be introduced to TCM preparations from manufacturing equipment or factory environment pollution, better manufacturing process would be crucial for ensuring the quality and safety of TCM preparations.

Conclusion

Our study has demonstrated the potentials of a metagenomic-based method (M-TCM) in analyzing the biological ingredients of TCM preparations. The eminent characteristic of M-TCM is that it can determine both prescribed and contaminated species simultaneously and indiscriminately. Our results have also illustrated the ability of M-TCM on evaluation of different TCM samples. It is particularly suitable for TCM preparations with multiple and complex ingredients. In addition, this method could be applied to biological ingredient analysis of other herbal and food products.

The HTS and metagenomic analysis have their limitations. First, it is unable to identify ingredients due to DNA degradation during TCM processing or the lack of reference sequence in the database. In addition, since PCR is used to amplify biomarkers before sequencing, the design of the "universal" primers could limit its resolution to a specific range of plant and animal. A balanced combination of biomarkers that could be used to identify all potential ingredients in TCM preparations is often difficult to achieve. Furthermore, this method is also not quantitative as the amount of remaining DNA does not correspond to the amount of an ingredient, while DNA-free components such as inorganic materials and extracts could not be quantified. Considering that TCM preparations would usually contain both biological and chemical ingredients, a combination of the genetic approaches for species identification with analytical chemistry approaches for compounds determination could better assess the quality of TCM preparations. In addition, it has been reported that the combination of TCM preparation assessment methods with clinical observations[32] would provide a holistic view for the TCM preparations, from their contents to effects.

Methods

Sample collection. 9 commercial LDW specimens were randomly purchased from 3 different Chinese manufacturers (namely MH, MS and MT) each with 3 lot numbers (A, B and C) (see Supplementary Table S2). Each batch was implemented with three biological replicates, therefore there were totally 3*3*3 = 27 commercial LDW samples. In addition, according to Chinese pharmacopeia[28], the reference LDW (RE) was made in-house with *R. glutinosa* (Rehmanniae radix preparata, 4 g), *C. officinalis* (Corni fructus preparata, 2 g), *P. suffruticosa* (1.5 g), *D.opposita* (2 g), *P. cocos* (1.5 g), *A. orientalis* (1.5 g) and refined honey (12 g). Above six herbal materials were purchased

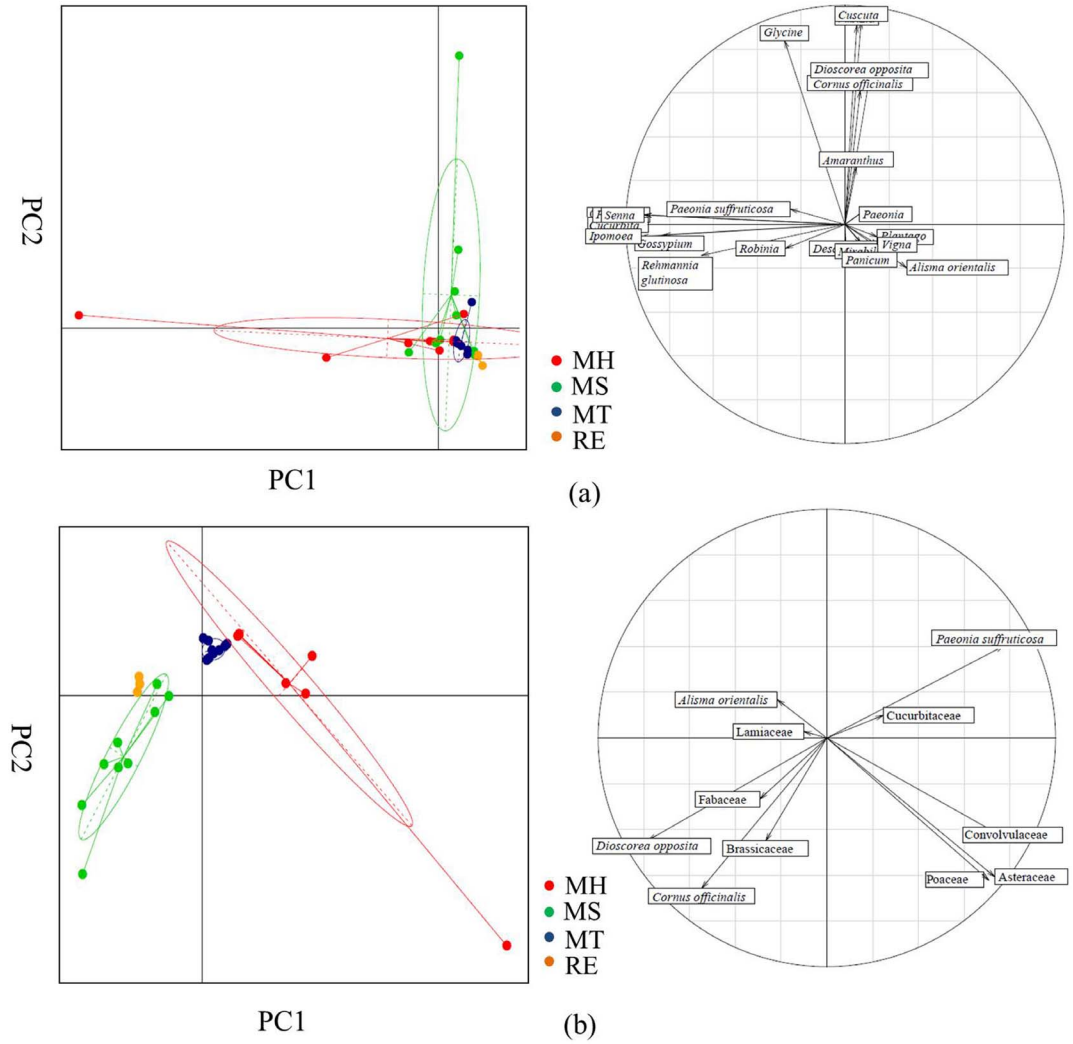


Figure 7 | Component species abundance differentiated LDW reference and commercial samples. Principal component analysis (PCA) of biological ingredients, based on the identity and abundance of prescribed and contaminated species, was carried out with 3 self-made reference LDW samples (RE) and 27 commercial LDW samples from 3 manufacturers (MH, MS and MT). (a) PCA analysis and loading figure based on ITS2. (b) PCA analysis and loading figure based on *trnL*. In PCA plot, x-axis and y-axis represent the discrimination results for PC1 and PC2 respectively.

from drugstore and identified by one of the authors B. H. The RE specimen was implemented with three biological replicates, therefore there are totally 3 reference LDW samples.

DNA extraction and quantification. Each sample (2.5 g) was completely dissolved with 0.1 M Tris-HCl (pH 8.0, 5 mL). Dissolved solution (1 mL) was diluted with extraction buffer (4 mL) consisting of 0.1 M Tris-HCl (pH 8.0), 20 mM EDTA (pH 8.0), 2% CTAB, 1.4 M NaCl and 1% SDS, then 10 μL 10 mg/mL proteinase K and 100 μL β-mercaptoethanol was added and incubated at 65°C for 3 h with occasional swirling. Protein was removed by extracting twice with an equal volume of phenol: chloroform: isoamyl-alcohol (25: 24: 1), and once with chloroform: isoamyl-alcohol (24: 1). The supernatant was incubated at −20°C with 0.6 folds of cold isopropanol for 1 h to precipitate DNA. The precipitate was washed with 70% ethanol and then dissolved and diluted to 5 ng/μL with TE buffer and used as template for PCR[33]. DNA concentration was quantified on Synergy HT Multi-Mode Microplate Reader (BioTek).

DNA amplification and DNA sequencing. Individual amplifications of ITS2 and *trnL* were carried out in 25 μL total volume including 10 ng template DNA, 5 × PrimeSTAR buffer (Mg^{2+} plus), 0.1 mM dNTPs, 0.2 μM of each primer, 1 μL DMSO, 0.25 μL BSA and 0.25 μL PrimeSTAR® HS DNA Polymerase (Takara, 5 U/μL). For amplification and sequencing of ITS2, primers S2F and S3R (see Supplementary Table S1) with 7 bp MID tags designed and used with the following cycling conditions: initial denaturation at 95°C for 5 min, followed by 40 cycles of 98°C for 15 s, 58°C for 10 s, 72°C for 30 s, and a final extension at 72°C for 10 min. For amplification and sequencing of *trnL*, primers *trnL* c and *trnL* h (see Supplementary Table S1) with 7 bp MID tags designed and used with the following cycling conditions: initial denaturation at 95°C for 5 min, followed by 40 cycles of 98°C for 15 s, 57°C for 10 s, 72°C for 30 s, and a final extension at 72°C for 10 min.

The PCR products (see Supplementary Figure S1 and S2) were electrophoresed on 2% agarose gel and purified with QIAquick Gel Extraction kit (QIAGEN). DNA concentration was measured on Synergy HT Multi-Mode Microplate Reader (BioTek).

trnL and ITS2 sequence amplicon reads were sequenced from 454 GS-Titanium sequencer, yielding a total of 285,720 processed reads (see Supplementary Note 1). These raw data would then be subject to quality control (QC) before sequencing data could be used.

Sequencing data analysis procedure and software configurations. To minimize the effects of random sequencing errors and avoid overestimation of the phylogenetic diversity of the raw data, relatively stringent quality-based trimming of the reads was performed using the MOTHUR software package[34] for quality control. First, we discarded sequences < 150 bp in ITS2 dataset and <75 bp in *trnL* dataset, and sequences that had an average quality score < 20 in each 5 bp-window rolling along the whole read. Then sequences that contained primer mismatches, uncorrectable barcodes, ambiguous bases, or homopolymer runs in excess of 8 bases were also removed from both ITS2 and *trnL* datasets. After that, reads were sorted by tag sequences.

Our reference database was composed of all *trnL* and ITS2 sequences downloaded from GenBank[29], together with *trnL* and ITS2 sequences of six herbal materials determined by Sanger sequencing. Then BLAST searches were performed using a 1E-10 E-value threshold for all datasets. We have checked the BLAST results, and observed that most of the top hits (96.95% of the matches for ITS2 reads and 97.15% of the matches for *trnL* reads) have been obtained with high identity (>98%) to the biomarker sequences of ITS2 and *trnL* in the reference databases. Therefore, query sequences were identified as the top hit in the reference database. Considering the incomplete *trnL* database comparing to ITS2 database, we filtered *trnL* sequence for which the corresponding possible species was evidenced (matched) by only 1 read,

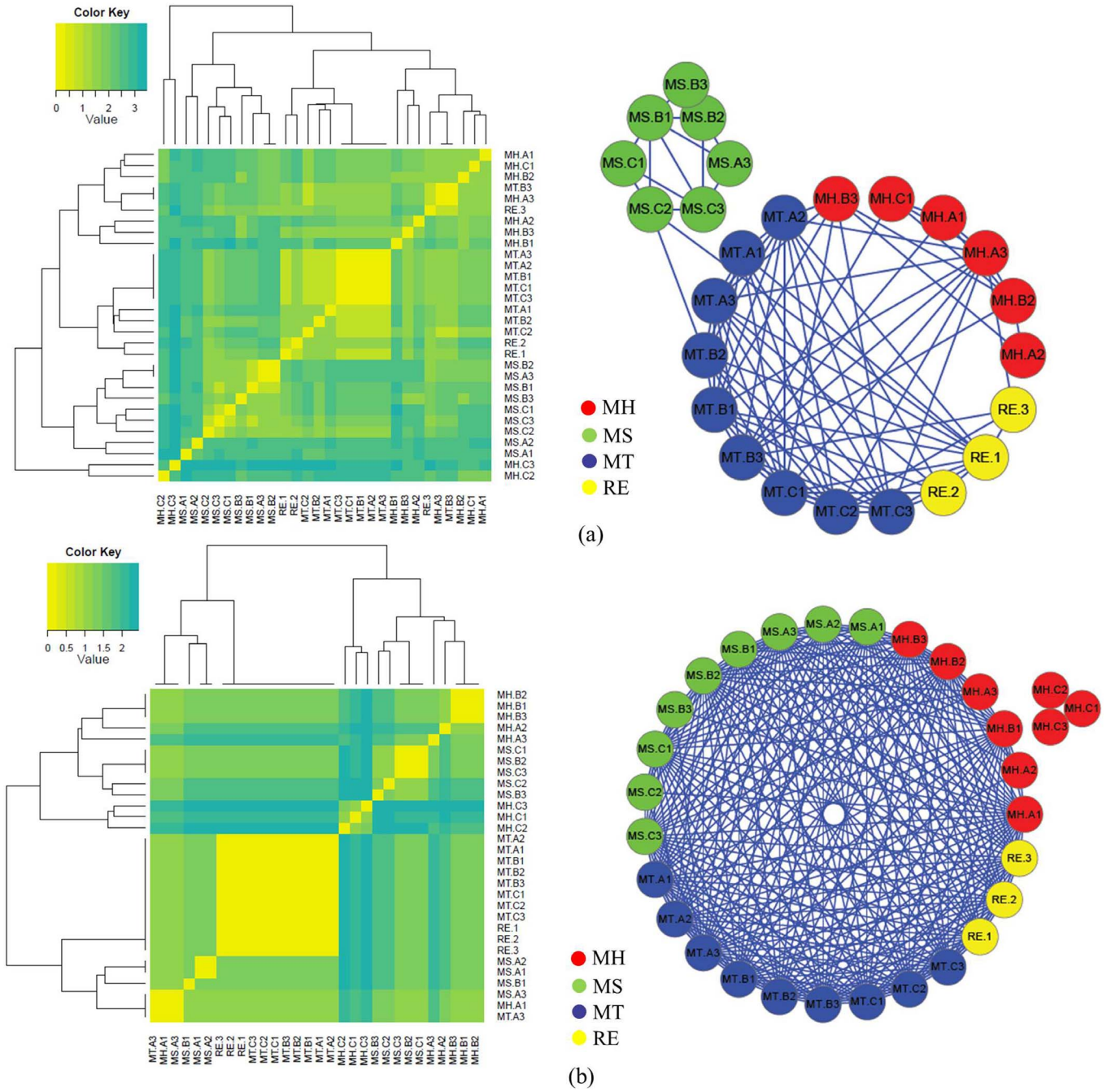


Figure 8 | Clustering of different samples based on their species composition reveals relationships among commercial and reference LDW samples. Heatmaps showing clusters of different samples based on their species composition using hierarchical clustering, and cluster network illustrated in Cytoscape based on (a) ITS2 and (b) *trnL* sequencing results. Each edge represents sample pair with distance less than 1.5 (as defined by Euclidean distance).

and ITS2 sequence for which the corresponding possible species was evidenced by 3 or less reads.

The rarefaction analysis was performed (with parameter setting to: abundance value = number of sequences, species number step = 1) and rarefaction curves were generated by R using the vegan package (http://CRAN.R-project.org/package=vegan, R package version 2.0-10 (2013)) to demonstrate the sequencing saturation for detecting all components in TCM samples. PCA analysis was performed in R[35] using the ade4 package[36] to cluster different TCM samples and visualize the difference of biological ingredient structures among different samples.

We have performed clustering analysis for the 30 samples based on their species composition profiles. Firstly, we transformed the existence information of different species into Boolean values (0 and 1). Secondly, differences between LDW samples were evaluated by the Euclidean distance with the Boolean values of the species exist in the samples. After that, we applied average-linkage hierarchical clustering algorithm using the "hclust()" function of R[34] to cluster these samples based on their Euclidean distances. A heat-map figure was generated based on clustering results by the "gplots" package of R (Figure 8) (http://CRAN.R-project.org/package=gplots, R package version 2.12.1 (2013)). Finally, we depicted clustered groups of LDW samples based on a graph-theoretic approach as follows: (i) Samples were represented as nodes in a network; edges between nodes indicated their distances that were <=1.5 (as defined by Euclidean distance); (ii) connected components could be extracted from the network; by definition, a connected components is a subgraph in which every node can be reached from any other node. The connected component (subgraph) was defined as a cluster group. Network and cluster visualization was generated by the software package Cytoscape[37].

The 454 sequencing data for 30 LDW samples were deposited to NCBI SRA database with accession number SRR1049940.

1. Yi, Y. D. & Chang, I. M. An overview of traditional Chinese herbal formulae and a proposal of a new code system for expressing the formula titles. *J. Evidence-Based Complementary Altern. Med.* **1**, 125–132 (2004).

2. Yu, Z. & Luo, Y. Import and export analysis of traditional Chinese medicine in 2012. *Mod. Chin. Med.* **15**, 143–146 (2013).
3. Jiang, Y., David, B., Tu, P. & Barbin, Y. Recent analytical approaches in quality control of traditional Chinese medicines–a review. *Anal. Chim. Acta.* **657**, 9–18 (2010).
4. Liang, X. *et al.* Qualitative and quantitative analysis in quality control of traditional Chinese medicines. *J. Chromatogr. A* **1216**, 2033–2044 (2009).
5. Li, S., Zhao, J. & Yang, B. Strategies for quality control of Chinese medicines. *J. Pharm. Biomed. Anal.* **55**, 802–809 (2011).
6. Jing, J., Parekh, H. S., Wei, M., Ren, W. & Chen, S. Advances in analytical technologies to evaluate the quality of traditional Chinese medicines. *TrAC Trends Anal. Chem.* **44**, 39–45 (2013).
7. Ram, M., Abdin, M. Z., Khan, M. A. & Jha, P. in *High-Performance Thin-Layer Chromatography (HPTLC)* (ed Srivastava, M. M.) Ch. 7, 105–116 (Springer Berlin Heidelberg, 2011).
8. Chen, S. *et al.* High-performance thin-layer chromatographic fingerprints of isoflavonoids for distinguishing between *Radix Puerariae lobate* and *Radix Puerariae thomsonii*. *J. Chromatogr. A* **1121**, 114–119 (2006).
9. 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. Pharmacal. Res.* **33**, 1355–1359 (2010).
10. Li, Z., Merfort, I. & Reich, E. High-performance thin layer chromatography for quality control of multicomponent herbal drugs: example of Cangzhu Xianglian San. *J. AOAC Int.* **93**, 1390–1398 (2010).
11. Sucher, N. J. & Carles, M. C. Genome-based approaches to the authentication of medicinal plants. *Planta Med.* **74**, 603–623 (2008).
12. Heubl, G. New aspects of DNA-based authentication of Chinese medicinal plants by molecular biological techniques. *Planta Med.* **76**, 1963–1974 (2010).
13. Kool, A. *et al.* Molecular identification of commercialized medicinal plants in southern Morocco. *PLoS One* **7**, e39459 (2012).
14. Yan, D. *et al.* Forensic DNA barcoding and bio-response studies of animal horn products used in traditional medicine. *PLoS One* **8**, e55854 (2013).
15. Gao, T. *et al.* Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large Asteraceae family. *BMC Evol. Biol.* **10**, 324 (2010).
16. Gao, T. *et al.* Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. *J. Ethnopharmacol.* **130**, 116–121 (2010).
17. Pang, X. *et al.* Applying plant DNA barcodes for Rosaceae species identification. *Cladistics* **27**, 165–170 (2011).
18. Coghlan, M. L. *et al.* Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. *PLoS Genet.* **8**, e1002657 (2012).
19. Newmaster, S. G., Grguric, M., Shanmughanandhan, D., Ramalingam, S. & Ragupathy, S. DNA barcoding detects contamination and substitution in North American herbal products. *BMC Med.* **11**, 222 (2013).
20. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6**, 805–814 (2005).
21. Chen, S. *et al.* Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* **5**, e8613 (2010).
22. Li, D. *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.* **108**, 19641–19646 (2011).
23. Yao, H. *et al.* Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One* **5**, e13102 (2010).
24. Keller, A. *et al.* 5.8S–28S rRNA interaction and HMM-based ITS2 annotation. *Gene* **430**, 50–57 (2009).
25. Ward, J., Peakall, R., Gilmore, S. R. & Robertson, J. Molecular identification system for grasses: a novel technology for forensic botany. *Forensic Sci. Int.* **152**, 121–131 (2005).
26. Taberlet, P. *et al.* Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* **35**, e14 (2007).
27. Wang, J. *et al.* Chinese patent medicine liu wei di huang wan combined with antihypertensive drugs, a new integrative medicine therapy, for the treatment of essential hypertension: a systematic review of randomized controlled trials. *J. Evidence-Based Complementary Altern. Med.* **2012** (2012).
28. Chinese Pharmacopoeia Commission. *Chinese Pharmacopoeia*. ver. 2010 (Chinese Medical Science and Technology Press, 2010).
29. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
30. McNeal, J. R. *et al.* Using partial genomic fosmid libraries for sequencing complete organellar genomes. *Biotechniques* **41**, 69–73 (2006).
31. Yagi, S. M., El Tigani, S. & Adam, S. E. I. Toxicity of *Senna obtusifolia* fresh and fermented leaves (kawal), *Senna alata* leaves and some products from *Senna alata* on rats. *Phytother. Res.* **12**, 324–330 (1998).
32. Wang, X. *et al.* Potential role of metabolomics apporoaches in the area of traditional Chinese medicine: As pillars of the bridge between Chinese and Western medicine. *J. Pharm. Biomed. Anal.* **55**, 859–868 (2011).
33. Cheng, X. *et al.* DNA extraction protocol for biological ingredient analysis of LiuWei DiHuang Wan. *Genomics, Proteomics & Bioinformatics,* DOI: 10.1016/j.gpb. 2014.03.002 (2014).
34. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**, e27310 (2011).
35. Dessau, R. B. & Pipper, C. B. "R"--project for statistical computing. *Ugeskr Laeger* **170**, 328–330 (2008).
36. Dray, S. & Dufour, A. B. The ade4 package: Implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).
37. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

Acknowledgments

This work is partially supported by the grant of National Natural Science Foundation of China (30870572, 61303161 and 61103167) and National High-tech R&D Program (863 Program) funded by Ministry of Science and Technology of China (2012AA02A707 and 2014AA021502). We also thank NVIDIA for their support through CUDA Research Center at QIBEBT-CAS, as well as the support from the Key Laboratory for Rare and Uncommon Diseases of Shandong Province, P.R. China.

Author contributions

H.B. and K.N. conceived of and proposed the idea. X.W.C., X.H.C., H.B. and K.N. designed the study. X.W.C. and C.P.B. performed the experiments. X.W.C., X.Q.S. and K.N. analyzed the sequencing data. X.W.C., X.Q.S., X.H.C., H.X.Z., J.X., H.B. and K.N. contributed to writing, revising and proof-reading the manuscript. All authors read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Cheng, X.W. *et al.* Biological ingredient analysis of traditional Chinese medicine preparation based on high-throughput sequencing: the story for Liuwei Dihuang Wan. *Sci. Rep.* **4**, 5147; DOI:10.1038/srep05147 (2014).