Food Chemistry: Molecular Sciences 3 (2021) 100044

Contents lists available at ScienceDirect

Food Chemistry: Molecular Sciences

journal homepage: www.sciencedirect.com/journal/food-chemistry-molecular-sciences

Evaluating the usefulness of next-generation sequencing for herb authentication

Anna Delgado-Tejedor [1,2], Pimlapas Leekitcharoenphon, Frank M. Aarestrup [3], Saria Otani [*,3]

Research Group for Genomic Epidemiology, Division for Global Surveillance, National Food Institute, Technical University of Denmark, 2800 Kgs Lyngby, Denmark

ARTICLE INFO

Keywords:
Herbs
Authenticity testing
Next generation sequencing
Barcodes
Food

ABSTRACT

Food authentication is a rapidly growing field driven by increasing public awareness of food quality and safety. Foods containing herbs are particularly prone to industrial fraud and adulteration. Several methodologies are currently used to evaluate food authenticity. DNA-based technologies have increased focus, with DNA barcoding the most widely used. DNA barcoding is based on the sequencing and comparison of orthologous DNA regions from all species in a sample, but the approach is limited by its low resolution to distinguish closely-related species. Here we developed a customised database and bioinformatics pipeline (Herbs Authenticity - GitHub) to identify herbal ingredients implemented as a metagenomics approach for plant-derived product authenticity testing. We evaluated the accuracy of the method by using publicly available plant genomes and databases to allow the construction of our customised database barcodes, which were also complemented with entries from publicly available resources (iBOL and ENA). The pipeline performance was then tested with new 47 de novo partly sequenced whole plant genomes or barcodes as query sequences. Our results show that using our mapping algorithm with the customised barcode database correctly identifies the main components of a wide range of plant-derived samples, albeit with variable additional noise across samples depending on the tested samples and barcodes. Our result also show that at the current stage the usefulness of metagenomics is limited by the availability of reference sequences and the needed sequencing depth. However, this method shows promise for evaluating the authenticity of different herbal products provided that the method is further refined to increase the qualitative and quantitative accuracy.

1. Introduction

There has been a significant increase in food fraud and adulteration for economic advantage over the last decade (Medina, Pereira, Silva, Perestrelo, & Câmara, 2019). As a result, food authentication is a rapidly growing field (Mishra et al., 2016). Foods can be misdescribed through substituting or mixing ingredients with cheaper alternatives, or including undeclared ingredients (Primrose, Woolfe, & Rollinson, 2010). Undeclared compounds can also represent a threat to public health; for example, foods adulterated with nut protein can cause anaphylactic reactions in susceptible individuals (Haynes, Jimenez, Pardo, & Helyar, 2019). There is therefore a need to develop accurate analytical methods to verify the type and quantity of ingredients in food products to verify manufacturer claims (Delia, 2019; Pamela, Haughey, & Elliott, 2018).

Plants and herbs are widely used in the food industry. Although often included in relatively small quantities, they are important and often expensive ingredients in many products, making them prone to industrial fraud (Black, Haughey, Chevallier, Christopher, & Elliott, 2016). Four main methods are currently used to evaluate food authenticity: morphology, chromatography/mass spectrometry, immunological assays, and DNA-based methodologies (Drouet et al., 2018).

While morphological identification is a low-cost approach, its accuracy depends on human expertise and its low resolution makes it unsuitable for powdered products or to distinguish closely-related species (Yat-Tung & Shaw, 2018). High-resolution chromatographic techniques such as gas chromatography (GC) and high-performance liquid chromatography (HPLC) coupled to mass spectrometry (MS) are also popular

* Corresponding author.
E-mail address: saot@food.dtu.dk (S. Otani).
[1] Present address 1: Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain.
[2] Present address 2: Universitat Pompeu Fabra, Barcelona, Spain.
[3] Contributed equally.

https://doi.org/10.1016/j.fochms.2021.100044
Received 6 May 2021; Received in revised form 13 September 2021; Accepted 2 October 2021
Available online 20 October 2021

in food authentication (Drouet et al., 2018). Authentication is performed by comparing the chemical fingerprints from standards with those obtained from the herbal product. However, it is expensive and requires specialist equipment and trained analysts for interpretation (Danezis, Tsagkaris, Camin, Brusic, & Georgiou, 2016). Enzyme-linked immunosorbent assays (ELISAs) are the most widely used immunological method in food authentication due to their high sensitivity (Sasikumar, Swetha, Parvathy, & Sheeja, 2016). However, their performance is lower in processed or powdered products, the design and use of specific antibodies can be expensive, and there may be cross-reactivity to proteins from closely-related species (Montowska, Fornal, Piątek, & Krzywdzińska-Bartkowiak, 2019; Walker et al., 2018).

Over the last few years, next-generation sequencing (NGS) has transformed genomics (Sara, McPherson, & Richard McCombie, 2016). While NGS has been applied to plant and herb identification, currently the most widely used method is DNA barcoding (Gerard, 2016), a technology based on PCR amplification followed by sequencing and comparison of orthologous DNA regions from all species in a sample (Böhme, Calo-Mata, Barros-Velázquez, & Ortea, 2019). The main limitation of DNA barcoding is its current gene bias due to the PCR-based approach. For example, DNA barcoding has insufficient resolution to differentiate *Mentha*, *Ocimum*, *Origanum*, *Salvia*, and *Thymus* species in the Lamiaceae family (Drouet et al., 2018).

Choosing appropriate barcoding genes for identification purposes is challenging, as they need to be present in a wide range of plants and herbs but harbour sufficient interspecies variability (Böhme et al., 2019). To overcome this limitation in resolution, multiple barcode methodologies have been developed and have improved product identification performance, instead of single-locus approaches (single barcode), however with bias that varies depending on the sample taxa (Mishra et al., 2016).

Another challenge in DNA barcoding for herb identification is the lack of a complete and accurate reference library (Coissac, Hollingsworth, Lavergne, & Taberlet, 2016; Hollingsworth, Li, van der Bank, & Twyford, 2016; Tnah et al., 2019). Several databases include different plant-derived genes based on their location or their usage. The most relevant to species identification is the International Barcode of Life project (iBOL) (Illuminate Biodiversity - International Barcode of Life. url: http://ibol. org/site/) from the Consortium for the Barcode of Life (CBOL) initiative (CBOL — iBOL. url: http://www. ibol. org/phase1/cbol/), which contains 450,581 entries. iBOL, however, unevenly represents species and genera, some annotations are poor quality, and several sequences are incomplete. No publicly available database only includes the reference barcodes from herbs and species used in Danish and European cuisine. A customised database would eventually increase the accuracy of identification.

Here we developed and evaluated a metagenomic approach for plant-derived sample authenticity, from both single plant species that are used as spices and commercially available herbal products. A customised alignment pipeline and plant-specific barcode database were built and the pipeline and barcode database were validated using publically available plant sequences of known origins. Finally, 47 herbal plant species and products of known composition and commercially available food preparations were sequenced at an approximately 80 million reads per sample to be evaluated and assessed using our novel metabarcoding analysis pipeline.

2. Materials and methods

2.1. Sample collection and DNA extraction

Forty-seven plants and herb products widely used in Danish and European cuisine were included: 33 plant species (**Table S1**) to build the barcode database and 14 herbal products (**Table S2**) for authenticity evaluation. Samples were seeds, powders, or fresh or dried plant tissue obtained both from a farm in Helsinge, Denmark (Fuglebjerggaard) and a Danish supermarket.

Total genomic DNA was extracted from all samples with the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) (Peter M Hollingsworth et al., 2011) following the manufacturer's instructions with the following modifications: 100–200 mg of sample was used as a starting material and mixed with 400 µl lysis buffer. TissueLyser (Qiagen) was used for bead treatment in two cycles of 1 min at 30 Hz. The final step of lysis was the addition of 4 µl RNase A stock solution to the sample followed by incubation for 15 min at 65 °C. After lysing and protein precipitation, the samples were centrifuged for 5 min at 20,000×*g*, and the supernatant was applied to the QIAshredder Mini spin column (Qiagen) and centrifuged for 5 min at 20,000×*g*. Then, the flow-through fraction excluding any precipitate was mixed with 1.5 volumes washing buffers AW1 and AW2. Finally, the DNA was eluted in two volumes of 50 µl of pre-heated (65 °C) AE buffer (**Table S3**).

2.2. Library preparation and sequencing

Genomic DNA quality was assessed with the TapeStation Genomic DNA Assay (Agilent Technologies, Santa Clara, CA). Library preparation was performed using the KAPA HyperPrep kit without PCR as per the manufacturer's recommendations (Kapa Biosystems, Roche, Basel, Switzerland). Library quality and quantity were assessed with the Qubit 2.0 DNA HS Assay (Thermo Fisher Scientific, Waltham, MA) and QuantStudio® 5 (Applied Biosystems, Foster City, CA). Libraries were loaded onto an Illumina HiSeq 2 × 150 bp format to target 80 M total reads (40 M reads each direction) per sample.

2.3. Pipeline and database

2.3.1. Pipeline implementation

The *HerbsAuthenticate* package was constructed. It processes trimmed FASTQ files from both single or paired-end reads to perform sequence mapping against barcode databases. It can also build customised databases based on the user's needs by extracting specific barcodes from the trimmed reads. The complete algorithm, scripts, and documentation are available online at GitHub (Anna Delgado. Herbs Authenticity - GitHub. 2019. url: https://github.com/ADelgadoT/HerbsAuthenticate.git).

2.3.2. Building a customized database: Barcode generation through the alignment of herbal plant sequences to the barcode backbone database

The script Barcodes.py was constructed and used to extract specific barcodes from trimmed reads to build a customised user database. The script uses KMA (Clausen, 2018) and the main workflow is shown in Fig. 1.

To obtain the consensus sequences of a set of specific barcodes (**Table S4**) for plant taxonomical identification, all trimmed reads from sequencing 33 plants (**Table S1**) were aligned to a barcode *backbone* database with KMA (Clausen, 2018). The default barcode backbones present in the database were: *matK* (iBOL accession number ABCBF144-11), *rbcL* (iBOL accession number AGOPO45-11), *ropC1* (GenBank accession number DQ886273.1), *rpoB* (GenBank accession number GU732808.1), *trnH-psbA* (iBOL accession number ALOAF030-10), *trnL-F* (GenBank accession number AF292404.1), *ycf1* (GenBank accession number JF289072.1), *ITS2* (iBOL accession number AGOPO45-11), and *COI* (GenBank accession number AY490250.1) (**Tables S4–S6**). The backbone barcodes were selected based on previously published data for plant identification (Anantha & Johnson, 2019; Dong et al., 2015; Li et al., 2015; Yu, 2018). This database contained the sequences of the nine different genes from several species in the Magnoliophyta phylum (**Tables S4–S6**), and all the included plant species and herbal products belonged to it those taxa.

Before performing alignment, the database was indexed with the default command and parameters, which can be accessed online (genomicepidemiology / kma — Bitbucket. url: https://bitbucket.org/genomicepidemiology/kma). To investigate the optimal set of

Fig. 1. Barcode extraction algorithm workflow.

parameters for barcode extraction, KMA was executed with different configurations (methodology validation section, **Supplementary Information**). The constructed plant barcode backbone was then validated using blastn search against *nt* database. Mapping trimmed reads against a specific barcode database was performed using the Mapping.py script (**Supplementary Information**).

2.4. Pipeline benchmarking procedure using publicly available plant sequence datasets

The performance of all used algorithms was evaluated using ten publicly available single-species datasets of plants downloaded from the European Nucleotide Archive (ENA) (European Nucleotide Archive EMBL-EBI. url: https://www.ebi.ac.uk/ena) (**Table S7**). Trimmed reads from those ten publicly available plant species (**Table S7**) were aligned against our barcode database (**Tables S4–S6**) to obtain all possible barcodes. The alignment was executed at five different level of stringency to evaluate which parameters lead to a better performance (**Supplementary Information**).

To test the performance of the algorithm solely, without the potential effect of our customised barcode database, all ten publicly available plant samples (**Table S7)** were mapped against the iBOL database, a publicly available resource from The International Barcode of Life Consortium (Illuminate Biodiversity - International Barcode of Life. url: http://ibol. org/site/), using KMA with default alignment parameters.

2.5. De novo sequencing and testing

2.5.1. Plant species and herbal product analysis

33 plant species and 14 herbal product sequences included in this study (**Table S1, S2)** were used to identify the barcode sequences for the taxonomical annotation **(Tables S4–S6).** These sequences, once validated with *blastn*, formed the customized databased that was used in this study. In the case that there were missing barcodes from specific species (**Table S12**), those sequences were supplemented with barcodes from public databases such as iBOL and ENA (**Table S8**) when not detected in our samples using the script Barcodes.py (Anna Delgado. Herbs Authenticity - GitHub. 2019. url: https://github.com/ADelgadoT/HerbsAuthenticate.git) with default user options. This was done to increase the detection coverage of a barcode if our sequencing depth was not sufficient to build the backbone barcode database.

3. Results

3.1. Sequencing data

47 herbal plants and products were subjected to next generation sequencing. Before trimming, the maximum number of reads in a sample was 118,252,290 and the minimum value was 80,868,466. On average, there were 97,197,328.69 reads per sample. After trimming, there were 91,881,393 reads per sample on average (94.57%) (**Table S9).** The raw reads from 47 plants were submitted in ENA under project number PRJEB44059.

3.2. Evaluation of the pipeline, HerbsAuthenticate

To test the efficiency of our mapping pipeline for plant detection, ten publicly available samples (**Table S7**) were mapped against our customised barcode database with five different levels of stringency. The fraction of validated barcodes in all five stringency-parameter sets was almost constant, suggesting that the number of validated sequences did not increase proportionally with the reduction in stringency of the alignment algorithm (**Fig. S1 and Supplementary Information**). This suggested that parameter set number three, with a medium level of stringency, had the best performance and was used for the downstream analyses.

To test the efficiency of our mapping pipeline solely, without the effect of our customised barcode database, for plant taxonomical identification, the ten publicly available samples (**Table S7**) were also mapped against the iBOL database. The mapping results are presented in Table 1 **(Figure S3)**.

The most abundant taxon in each sample outputs corresponded to the sample origin (Table 1**, Figure S3**). For example the most abundant hit in garlic (genus *Allium*) (**Figure S3-A**) corresponded to the expected genus, with a relative read abundance of 92%. Although *Asparagales* and *Phoenix* were detected, they were present at largely lower abundances than *Allium*.

3.3. Barcode database construction

The plant samples included in this study (**Table S1**) were mapped to our customised barcode database (**Tables S4–S6**) to obtain their respective barcodes. All the generated barcodes from this mapping are shown in Fig. 2. Only 34% of the total possible sequences were assigned to the backbone database. No barcode was consistently found in all samples. Their recovery rates were: *rbcL* 78.8%, *trnL-F* 51.5%, *trnH-psbA* 42.4%, *ITS2* and *ropC1* 36.4%, *matK* and *rpoB* 24.2%, *ycf1* 15.2%, and *COI* 3%.

To generate a complete database covering as many plant species as possible, the three barcodes with the highest recovery rates (*rbcL*, *trnLF*, and *trnH-psbA*) were selected for further database construction and the missing entries were incorporated manually from publicly available databases (iBOL and ENA; Fig. 3**, Table S10**). The final customised database included 99 sequences (Fig. 3), and contained data from *rbcL*, *trnL-F*, and *trnH-psbA* from all 33 plant samples (**Table S1).**

3.4. Taxonomic composition of 33 plant samples using our customised barcode database and pipeline

Our validated and customised barcode database allowed for good quality plant-taxonomic assignments of the 33 single-species plants included in this study. The compositional taxonomic assignments are all shown as relative abundances, which refer to the proportional abundance of a plant taxon out of the entire identified taxa in one sample. Results are shown in Table 2 (**Figure S4**), and the raw mapped data, before relative abundance calculations, are presented in Table S13.

The most abundant plant taxon that was found in each sample in this collection corresponded to the expected sample taxon (Table 2**, Figure S4**). For example, the basil sample (*Ocimum*) (Table 2**, Figure S4**) had the highest relative read abundance for *Ocimum* (66.0%) followed by *Thymus* (9.3%), *Mentha* (8.2%), *Salvia* (5.9%), and *Anethum* (4.9%). The number of assigned taxa in each sample was between 4 and 6 except for mustard and vanilla were only 3 and 2 taxa were assigned for each sample (Table 2**, Figure S4**).

Table 1
Relative abundances of plant taxa that were identified in each of the ten publicly available plant species when mapped using our pipeline against iBOL database.

Sample name	Basil - *Ocimum*						Brown mustard - *Brassica*					
Genus	*Ocimum*	*Vitex*	*Lycopus*	*Glechoma*	*Lamiales*	*Others*	*Brassica*	*Erucastrum*	*Lepidium*	*Ulva*	*Sinapis*	*Others*
relative abundance	69	10	7	3	2	9	53	11	9	6	5	16
Sample name	Cinnamon - *Cinnamomum*						Fennel - *Daucus*					
Genus	*Cinnamomum*	*Machilus*	*Lindera*	*Laurus*	*Laurales*	*Others*	*Daucus*	*Foeniculum*	*Osmorhiza*	*Anethum*	*Cryptotaenia*	
relative abundance	24	23	9	8	8	28	52	33	6	5	3	
Sample name	Garlic - *Allium*			Liquorice - *Glycyrrhiza*			Thyme - *Thymus*					
Genus	*Allium*	*Asparagales*	*Phoenix*	*Glycyrrhiza*	*Wisteria*	*Medicago*	*Millettia*	*Lotus*	*Others*			
relative abundance	92	5	2	75	10	7	4	3	1			
Sample name	Paprika - *Capsicum*						Vanilla - *Vanilla*					
Genus	*Capsicum*	*Datura*	*Solanum*	*Nicotiana*	*Calibrachoa*	*Others*	*Thymus*	*Origanum*	*Mimulus*	*Mentha*	*Lycopus*	*Others*
relative abundance	40	17	15	10	6	12	50	17	15	4	4	10
Sample name	Tomato - *Solanum*						*Vanilla*	*Cypripedium*	*Phoenix*	*Epipactis*	*Nannochloris*	*Others*
Genus	*Solanum*	*Nicotiana*	*Calibrachoa*	*Datura*	*Others*		53	16	15	10	2	4
relative abundance	69	12	4	3	12							

3.5. Taxonomic composition of herbal products using our customised barcode database and pipeline

To evaluate the performance of our pipeline and customised barcode database methods with complex herbal samples, authenticities of 14 herbal products were tested (**Table S2**). Results are shown in Table 3, **Figure S5**. In the seven dried products where each sample represents mainly one herbal plant species (*e.g.*, dried basil and dried tarragon **Table S2**, Table 3, **Figure S5**), the most abundant plant taxon in each sample corresponded to the expected herbal plant taxon (Table 3, **Figure S5**). For example, in the dried saffron sample (*Crocus*), *Crocus* was the first hit, with a relative read abundance of 95.9%. *Zingiber* (3.4%) (Table 3, **Figure S5**).

In the remaining seven herbal products (**Table S2**) where each sample is a mixture of spices and herbal products, several plant taxa were identified in the mapped sequences when using our customised pipeline and barcode database (Table 3, **Figure S5**). For example, the red curry sample (Table 3, **Figure S5**) is a mix of herbal products and, as expected, contained several plant genera with different relative read abundance values: *Allium* (40.0%), *Capsicum* (15.0%), *Armoracia* (9.6%), *Crocus* (9.2%), and *Elettaria* (6.7%). Finally, the roasted garlic pepper sample (Table 3, **Figure S5**) had *Allium* as the highest relative read abundance (76.5%) followed by *Anethum* (5.9%), *Elettaria* (3.1%), *Petroselinum* (2.1%), and *Curcuma* (1.6%)

4. Discussion

Here we present a novel methodology and algorithm-based analysis to metagenomically identify edible herb and plant taxonomy as a product authenticity assay. Our method is cutting-edge and shows high performance at correctly identifying the most dominant genera in a sample. The method also allowed us to molecularly identify the main components in herbal species and commercially-available plant-derived products. However, our study also highlight the current limitations of using NGS, namely the lack of sufficient high quality reference genomes and the need to perform deep sequence which at the current prices will limit the routinely use of NGS for food authenticity.

4.1. Algorithm validation

Ten publicly available datasets were included in the validation of the mapping script against iBOL database only without the influence of our customised barcode database. In eight of them, the first hit matched the expected genera, with relative read abundance values higher than the remaining identified genera (Table 1, **Figure S3**). Conversely, the results from the last two samples did not fit the previous description, in the case of cinnamon and fennel. Thus, taking only the first hits from all samples into consideration, the accuracy of the method was 80%. Nevertheless, noise was apparent in all the tested samples, which was sample dependent. This might be for two main reasons. First, there may have been contamination from other plants or species in the trimmed reads from each publicly available dataset. Second, closely related genera can have a different number of barcodes in the iBOL database, which may explain the noise within the results (Table 1, **Figure S3**). As a consequence, a customised database containing the limited number of genera present in species and herbal products could reduce noise.

To examine the impact of closely-related species, a phylogenetic tree containing all genera described in Table 1 (**Table S7**, **Figure S3**) was obtained from NCBI taxonomy. The false positives in each sample were mostly from closely related genera belonging to the same taxonomical order of the expected genus (**Figure S7**). Therefore, the alignment algorithm might incorrectly assign reads to the wrong species due to the high similarity between their genetic background presented in their barcode sequences.

If several of our barcodes were found in both genera, the alignment algorithm would ideally map them to the correct genus, even though

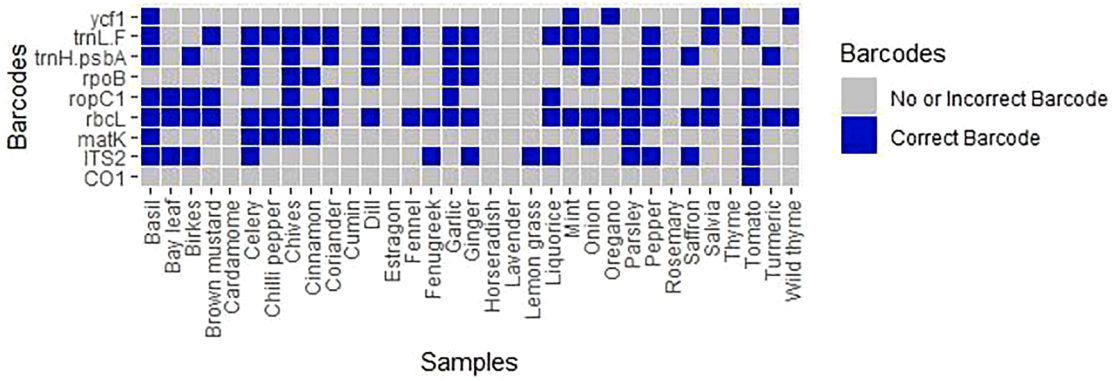

Fig. 2. Heatmap shows the obtained barcodes by mapping sequences from 33 single species plant using our customised algorithm.

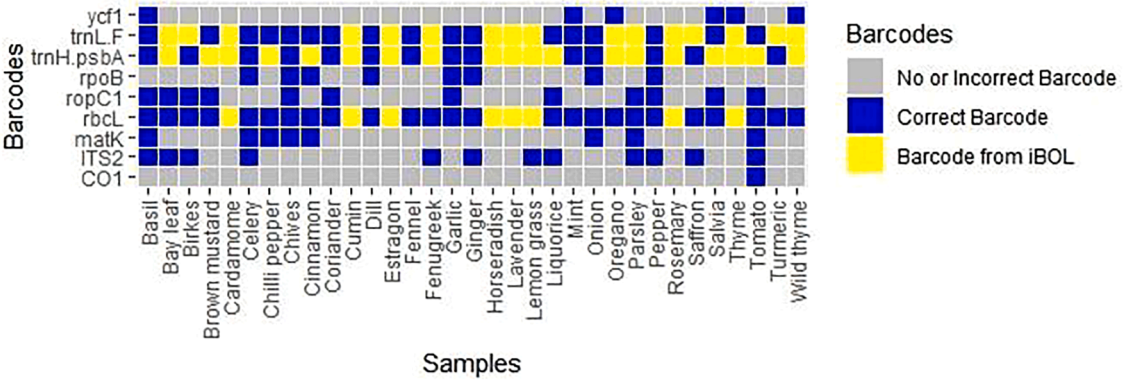

Fig. 3. Heatmap shows the structure of the customised barcode database that was obtained from 33 single species plant sequences. The barcode database was also supported by barcodes obtained from the publicly available database iBOL.

false positives could arise due to the high similarity between the query sequences. However, if the regions were only present in the closely related genus, the trimmed reads would map to these barcode sequences. Then, the hits assigned to this genus could be higher than the true value and, as a consequence, its relative read abundance would also be higher. These features also support the creation of a customised database in which each genus includes an equal number of sequences representing the same set of barcodes, as it could decrease the number of false positive hits. Therefore a barcode-based database for plant identification was generated here.

4.2. *Species and herbal product analysis*

The first hit for almost all of the 33 single-species samples (Table 2, **Table S1,** Figure S4) was the expected genus, with relative abundance values higher than the second hit. Hence, the overall accuracy of our pipeline and customised database was 93.9%. These data suggest that the customised barcode database has sufficient resolution to distinguish all the included genera. Although our methods showed high accuracy rates, signals of mismatches between the sequenced data and taxonomical hits were apparent (for example, authenticity test of Red Curry product Table 3, **Figure S5**). This is likely due to the low resolution of the barcodes for closely related species.

Authenticity analysis was based on comparisons between 14 herbal product compositional profiles (Table 3, **Figure S5**). Regarding the four herbal products (Basil, Estragon, Dill and Saffron), the two first hits of all samples matched with their respective plant taxonomical hits. Their main component complied with their label specifications along with the presence of other non-specific taxon hits.

With respect to the powdered products, we examined three single-species herbal products and three mixes. Cinnamon and paprika powder results were consistent with the data in their labels (Table 3, **Figure S5**). The garlic powder composition had more than one abundant plant genus in its plant taxonomical placement, even though the first hit was *Allium* (Table 3, **Figure S5**). Thus, a number of substitutes or non-declared species may be present in this powdered product.

The three types of curry powder contained several plant species representing different genera and, as mentioned above, the overall accuracy of the algorithm is decreased when several genera are included in a single sample. Nevertheless, the main components in each mix should be readily identifiable. Genera such as *Trigonella, Curcuma, Coriandrum*, and *Cuminum* were detected in the curry powders, which was consistent with the product label data. Furthermore, the other detected genera were closely related to the ones previously mentioned. Thus, the composition was consistent with the product description. The red and green curry samples had different plant taxonomical profiles, with their respective labels supporting these differences as the main products in the red curry were onion, garlic (*Allium*), and paprika (*Capsicum*) and the main products in green curry were garlic (*Allium*), ginger (*Zingiber*), and coriander (*Coriandrum*). The inclusion of other genera in their profiles could be due to the low resolution of the mapping algorithm between closely related genera or could indicate adulteration of the product.

Finally, we tested four ground products (Table 3, **Table S2,** **Figure S5**). The first and second hits from chili explosion were *Brassica* and *Armoracia*, which are not closely related to the main components reported on the product label as they belong to taxonomically different families. The main expected components, *Capsicum* and *Solanum*, were found in our data, although they were the least abundant genera. Regarding citron pepper, the only genus consistent with the label information was *Allium*. These results could indicate adulteration in these

Table 2
Relative abundances of plant taxa that were identified in 33 single plant species when mapped using our pipeline against our customised barcode database.

Sample name	Basil - *Ocimum*					Bay leaf - *Laurus*					Birkes - *Papaver*				
Genus	*Ocimum*	*Thymus*	*Mentha*	*Salvia*	*Anethum*	*Laurus*	*Cymbopogon*	*Cinnamomum*	*Elettaria*	*Papaver*	*Papaver*	*Artemisia*	*Cymbopogon*	*Crocus*	*Allium*
relative abundance	65.95	9.34	8.22	5.86	4.88	54.28	16.66	13.27	8.42	4.2	69.9	8.02	7.6	5.81	5.23
Sample name	Mustard - *Brassica*			Cardamome - *Elettaria*					Celery - *Apium*						
Genus	*Brassica*	*Anethum*	*Allium*	*Elettaria*	*Curcuma*	*Zingiber*	*Cinnamomum*	*Laurus*	*Apium*	*Petroselinum*	*Laurus*	*Brassica*	*Elettaria*		
relative abundance	82.14	17.83	0.01	63.15	26.53	8.23	0.86	0.53	71.94	25.16	1.08	0.73	0.49		
Sample name	Chili - *Capsicum*					Chives - *Allium*					Cinnamon - *Cinnamomum*				
Genus	*Capsicum*	*Solanum*	*Apium*	*Coriandrum*	*Elettaria*	*Allium*	*Elettaria*	*Laurus*	*Armoracia*	*Brassica*	*Cinnamomum*	*Laurus*	*Elettaria*	*Papaver*	*Brassica*
relative abundance	46.49	12.11	11.25	6.28	5.38	86.53	7.7	2.55	1.43	0.87	57.42	30.18	6.31	2.44	1.27
Sample name	Coriander - *Coriandrum*					Cumin - *Anethum*					Dill - *Anethum*				
Genus	*Coriandrum*	*Anethum*	*Armoracia*	*Apium*	*Artemisia*	*Anethum*	*Petroselinum*	*Artemisia*	*Coriandrum*	*Foeniculum*	*Anethum*	*Petroselinum*	*Artemisia*	*Cymbopogon*	*Crocus*
relative abundance	31.1	19.26	12.78	12.36	11.37	40.77	20.48	16.95	16.84	1.48	52.26	21.42	10.97	9.24	4.41
Sample name	Estragon - *Artemisia*					Fennel - *Foeniculum*					Fenugreek - *Trigonella*				
Genus	*Artemisia*	*Anethum*	*Zingiber*	*Elettaria*	*Crocus*	*Foeniculum*	*Petroselinum*	*Artemisia*	*Anethum*	*Coriandrum*	*Trigonella*	*Artemisia*	*Glycyrrhiza*	*Elettaria*	*Allium*
relative abundance	64.52	17.5	7.87	5	2.18	49.18	28.74	14.37	3.95	1.82	75.07	13.2	5.15	3.63	1.05
Sample name	Garlic - *Allium*					Ginger - *Zingiber*					Horseradish - *Armoracia*				
Genus	*Allium*	*Artemisia*	*Elettaria*	*Petroselinum*	*Zingiber*	*Zingiber*	*Elettaria*	*Curcuma*	*Petroselinum*	*Artemisia*	*Armoracia*	*Brassica*	*Cymbopogon*	*Anethum*	*Elettaria*
relative abundance	70.02	15.78	6.1	2.85	1.99	53.45	35.26	9.7	0.6	0.42	52.96	20.13	11.39	10.51	2.67
Sample name	Lavander - *Cymbopogon*					Lemongrass - *Cymbopogon*					Liquorice - *Glycyrrhiza*				
Genus	*Cymbopogon*	*Coriandrum*	*Glycyrrhiza*	*Crocus*	*Laurus*	*Cymbopogon*	*Mentha*	*Allium*	*Glycyrrhiza*	*Artemisia*	*Glycyrrhiza*	*Anethum*	*Cymbopogon*	*Crocus*	*Artemisia*
relative abundance	32.22	17.45	12.34	11.39	7.8	84.36	7.45	4.3	1.2	0.76	60.62	14.84	8.87	8.22	6.95
Sample name	Mint - *Mentha*					Onion - *Allium*					Oregano - *Origanum*				
Genus	*Mentha*	*Cymbopogon*	*Origanum*	*Thymus*	*Ocimum*	*Allium*	*Crocus*	*Cymbopogon*	*Mentha*	*Glycyrrhiza*	*Origanum*	*Mentha*	*Cymbopogon*	*Ocimum*	*Lavandula*
relative abundance	51.31	14.81	14.7	8.87	3.31	78	9.59	7.45	1.75	1.13	48.29	23.25	12.25	3.33	2.73
Sample name	Parsley - *Petroselinum*					Pepper - *Piper*					Rosemary - *Salvia*				
Genus	*Petroselinum*	*Foeniculum*	*Crocus*	*Salvia*	*Solanum*	*Piper*	*Crocus*	*Ocimum*	*Petroselinum*	*Anethum*	*Salvia*	*Mentha*	*Crocus*	*Thymus*	*Ocimum*
relative abundance	72.61	19.21	5.67	0.61	0.55	90.88	5.54	1.67	0.73	0.46	52.96	17.84	10.13	5.95	5.78
Sample name	Saffron - *Crocus*					Salvia - *Salvia*					Thyme - *Thymus*				
Genus	*Crocus*	*Solanum*	*Salvia*	*Petroselinum*	*Mentha*	*Salvia*	*Ocimum*	*Mentha*	*Crocus*	*Thymus*	*Thymus*	*Mentha*	*Origanum*	*Crocus*	*Salvia*
relative abundance	97.96	0.43	0.4	0.27	0.26	56.04	13.76	8.91	8.74	5.42	40.78	27.69	14.8	7.58	3.74
Sample name	Tomato - *Solanum*					Turmeric - *Curcuma*					Vanilla - *Vanilla*				
Genus	*Solanum*	*Crocus*	*Elettaria*	*Salvia*	*Origanum*	*Curcuma*	*Elettaria*	*Crocus*	*Zingiber*	*Petroselinum*	*Vanilla*	*Zingiber*			
relative abundance	82.81	9.48	5.93	1.42	0.08	53.92	32.01	7.48	5.36	0.36	96.25	3.74			

Table 3
Relative abundances of plant taxa that were identified in 14 herbal products for authenticity testing when mapped using our pipeline against our customised barcode database.

Herbal product name	Chilli explosion (mixture)					Cinnamon powder					
Genus	*Brassica*	*Armoracia*	*Anethum*	*Capsicum*	*Solanum*	*Cinnamomum*	*Laurus*	*Elettaria*	*Curcuma*	*Thymus*	
relative abundance	33.21	24.58	16.28	14.23	8.78	58.47	31.33	6.8	0.98	0.85	
Herbal product name	Citronpeppar (mixture)					Curry (mixture)					
Genus	*Allium*	*Laurus*	*Crocus*	*Ocimum*	*Anethum*	*Trigonella*	*Curcuma*	*Elettaria*	*Anethum*	*Foeniculum*	*Coriandrum*	*Cuminum*
relative abundance	35.42	11.93	8.89	8.69	7.95	29.35	21.39	10.06	7.87	6.76	5.19	4.49
Herbal product name	Dried basil					Dried dill						
Genus	*Ocimum*	*Thymus*	*Mentha*	*Salvia*	*Lavandula*	*Anethum*	*Petroselinum*	*Foeniculum*	*Crocus*	*Glycyrrhiza*		
relative abundance	64.13	9.1	8.27	5.93	4.49	59.88	22.55	11.72	4.46	0.42		
Herbal product name	Dried estragon					Garlic pepper (mixture)						
Genus	*Artemisia*	*Anethum*	*Cymbopogon*	*Elettaria*	*Crocus*	*Allium*	*Petroselinum*	*Artemisia*	*Cymbopogon*	*Crocus*		
relative abundance	75.49	8.49	8.26	4.97	1.51	43.73	16.85	12.39	11.81	7.29		
Herbal product name	Garlic powder					Green curry (mixture)						
Genus	*Allium*	*Glycyrrhiza*	*Crocus*	*Anethum*	*Artemisia*	*Allium*	*Coriandrum*	*Elettaria*	*Crocus*	*Zingiber*		
relative abundance	38.69	26.36	9.9	8.06	5.45	38.51	9.42	8.34	8.17	7.76		
Herbal product name	Paprika powder					Red curry (mixture)						
Genus	*Capsicum*	*Solanum*	*Elettaria*	*Coriandrum*	*Anethum*	*Allium*	*Capsicum*	*Armoracia*	*Crocus*	*Elettaria*		
relative abundance	53.71	27.07	6.86	6.16	2.06	40.02	15	9.63	9.24	6.7		
Herbal product name	Red garlic and pepper (mixture)					Saffron						
Genus	*Allium*	*Anethum*	*Elettaria*	*Petroselinum*	*Curcuma*	*Crocus*	*Zingiber*	*Salvia*	*Glycyrrhiza*	*Mentha*		
relative abundance	76.52	5.89	3.13	2.09	1.58	95.86	3.4	0.13	0.1	0.08		

products. In at least three herbal products (*e.g.*, citron pepper, red garlic and pepper Table 3, **Figure S5**), all the main components were detected apart from the pepper (*Piper nigrum*). These patterns support the hypothesis that this genus might not be detectable in the results due to its initial absence in the trimmed reads of the sequenced samples. The DNA extraction, library preparation, or the sequencing technology might be unsuitable for the detection of dry pepper in a mixed product.

This is a proof of concept for improving food authentication using sequencing technology. Our mapping algorithm with the customised barcode database correctly identified the components in most of the samples, albeit with noisy signals from mismatched taxa. Therefore, our mapping algorithm based on the alignment of trimmed reads against a barcode database was able to identify in details the herbal components in both plant species and commercial herbal products for food authentication, which can pave the way for evaluating the authenticity of different species or commercially available herbal products. Sequence resolution must, however, be improved to increase the qualitative and quantitative accuracy of the method, especially for complex samples.

5. Conclusions

Here we aimed to design and develop bioinformatics tools to evaluate the authenticity of commercially available herbs and products used in European and Danish cuisine. First, we implemented a metagenomic-based pipeline in Python3 and UNIX (Anna Delgado. Herbs Authenticity - GitHub. 2019. url: https://github.com/ADelgadoT/HerbsAuthenticate.git). The pipeline was tested with publicly available data to demonstrate its capacity to extract barcodes from trimmed reads and also to map them against public or customised barcode databases.

The use of the pipeline allowed the construction of a customised database barcode through the extraction of select marker genes (*rbcL, trnH-psbA*, and *trnL-F*) from all single species included in the study. Moreover, missing entries were incorporated from publicly available resources (iBOL and ENA). The final database contained 99 sequences with an average length of 465 bp.

All trimmed reads from single plant species were mapped against our customised barcode databases to evaluate the methods performance. All expected genera were detected in all the plant samples. Other plant taxa can still appear due to their low discrimination rate, as short sequences representing the same region of the genome tend to have high similarity in closely-related species.

Lastly, our barcoding mapping method could identify the predominant genera according to their respective labels in commercially available herbal samples where authenticity was evaluated.

Further research is required to improve the qualitative and quantitative accuracy of the approach and decrease noise, for example by using of full chloroplast sequences as queries in the alignment procedure.

Declaration of Competing Interest

The authors declare no conflict of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Jacob Dyring Jensen and Christina Aaby Svendsen for laboratory assistance. This work was partially supported by CIRCLES (Controlling mIcRobiomes CircuLations for bEtter food Systems) funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 818290 to FMA.

Data Availability

The datasets generated and analysed in this study are submitted to

ENA, project number PRJEB44059. All samples ENA accession numbers are in Table S9.

Author Contributions

ADT performed DNA extractions, prepared libraries, designed the customised barcode database and pipeline, performed bioinformatics analyses, interpreted the data and prepared first drafts of figures and the manuscript. PL designed the customised barcode database and pipeline, performed bioinformatics analyses, interpreted the data and contributed to the writing of the manuscript. FMA and SO designed the study, contributed to DNA extractions, analysed and interpreted the data and prepared the final version of the figures and the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochms.2021.100044.

References

Anantha, N. D. B., & Johnson, S. T. (2019). DNA barcoding in authentication of herbal raw materials, extracts and dietary supplements: a perspective. issn: 1863-5466. In *Plant Biotechnology Reports* (pp. 1–10). https://doi.org/10.1007/s11816-019-00538-z.

Black, C., Haughey, S. A., Chevallier, O. P., Christopher, P.-K., & Elliott, T. (2016). A comprehensive strategy to detect the fraudulent adulteration of herbs: The oregano approach. *Food Chemistry, 210*, 551–557. https://doi.org/10.1016/J.FOODCHEM.2016.05.004

Böhme, K., Calo-Mata, P., Barros-Velázquez, J., & Ortea, I. (2019). Review of Recent DNA-Based Methods for Main Food- Authentication Topics. *Journal of Agricultural and Food Chemistry, 67*(14), 3854–3864. https://doi.org/10.1021/acs.jafc.8b07016

CBOL — iBOL. url: http://www.ibol.org/phase1/cbol/ (visited on 06/19/2019).

Clausen Philip T. L. C., Frank M. Aarestrup, and Ole Lund (2018). "Rapid and pre- cise alignment of raw reads against redundant databases with KMA". In: BMC Bioinformatics 19.1, p. 307. issn: 1471-2105. doi: 10.1186/s12859- 018 - 2336 - 6. url: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2336-6.

Coissac, E., Hollingsworth, P. M., Lavergne, Sébastien, & Taberlet, P. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology, 25*(7), 1423–1428. https://doi.org/10.1111/mec.13549

Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. *TrAC - Trends in Analytical Chemistry, 85*, 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Delgado Anna. Herbs Authenticity - GitHub. 2019. url: https://github.com/ADelgadoT/HerbsAuthenticate.git.

Delia, G. (2019). Food Fraud. In *Encyclopedia of Food Security and Sustainability* (pp. 238–248). Elsevier. https://linkinghub.elsevier.com/retrieve/pii/B9780081005965215771. https://doi.org/10.1016/B978-0-08-100596-5.21577-1.

Dong, W., Chao, X., Li, C., Sun, J., Zuo, Y., Shi, S., Cheng, T., Guo, J., & Zhou, S. (2015). ycf1, the most promising plastid DNA barcode of land plants. *Scientific Reports, 5*(1), 8348. http://www.nature.com/articles/srep08348. https://doi.org/10.1038/srep08348

Drouet, S., Garros, L., Hano, C., Tungmunnithum, D., Renouard, S., Hagège, D., et al. (2018). A critical view of different botanical, molecular, and chemical techniques used in authentication of plant materials for cosmetic applications. *Cosmetics, 5*(2), 30. http://www.mdpi.com/2079-9284/5/2/30. https://doi.org/10.3390/cosmetics5020030

European Nucleotide Archive EMBL-EBI. url: https://www.ebi.ac.uk/ena (visited on 05/21/2019).

genomicepidemiology / kma — Bitbucket. url: https : / / bitbucket . org / genomicepidemiology/kma (visited on 04/28/2019).

Downey Gerard (2016). Advances in food authenticity testing. Woodhead Publishing is an imprint of Elsevier, isbn: 9780081002339. url: https://books. google.es/books?hl=es&lr=&id=Q-8QCgAAQBAJ&oi=fnd&pg=PP1&dq=herbs+ food+authenticity&ots=Yao44MySJC&sig=v4Xuw_4pyErJlXPvX6LYFfFN7oM# v=onepage&q=herbsfoodauthenticity&f=false.

Haynes, E., Jimenez, E., Pardo, M. A., & Helyar, S. J. (2019). The future of NGS (Next Generation Sequencing) analysis in testing food authenticity. issn: 0956-7135. In *Food Control* (pp. 134–143) https://www.sciencedirect.com/science/article/pii/S0956713519300581. https://doi.org/10.1016/J.FOODCONT.2019.02.010.

Hollingsworth, P. M., Li, D.-Z., van der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1702), 20150338. https://doi.org/10.1098/rstb.2015.0338

Illuminate Biodiversity - International Barcode of Life. url: http://ibol. org/site/ (visited on 12/17/2018).

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: from gene to genome. issn: 14647931. In *Biological Reviews* (pp. 157–166). https://doi.org/10.1111/brv.12104.

Medina, S., Pereira, J. A., Silva, P., Perestrelo, R., & Câmara, J. S. (2019). Food fingerprints – A valuable tool to monitor food authen- ticity and safety. issn: 0308-8146. In *Food Chemistry* (pp. 44–162) https://www.sciencedirect. com/science/article/pii/S0308814618319848. https://doi.org/10.1016/J.FOODCHEM.2018.11.046.

Mishra, P., Kumar, A., Nagireddy, A., Mani, D. N., Shukla, A. K., Tiwari, R., et al. (2016). DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnology Journal, 14*(1), 8–21. https://doi.org/10.1111/pbi.12419

Montowska, M., Fornal, E., Piątek, M., & Krzywdzińska-Bartkowiak, M. (2019). Mass spectrometry detection of protein allergenic additives in emulsion-type pork sausages. issn: 09567135. In *Food Control* (pp. 122–131) https://linkinghub.elsevier.com/retrieve/pii/S0956713519301835. https://doi.org/10.1016/j.foodcont.2019.04.022.

Pamela, G.-K., Haughey, S. A., & Elliott, C. T. (2018). Herb and spice fraud; the drivers, challenges and detection. issn: 09567135. In *Food Control* (pp. 85–97) https://linkinghub.elsevier.com/retrieve/pii/S0956713517306102. https://doi.org/10.1016/j.foodcont.2017.12.031.

Primrose, S., Woolfe, M., & Rollinson, S. (2010). Food forensics: Methods for determining the authenticity of foodstuffs. *Trends in Food Science and Technology, 21*(12), 582–590. https://doi.org/10.1016/j.tifs.2010.09.006

Sara, G., McPherson, J. D., & Richard McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333–351. http://www.nature.com/articles/nrg.2016.49. https://doi.org/10.1038/nrg.2016.49

Sasikumar, B., Swetha, V. P., Parvathy, V. A., & Sheeja, T. E. (2016). Advances in adulteration and authenticity testing of herbs and spices. *Advances in Food Authenticity Testing*, 585–624. https:// www.sciencedirect.com/science/article/pii/B9780081002209000229. https://doi.org/10.1016/B978-0-08-100220-9.00022-9

Tnah, L. H., Lee, S. L., Tan, A. L., Lee, C. T., Ng, K. K. S., Ng, C. H., et al. (2019). DNA barcode database of common herbal plants in the tropics: a resource for herbal product authentication. issn: 0956-7135. In *Food Control* (pp. 318–326) https://www .sciencedirect.com /science/article/pii/ S0956713518304298. https://doi.org/10.1016/J.FOODCONT.2018.08.022.

Walker, M. J., Burns, M., Quaglia, M., Nixon, G., Hopley, C. J., Gray, K. M., et al. (2018). Almond or Mahaleb? Orthogonal Allergen Analysis During a Live Incident Investigation by ELISA, Molecular Biology, and Protein Mass Spectrometry, 8. In *Journal of AOAC International* (pp. 162–169). https://doi.org/10.5740/jaoacint.17-0405.

Yat-Tung, L., & Shaw, P.-C. (2018). DNA-based techniques for authentication of processed food and food supplements. issn: 03088146. In *Food Chemistry* (pp. 767–774) https://linkinghub.elsevier.com/retrieve/pii/S0308814617313407. https://doi.org/10.1016/j.foodchem.2017.08.022.

Yu, H. P. (2018). DNA barcoding angiosperms: identifying porulaca species us- ing matK and ITS2. *Inti International University*. http://eprints.intimal.edu.my/1171/.