Chapter 3

Designing of new metabarcode and evaluation of their efficacy through NGS platform

3.1 Introduction

DNA metabarcoding combined the strengths of next-generation sequencing and barcoding for detecting multiple taxa in samples (Coghlan et al., 2012). Using a single universal plant barcode for species-level identification has proven challenging due to the great diversity, relatively slow molecular evolution, frequent cross-pollinations, and hybridization in the plant kingdom; henceforth, different barcodes show different degrees of taxon specificity (Fazekas et al., 2009). To precisely identify plant species within complex herbal formulations, and to detect potential adulteration, the power of parallel sequencing was employed (Arulanandhu et al., 2017; Pandit et al., 2021; Seethapathy et al., 2019; Wu & Shaw, 2022; Yao et al., 2022). However, these studies also highlighted the limitations of DNA metabarcoding applications for the authentication of herbal products due to variability in degrees of universality and resolution power of barcodes for specific taxa, a lack of a curated database, and a robust bioinformatics pipeline (Raclariu et al., 2018; Wu & Shaw, 2022). To overcome these constraints, there is a need for screening of new barcodes and new variable regions within the same barcode for authentication of the herbal products.

This study aims to develop and validate a new *ITS2* metabarcoding method for accurately identifying medicinal plant species in complex mixtures. This involves designing degenerate primers, chosen for their shorter length and high-resolution power, capable of discriminating up to species level. To evaluate primer efficiency, two types of mock controls were designed to ensure accurate and sensitive species detection. The ultimate goal is to provide a reliable tool for authenticating herbal products and ensuring their quality.

3.2Materials and Methods

3.2.1 Collection of Reference Plant Material

Reference plants were collected with assistance from an expert taxonomist associated with Maharaja Sayajirao University (MSU), Vadodara (Gujarat, India) and the Directorate of Medicinal and Aromatic Plants Research (DMAPR), Anand (Gujarat, India).

3.2.1.1 Molecular Authentication of Reference Plant

DNA was extracted from leaf samples of the reference plant leaf using the DNeasy Plant mini kit, Qiagen, India. These reference plant species were definitively identified through Sanger sequencing of the *rbcL* gene, amplified using the universal primers listed in Table 2.1. Subsequently, the obtained sequences were deposited into the NCBI database with accession numbers MW628906 to MW628936. Voucher specimens were developed and submitted to our institutional herbarium.

3.2.2 Primer Designing

The aim was to design metabarcode for the *ITS2* genes that would be approximately 300 base pairs in length, enabling amplification from degraded DNA isolated from herbal products and allowing discrimination and identification of species at lower taxa levels. The metabarcode for the *ITS2* gene were designed by downloading and curating *ITS2* sequences of Magnoliophyta from the BOLD database. In the end, 1,465 *ITS2* sequences were retained to design the metabarcode after trimming and curation of sequences in the CLC workbench. These sequences were preceded for multiple sequence alignment separately using BioEdit 7.2. HYDEN (HighlY DEgeNerate primers) software (Linhart & Shamir, 2005) to design degenerate primers, where a maximum of 3 degeneracy per primer was allowed. The designed primers were checked for amplicon length using NCBI primer BLAST (Altschul et al., 1990) (Figure 3.1). To synthesize fusion primers, forward primers of *ITS2* were tagged with the Ion torrent adapter and a ten bp multiplex identifier barcode. In contrast, reverse primers were tagged with the P1 adapter. The nucleotide sequence of the designed primers and their amplicon length are shown in Table 2.1.

3.2.3 Preparation of Different Mock Controls

Two different types of controls were prepared as follows: Control 1) Genomic DNA from plant leaves from different genera belonging to diverse families has been first isolated and pooled into three different groups. A compromise of three different groups was prepared for the first type of control with the plant species of a different genus. Group 1 (5P) contained DNA from five species in equal proportions. Group 2 (10P) included DNA from ten species, with TB as a target. Group 3 (15P) comprised DNA from fifteen species, targeting TB, TC and PE (Figure 3.2; 3.9). High-quality DNA of all the species have been isolated individually from leaf tissue and pooled together in equal proportion to make these groups. The group's diversity has been increased by adding species from diverse genera that belong to diverse families to evaluate the resolution power and universality of the primers for the maximum number of species. Control 2) genomic DNA (Isolated from plant leaves) pool from different species of the two genera (Figure 3.2). As mentioned above, for the second type of control, two groups were prepared. Group one comprises six plant species of the two genera, including *Asparagus* and *Terminalia* (Figure 3.9b). The second group includes seven plant species of the two genera, including *Piper* and *Phyllanthus* (Figure 3.9b). Similar to the first control, high-quality DNA was individually isolated from each species and pooled in equal amounts. These controls were utilized to obtain insight into the resolving strength of our newly designed *ITS2* metabarcode at lower taxa levels.

3.2.4 PCR Optimization and Library Preparation

The library preparation process became a single-step process with barcoded fusion primers. The PCR optimization with each barcoded fusion primer was done with 45 different plant DNA listed in Table 3.1. PCR amplification of *ITS2* was performed using a temperature gradient to optimize primer annealing. The 20 µL reaction mixture contained 10 µL Emerald Master mix (2X), 2 µL genomic DNA (10-15 ng/µL), 1 µL each of forward and reverse primers (5 pmol), 1 µL BSA (2 mg/mL), and 5 µL PCR-grade water (Table 3.1a). Thermal cycling conditions were: initial denaturation at 95°C for 5 minutes; 30 cycles of 95°C for 5 minutes, 50-60°C (2°C intervals) for 30 seconds, and 72°C for 1 minute; and a final extension at 72°C for 5 minutes (Table 3.1b).

3.2.5 Metabarcoding

DNA from plant materials, blended formulations, and herbal products were extracted in duplicate using the DNeasy Plant Mini Kit (QIAGEN, Germany) following the manufacturer's instructions. The library was prepared from each DNA sample using *ITS2* fusion primers with optimized PCR conditions. The libraries were purified using AMPure XP beads (Beckman Coulter, CA, USA), and the quality of some of the libraries was checked using an Agilent high-sensitivity DNA kit on an Agilent 2100 Bioanalyzer. For each sample, libraries from two replicates were pooled. Further, all libraries were diluted to 100 pM and pooled in equimolar concentration. Emulsion PCR was carried out using Ion 520™ & Ion 530™ Kit-OT2 with 400 bp chemistry (Thermo Fisher Scientific, MA, USA). Sequencing was performed on the Ion S5 system using a 520/530 chip (Thermo Fisher Scientific, MA, USA).

3.2.6 Metabarcoding data analysis

Optimization was done with three parameters for establishing the metabarcoding data analysis pipeline. 1) filtering criteria includes discarding reads with length <280 and >300 bp for *ITS2* discarding reads 2) OTU clustering with 97, 98 and 99% similarity in reads, and 3) discarding OTU clusters having <5 and <10 reads. Obtained reads were filtered based on the quality score (Q >=20) and read length using PRINSEQ (Schmieder & Edwards 2011). Clustering filtered reads were performed using CD-HIT-EST (Huang et al., 2010). After that, the taxonomic assignment of OTU clusters having ≥5 and ≥10 reads were done using BLASTn (Altschul et al., 1990) (NCBI) with minimum E value $10E^{-5}$. For each sequence, ten hits were retrieved, and each hit was inspected and evaluated manually for the assigned plant genus and species. To analyze read abundance of each plant species, the number of reads was normalized by considering the total number of reads obtained after discarding clusters with <5 reads and <10 reads as 100% (Figure 3.3).

3.3 Results and Discussion

3.3.1 PCR assays using Newly Designed *ITS2* Primers and Fusion Primers

The *ITS2* region has proven valuable for detecting adulteration in herbal products, both in single and multi-barcode approaches (Frigerio et al., 2021; Seethapathy et al., 2019; Urumarudappa et al., 2020; Yao et al., 2022). However, DNA degradation during herbal product processing can hinder PCR amplification, particularly with longer barcodes (Mishra et al., 2016; Parveen et al., 2016; Raclariu et al., 2018; Wu & Shaw, 2022). To address this, the *ITS2* metabarcode length was minimized to approximately 300 bp, aiming to improve amplification efficiency from potentially degraded DNA samples, particularly targeting TB, TC and PE species. Minimal criteria, such as routine applicability and minimum intraspecific but maximum interspecific divergence at the taxon level, must be followed in the search for the appropriate barcode region. Hence, degenerated *ITS2* metabarcode primers were designed for high amplification efficiency, universality, and resolution power. In total, 45 medicinal plants from diverse families, genera, and species were taken to confirm and optimize the newly designed *ITS2* primer sets for PCR amplification experimentally (Table 3.2). The annealing temperature was optimized, and the results showed that the *ITS2* primer sets performed optimum at 56 °C (Figure 3.4). Among newly designed *ITS2* metabarcode, gives 88.9% amplification efficiency and is not able to provide amplification in 5 plant species including *Ailanthus excelsa, Andrographis paniculata, Adhatoda vasica, Ocimum tenuiflorum, and Ocimum canum* (Table 3.2, Figure 3.5). This lack of amplification could be attributed to the reduced length of the metabarcode and primer fit compatibility with these particular species (Särkinen et al., 2012; Techen et al., 2014). The amplification effectiveness of " fusion primers" (tagged with Ion torrent adapter and barcodes) of *ITS2* metabarcode remained unchanged. However, the appearance of non-specific amplification in some barcodes suggests that 56 °C annealing temperature is not optimal for the fusion primers. Therefore, further optimization of annealing temperature revealed that non-specific amplification was overcome by increasing the annealing temperature to 60 °C (Figure 3.6).

3.3.2 Establishing Data Analysis Pipeline using Mock Controls

A gDNA pooled control comprising different genera was used to establish the *ITS2* metabarcoding data analysis pipeline. A 300 bp read length filter was chosen, yielding more curated sequences than a 280 bp filter, likely due to the variable nature of the *ITS2* region and increased stringency afforded by the higher threshold (Table 3.2). For OTU clustering, both 97% and 98% similarity performed equally well in resolving plant species (Figure 3.7); 98% was selected for subsequent analyses. Discarding OTU clusters with <5 reads detected more plant species than discarding those with <10 reads (Figure 3.8). Therefore, the finalized *ITS2* metabarcoding analysis parameters were: >300 bp reads, 98% OTU clustering, and discarding OTUs with <5 reads. This optimized pipeline was then applied to the second mock control.

3.3.3 Metabarcoding of Different Types of Mock Controls

In the first type of mock control, the 5P, 10P, and 15P controls were designed to assess the specificity and potential for cross-amplification for *ITS2* metabarcode. The 5P control examines the ability of primer to discriminate against a broad range of plant species by incorporating DNA from five families unrelated to the target species. The 10P control includes one target species alongside nine others, evaluating the primer dynamics. Finally, the 15P control tests the capacity of the primer for simultaneous amplification of all three target species, a crucial factor in identifying adulterants in complex herbal mixtures. Table 3.2 presents the total reads obtained after filtering and the percentage of reads analyzed (post-filtering) after discarding OTU clusters with <5 reads for *ITS2* in the gDNA pooled controls (5P, 10P, and 15P). A total of 452,380 *ITS2* reads were obtained. The newly designed *ITS2* metabarcode exhibited varying detection success across the different control groups. *Eclipta prostrata* showed the highest percentage of reads in both the 5P (97.02%) and 10P (79.5%) controls. In contrast, one of the targeted species PE dominated the 15P control with the highest percentage of reads (40.7%). However, Aegle *marmelos* and *Azadirachta indica* were undetectable in the 5P, 10P, and 15P groups. This suggests potential primer incompatibility or significant DNA sequence divergence in the *ITS2* region for these species. Interestingly, the target species exhibited varying detection rates depending on the control group. TB showed 8.5% of reads in the 10P control but only 1.77% in the 15P control. Similarly,

TC yielded a mere 0.3% of reads in the 15P control. This variation might be attributed to PCR bias stemming from differences in DNA quality and quantity among the samples. High polysaccharide and polyphenol content in certain species could interfere with PCR amplification, potentially leading to inaccurate results. Overall, the *ITS2* metabarcode successfully detected three out of five species in the 5P control, seven out of ten in the 10P control, and twelve out of fifteen in the 15P control (Table 3.4; Figure 3.9). These findings highlight the influence of species composition and DNA quality on metabarcoding success and underscore the need for further optimization.

For the second type of control, gDNA pooled from different species of two genera, 164358 reads were obtained for the *ITS2* metabarcode (Table 3.3). The *ITS2* metabarcode demonstrated its ability to differentiate closely related species within these complex mixtures. In the 6SP control, it successfully identified five out of six species, including *T. bellirica* (17.3%), *T. chebula* (0.1%), and *T. arjuna* (6.8%), highlighting its ability to distinguish between members of the same genus. Similarly, in the 7SP control, *ITS2* resolved four out of seven species, including *P. nigrum* (0.8%), *P. longum* (21%), *P. emblica* (62.5%), and *P. amarus* (13.5%) (Table 3.5; Figure 3.10). The results indicate that the *ITS2* metabarcode can successfully distinguish between species, even when they belong to the same genus. However, variation in read percentages between congeneric species influences species abundance estimations. This reinforces previous reports highlighting the strong species-level discrimination power of *ITS2* metabarcoding, aligning with the primary goal of species detection and differentiation within genera (Newmaster et al., 2013; Yao et al., 2022).

3.4 Conclusion

This study employed two in-house designed mock controls with equal DNA proportions from various plant species to rigorously evaluate the *ITS2* metabarcoding approach. By analyzing read abundance, detection efficiency, primer compatibility, and PCR dynamics, the researchers sought to comprehensively evaluate the performance and identify potential limitations of *ITS2* metabarcoding. Despite inherent challenges in metabarcoding complex mixtures, the *ITS2* approach demonstrated high fidelity in resolving plant species. This success in controlled experiments highlights its robustness

and reliability for species-level identification, particularly in authenticating herbal products and ensuring their quality, safety, and efficacy. However, the development of standardized methods for metabarcoding sequencing and bioinformatics analysis pipeline and curated database is needed for effective use as a regulatory tool to authenticate herbal products in combination with advanced chemical methods to identify bioactive therapeutics.

Table 3.1a PCR reaction set up using *ITS2* metabarcode.

Component	**Volume (µL)**
Emerald Takara Master Mix (2X)	10.0
Template DNA	2.0
Forward Primer (5 pmol)	1.0
Reverse Primer (5 pmol)	1.0
BSA (2 mg/mL)	2.0
PCR-grade Water	4.0

*Forward and reverse *ITS2* metabarcode primer pairs are used as mentioned in Table 3.1.

Table 3.1b PCR cycling condition for *ITS2* metabarcode**.**

Steps	**Temperature (°C)**	**Duration**	**Number of Cycles**
Initial Denaturation	95	1 minutes	1
Denaturation	95	30 seconds	30
Annealing	60		
Extension	72		
Final Extension	72	5 minutes	1

Table 3.2 PCR amplification of 45 medicinal plant species with new *ITS2* metabarcode.

Sr. No.	Plant Species	*ITS2*	Sr. No.	Plant Species	*ITS2*
1	*Asparagus adscendens*	+	24	*Piper longum*	+
2	*Eclipta alba*	+	25	*Phyllanthus madresapatensis*	+
3	*Senna alexandrina*	+	26	*Aegle marmelos*	+
4	*Phyllanthus amarus*	+	27	*Bacopa monnieri*	+
5	*Cassia angustifolia*	+	28	*Ocimum canum*	-
6	*Piper arabarum*	+	29	*Piper nigrum*	+
7	*Terminalia arjuna*	+	30	*Solanum nigrum*	+
8	*Mentha arvensis*	+	31	*Ocimum tenuiflorum*	-
9	*Ailanthus excelsa*	-	32	*Zingiber officinale*	+
10	*Andrographis paniculata*	-	33	*Plantago ovata*	+
11	*Melia azedarach*	+	34	*Carica papaya*	+
12	*Ocimum basilicum*	+	35	*Asparagus racemosus*	+
13	*Ipomoea batatas*	+	36	*Withania somnifera*	+
14	*Terminalia bellerica*	+	37	*Lagerstroemia speciosa*	+
15	*Justicia adhatoda*	-	38	*Tribulus terrestris*	+
16	*Lantana camara*	+	39	*Cassia tora*	+
17	*Terminalia chebula*	+	40	*Aloe vera*	+
18	*Asparagua dumosus*	+	41	*Centella asiatica*	+
19	*Phyllanthus emblica*	+	42	*Hygrophila auriculata*	+
20	*Ocimum gratissimum*	+	43	*Piper betle*	+
21	*Gossypium herbaceum*	+	44	*Vitex negundo*	+
22	*Azadirachta indica*	+	45	*Senna occidentalis*	+
23	*Feronia limonia*	+			

Table 3.3. Details of reads obtained through *ITS2* metabarcoding for mock controls.

Mock control description	**Sample ID**	**Reads before filtering**	**Cumulative reads per control**	**Reads after filtering**		**Percentage of analyzed reads from filtered reads[#]**
				Discarding reads (>280 bp)	**Discarding reads (>300 bp)**	
Control 1: Genomic DNA pool of different genus (P)	5P	113942	452380	91564	90900	79.8
	10P	140647		85547	83209	59.2
	15P	197791		141856	134011	67.8
Control 2: Genomic DNA pooled from different species of the genus (SP)	6SP	63249	164358	40611	38013	60.1
	7SP	101109		66248	60328	59.7

[#]Analysed reads (from filtered reads) after discarding OTU clusters having <5 reads.

Table 3.4 Detection of medicinal plant species using *ITS2* metabarcoding in type 1 mock controls.

Sr. no.	Plant Species	Family	Detected and not detected plants in metabarcoding		
			5P	10P	15P
1	*Andrographis paniculata*	Acanthaceae	ND	ND	ND
2	*Azadirachcta indica*	Meliaceae	ND	ND	ND
3	*Eclipta alba*	Asteraceae	D	D	D
4	*Piper nigrum*	Piperaceae	D	D	D
5	*Zingiber officinale*	Zingiberaceae	D	D	D
6	*Aegale marmelous*	Rutaceae		ND	D
7	*Centella asiatica*	Apiaceae		D	D
8	*Terminalia arjuna*	Combretaceae		D	D
9	*Terminalia bellerica*	Combretaceae		D	D
10	*Vitex negundo*	Lamiaceae		D	D
11	*Bacopa monneri*	Plantaginaceae			D
12	*Sena tora*	Fabaceae			D
13	*Phyllanthus embilica*	Phyllanthaceae			D
14	*Terminalia chebula*	Combretaceae			D
15	*Justicia adhatoda*	Acanthaceae			ND

D = detected; ND = not detected

Table 3.5 Detection of medicinal plant Species using *ITS2* metabarcoding in in type 2 mock controls.

Sr.no	Plant species	Family	Mock control ID	*ITS2* metabarcode
1	*Asparagus racemosus*	Asparagaceae	6SP	D
2	*Asparagua dumosus*			ND
3	*Asparagus adscendens*			D
4	*Terminalia arjuna*	Combretaceae		D
5	*Terminalia bellerica*			D
6	*Terminalia chebula*			D
1	*Pipper nigrum*	Piperaceae	7SP	D
2	*Pipper logume*			D
3	*Piper betle*			ND
4	*Pipper arabarum*			ND
5	*Phyllanthus embica*	Phyllanthaceae		D
6	*Phyllanthus amaras*			D
7	*Phyllanthus madresapatensis*			ND

D = detected; ND = not detected

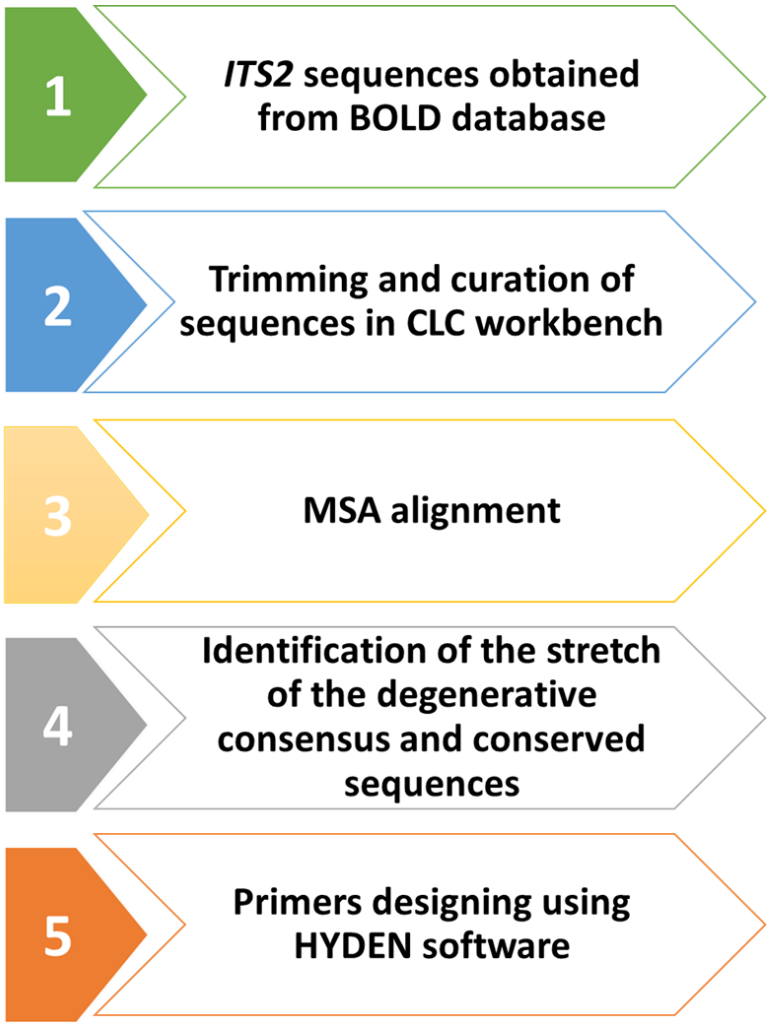

Figure 3.1 Bioinformatics pipeline for designing of *ITS2* Metabarcode primer.

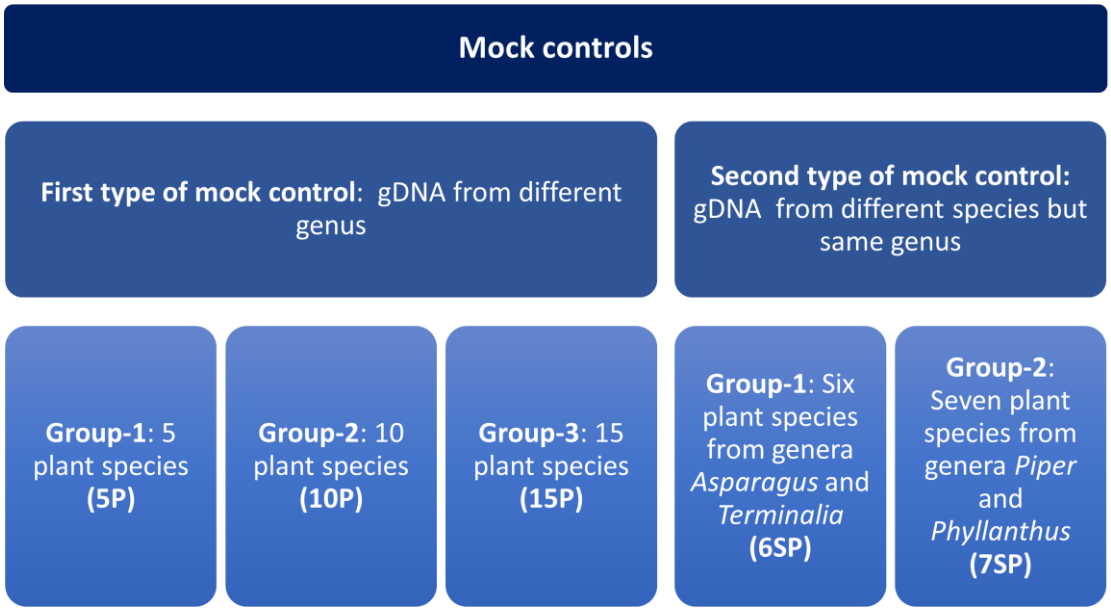

Figure 3.2 Schematic representation of different types of mock controls prepared in this study.

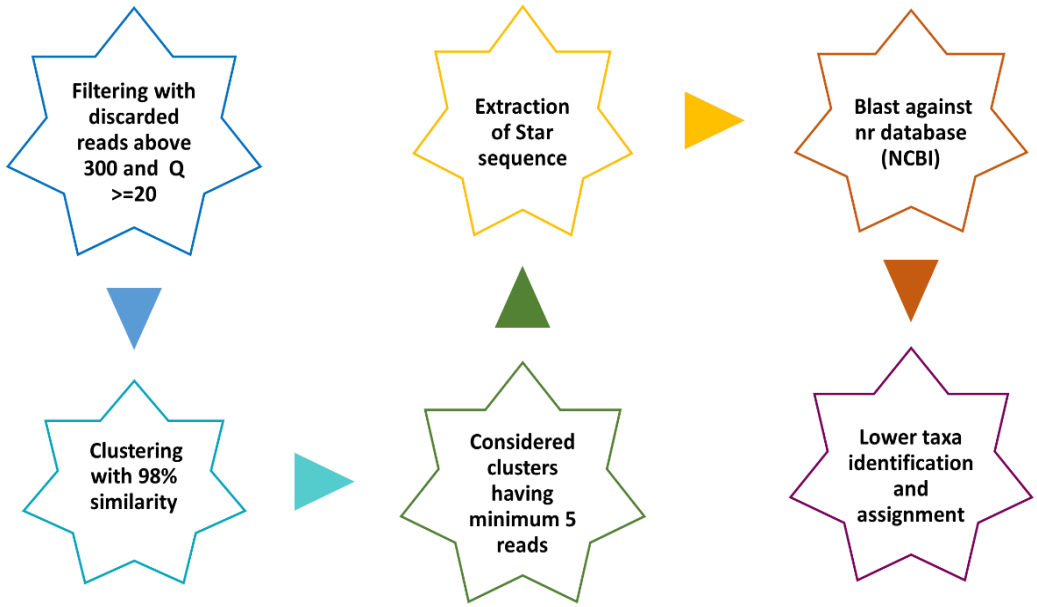

Figure 3.3 Optimized bioinformatics pipeline for *ITS2* metabarcoding data analysis.

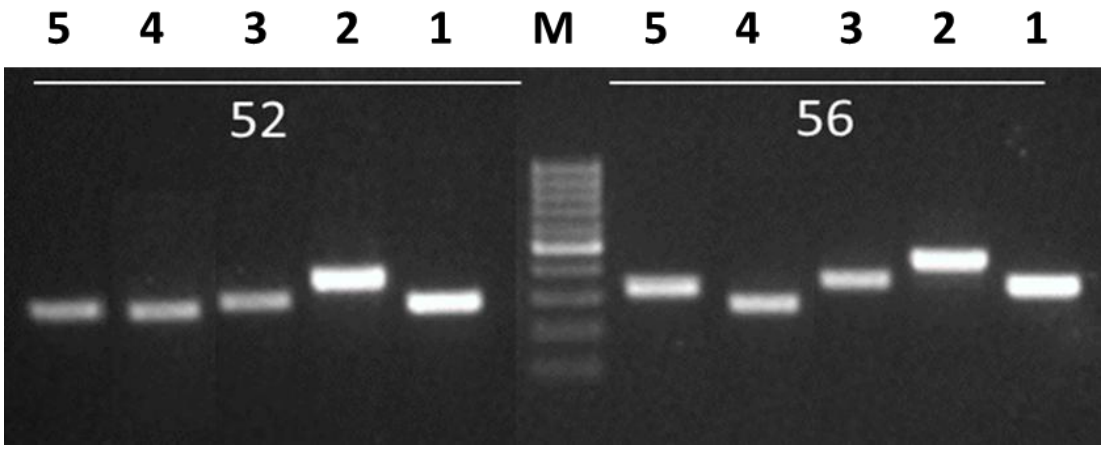

Figure 3.4 Optimization of *ITS2* metabarcode using DNA from leaves of 5 plant species at 52 and 56°C annealing temperature. M = 100 bp DNA marker (100-1000 bp). Expected amplicon size range: 290-320 bp. Sample IDs 1-5 are listed in Table 3.2.

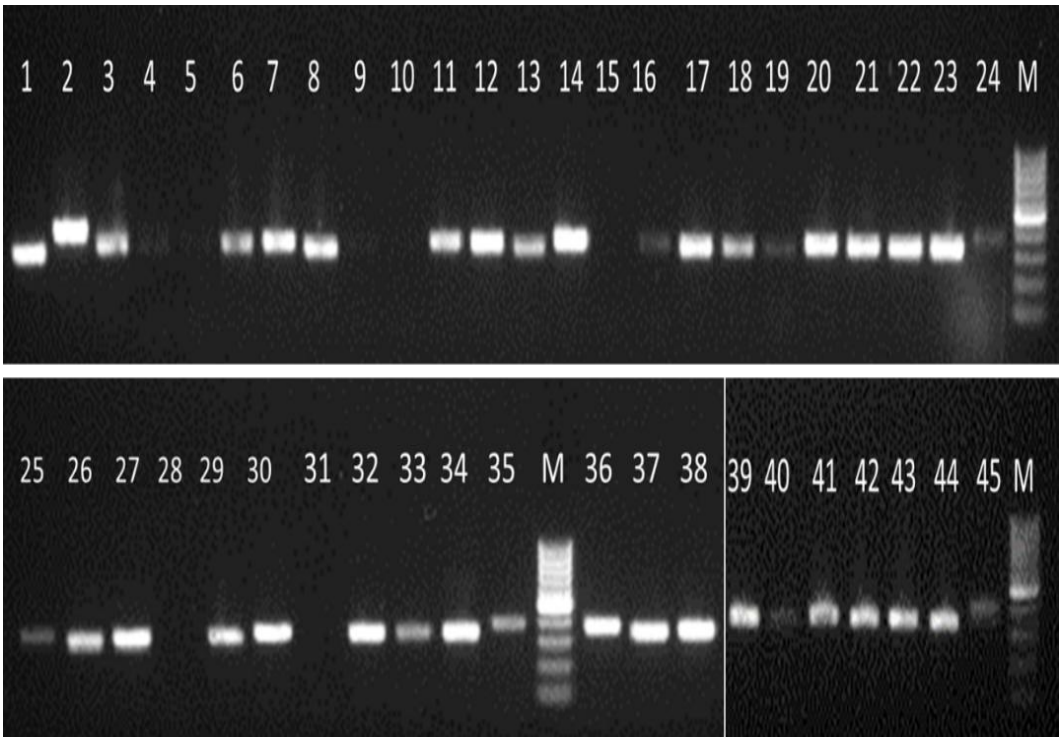

Figure 3.5 *ITS2* metabarcode amplification using DNA from leaves of 45 plant species belonging to different families and genera. M = 100 bp. DNA marker (100-1000 bp). Expected amplicon size range: 290-320 bp. Plant species 1-45 listed in Table 3.2.

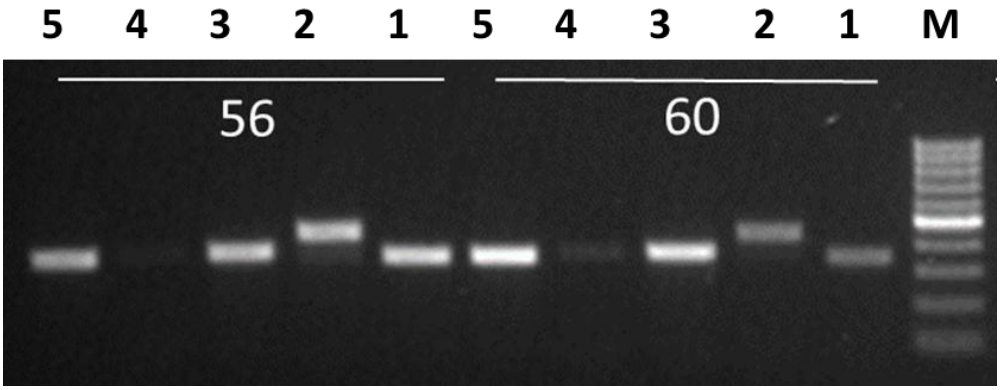

Figure 3.6 *ITS2* metabarcode fusion primers optimization assay using DNA from leaves of 5 different plant species with at 52 and 56 ℃ annealing temperature. M = 100 bp DNA marker (100-1000 bp). Expected amplicon size range: 360-380 bp. Sample IDs 1-5 Plant species listed in Table 3.2.

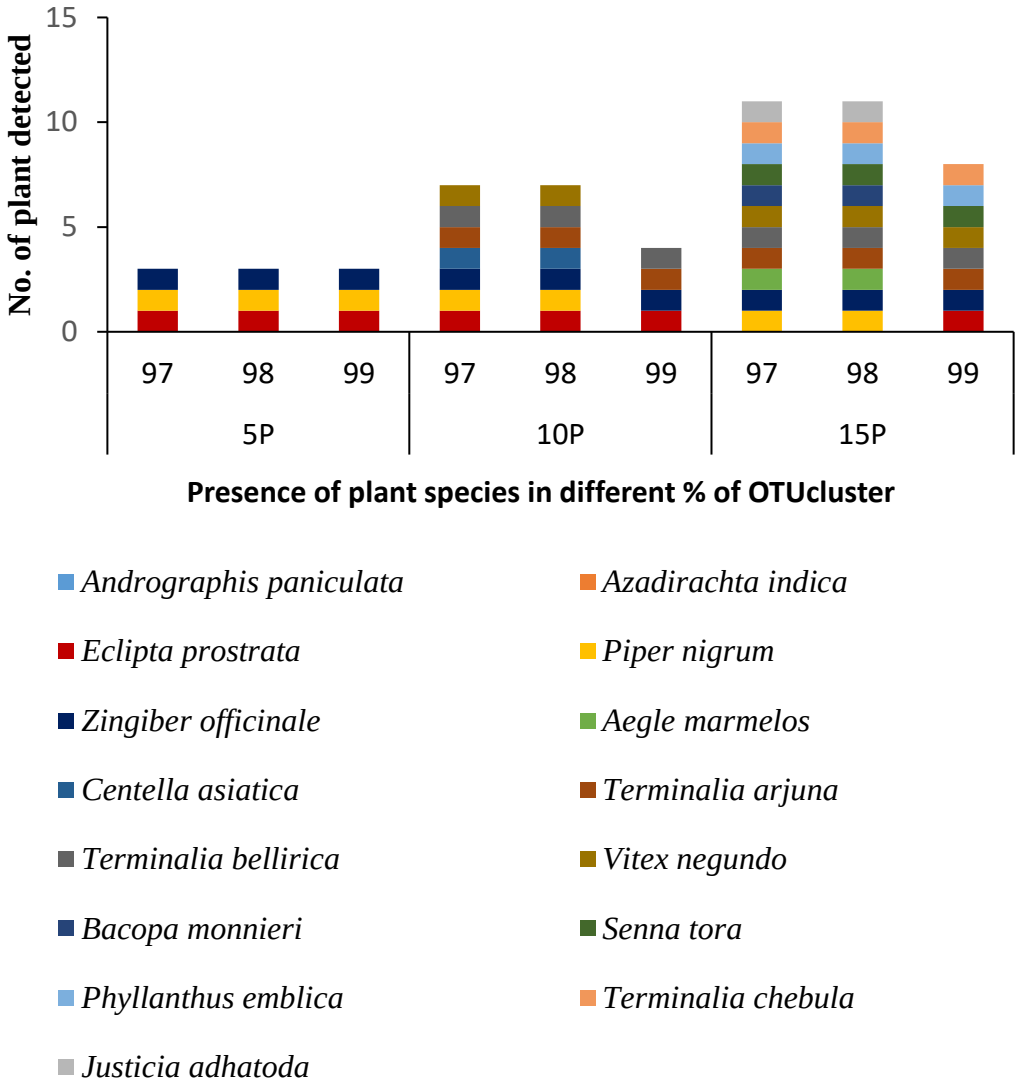

Figure 3.7 Detection of medicinal plant species using *ITS2* metabarcoding in type 1 mock control with 97%, 98%, and 99% OTU clustering to optimize the bioinformatics pipeline.

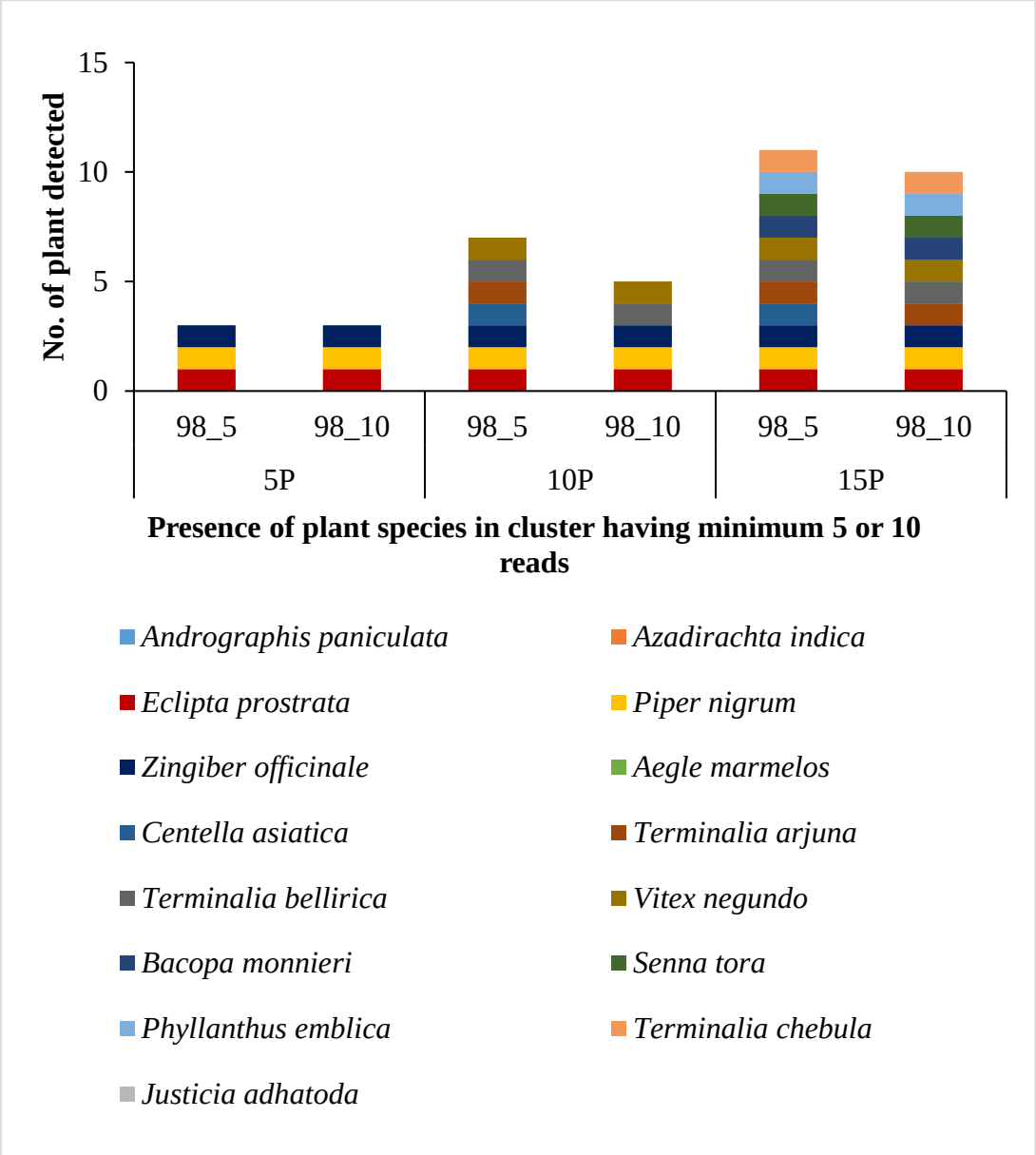

Figure 3.8 Detection of medicinal plant species using *ITS2* metabarcoding in type 1 mock, after discarding clusters with fewer than 5 or 10 reads, to optimize the bioinformatics pipeline.

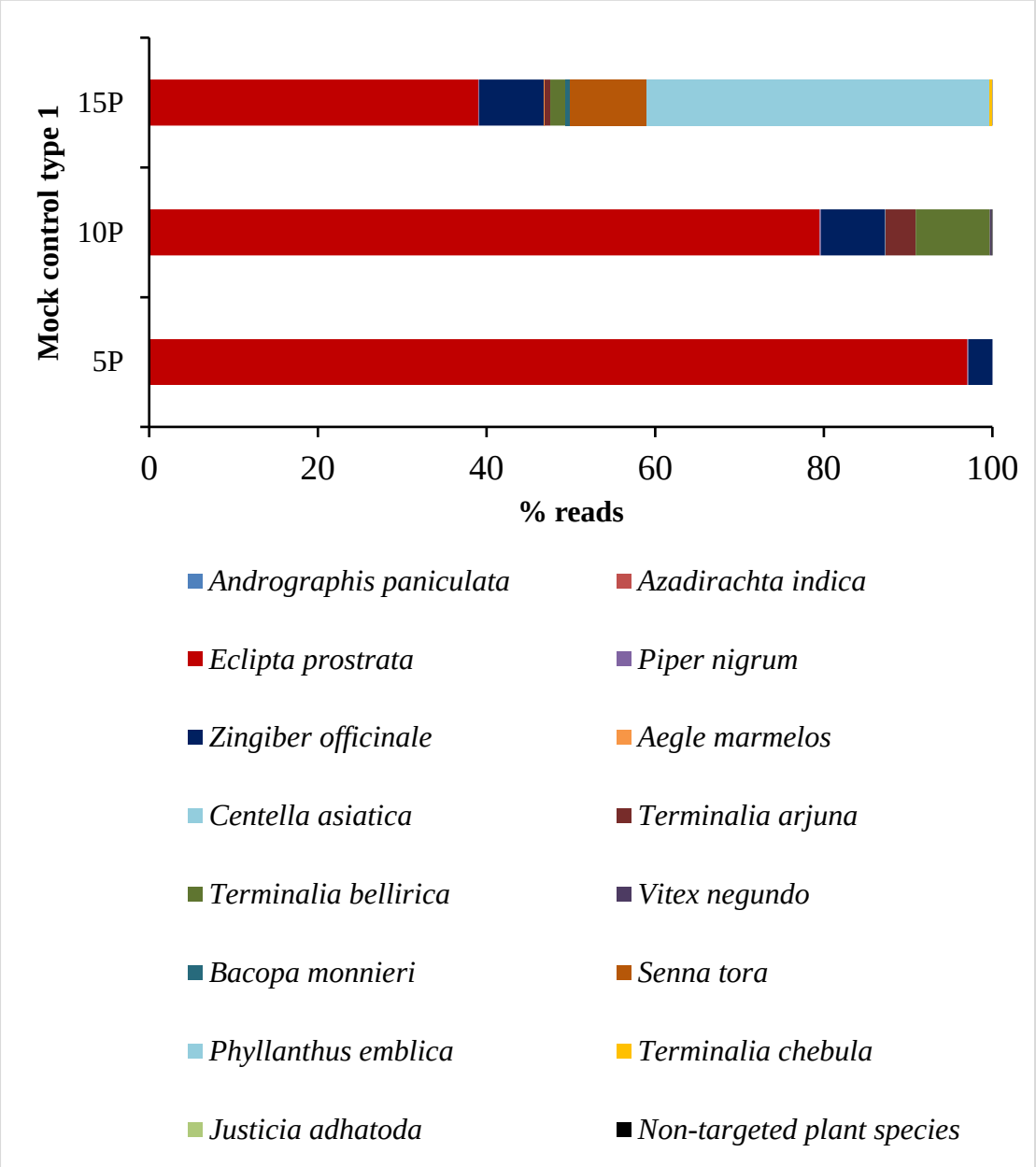

Figure 3.9 Relative abundance (% reads) of medicinal plant species detected in the first type of mock controls through *ITS2* metabarcoding. Plant species details are provided in Table 3.4

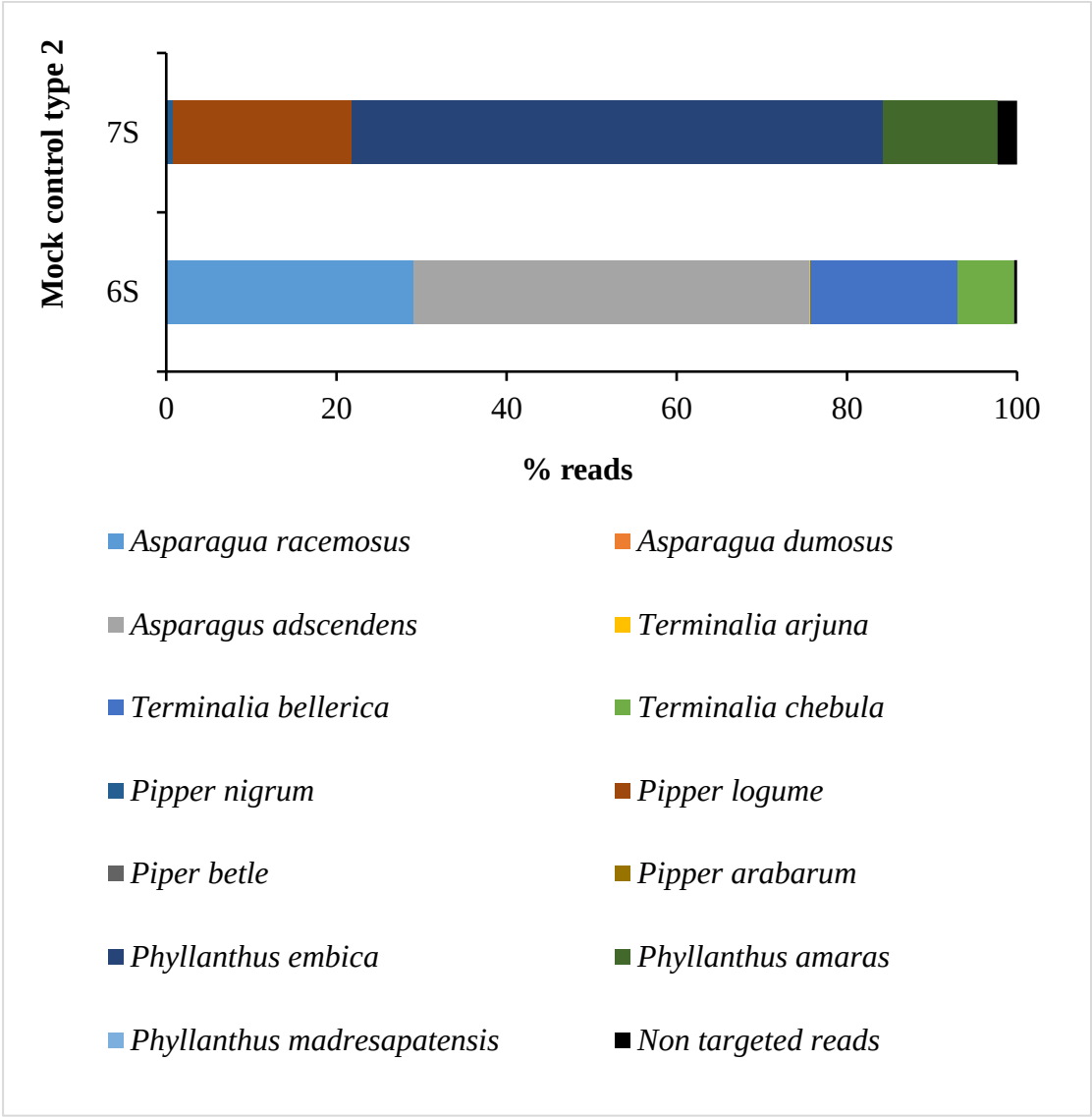

Figure 3.10 Relative abundance (% reads) of medicinal plant species detected in the second type of mock controls through *ITS2* metabarcoding. Plant species details are provided in Table 3.5.