

## Article

# Utilizing Big Data to Identify Tiny Toxic Components: *Digitalis*

Elizabeth Sage Hunter, Robert Literman and Sara M. Handy \* 

Center for Food Safety and Applied Nutrition, Office of Regulatory Science, U.S. Food and Drug Administration, College Park, MD 20740, USA; Elizabeth.Hunter@fda.hhs.gov (E.S.H.); Robert.Literman@fda.hhs.gov (R.L.)

\* Correspondence: sara.handy@fda.hhs.gov

**Abstract:** The botanical genus *Digitalis* is equal parts colorful, toxic, and medicinal, and its bioactive compounds have a long history of therapeutic use. However, with an extremely narrow therapeutic range, even trace amounts of *Digitalis* can cause adverse effects. Using chemical methods, the United States Food and Drug Administration traced a 1997 case of *Digitalis* toxicity to a shipment of *Plantago* (a common ingredient in dietary supplements marketed to improve digestion) contaminated with *Digitalis lanata*. With increased accessibility to next generation sequencing technology, here we ask whether this case could have been cracked rapidly using shallow genome sequencing strategies (e.g., genome skims). Using a modified implementation of the Site Identification from Short Read Sequences (SISRS) bioinformatics pipeline with whole-genome sequence data, we generated over 2 M genus-level single nucleotide polymorphisms in addition to species-informative single nucleotide polymorphisms. We simulated dietary supplement contamination by spiking low quantities (0–10%) of *Digitalis* whole-genome sequence data into a background of commonly used ingredients in products marketed for “digestive cleansing” and reliably detected *Digitalis* at the genus level while also discriminating between *Digitalis* species. This work serves as a roadmap for the development of novel DNA-based assays to quickly and reliably detect the presence of toxic species such as *Digitalis* in food products or dietary supplements using genomic methods and highlights the power of harnessing the entire genome to identify botanical species.

**Keywords:** dietary supplements; genome skimming; *Digitalis*; toxic botanicals



**Citation:** Hunter, E.S.; Literman, R.; Handy, S.M. Utilizing Big Data to Identify Tiny Toxic Components: *Digitalis*. *Foods* **2021**, *10*, 1794. <https://doi.org/10.3390/foods10081794>

Academic Editors: Rosalee S. Hellberg and Robert Hanner

Received: 15 June 2021  
Accepted: 27 July 2021  
Published: 3 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In 1785, physician and amateur botanist William Withering introduced the medical world to *Digitalis purpurea*. While he successfully used *Digitalis* preparations to treat a broad variety of cardiac conditions related to fluid retention, withering himself also noted the toxicity of this plant alongside its medicinal value, cautioning its use be discontinued as soon as symptoms of toxicity arose [1]. At the height of its use in the 1960s and 1970s, surveys estimated that up to 15% of all hospital patients were using digitalis-derived medications, and 20–30% of those patients were likely experiencing symptoms of toxicity [2]. These symptoms include cardiac manifestations, vomiting, anorexia, diarrhea, dizziness, fatigue, delirium, syncope, and visual disturbances such as xanthopsia (yellow vision) [1,3,4], and it has even been suggested that Vincent Van Gogh was experiencing xanthopsia due to digitalis toxicity during the last years of his life [5]. Today, two cardiac glycosides are still isolated for use from *Digitalis*: digitoxin and digoxin, and Withering's warning regarding their narrow therapeutic range holds true almost 250 years later. Due to the risks of toxicity, *Digitalis* derived drugs have become less popular in modern medicine [6], but its potent toxic effects still pose risks in cases of accidental contamination.

In 1997, two women were hospitalized with symptoms consistent with *Digitalis* toxicity several days after starting a dietary supplement regimen intended to provide “internal cleansing” [7]. The cases were brought to the United States Food and Drug Administration (FDA), which investigated them using a chemical-based method including a Kedde reaction and thin layer chromatography. The presence of cardiac glycosides was confirmed in the

raw materials indicated by the initial analyses using liquid chromatography and mass spectroscopy, and the species (*Digitalis lanata*) was confirmed using microscopy [7]. The source was identified as a batch of *Plantago*, which is commonly known as plantain, although not related to the tropical banana plant [7].

The world of eukaryotic DNA-based species identification in food and dietary supplements products has undergone a revolution in the past 10 years, and while progress has been swift it has not been without challenges [8–10]. Contemporary methods of DNA-based identification include DNA fingerprinting techniques like restriction fragment length polymorphisms [11], microarrays [12], a full suite of polymerase chain reaction (PCR) methods, [13–17], in addition to traditional DNA barcoding [9,18] and a variety of next generation sequencing (NGS) approaches [19–21].

NGS methods have recently been investigated to distinguish between closely related eukaryotic species [21,22] and to detect food origins [23], the presence of toxic species [24], allergens [25], as well as other cases of adulteration [20,26,27], both accidental and economically motivated. However, the application of NGS technology is not without limitations. Targeted barcoding regions can suffer from amplification and primer bias, while whole-genome sequencing (WGS) methods are often restricted by taxonomically limited sequence databases and lack of well-annotated reference genomes, especially for non-model organisms. Chloroplast databases can provide a uniform and informative reference for identification, but certain closely related or hybridized species are indistinguishable or confounded using these data [25]. Additionally, although chloroplast DNA is typically found in high copy numbers in many plant cells, plastid DNA proportions have been shown to vary due to both tissue type [28] and plant age [29], and large nuclear genomes can decrease the relative representation of plastid DNA in a genome skimming run [30–32].

Nuclear DNA methods may be required for robust detection of trace adulteration, but quality whole-genome references for flowering plants are still few and far between due to their size and repetitive structure. As of 2017, only 236 angiosperm genomes had been sequenced [33] out of an estimated 350,699 species ([www.theplantlist.org](http://www.theplantlist.org), accessed on 17 May 2021). This represents less than 0.07% of angiosperms, and a curated but comprehensive database remains an elusive goal. Data from the broadest available nuclear databases (e.g., NCBI) can suffer from poor sequence quality, inaccurate identification, and contamination [34]. Voucher collections have been leveraged to mediate some of these problems by providing traceable, expertly identified reference material for chemical analyses [35]. At the FDA, these samples are also being used to facilitate the creation of DNA-based reference databases [36] and novel molecular pipelines [25] that could circumvent complications involved with chemical analyses [10].

In this study we present a modified implementation of Site Identification from Short Read Sequences (SISRS) [37], a reference-free ortholog discovery pipeline that takes WGS data as input, to identify toxic adulterants (*Digitalis*) in mixed samples both at the genus and species level. SISRS does not rely on external databases or annotated reference genomes (which do not exist for *Digitalis*) when identifying loci useful for classification, overcoming a major hurdle in many contemporary NGS pipelines [34]. From these SISRS loci we identify over 2 million genus-informative single nucleotide polymorphisms (SNPs) for *Digitalis* and thousands of species-informative SNPs. As in the real-world case of *Digitalis* toxicity, we spike *Digitalis* WGS data into simulated mixtures containing *Plantago*, another member of the Plantaginaceae family, along with other commonly used botanical ingredients in products marketed for digestive health. SNPs generated in this study facilitated robust detection of *Digitalis* adulteration at read proportions of 0.05%, and accurate species identification when datasets contained ~4 Mb of *Digitalis* data. Using test species that allow us to assess the sensitivity and specificity of our markers even when differentiating amid closely related botanicals, here we examine how methods utilizing the nuclear data from low-coverage WGS datasets (i.e., ‘genome skims’) might be used in the future to detect even minute amounts of contamination with toxic plant species in foods and dietary supplements.

## 2. Materials and Methods

### 2.1. Sample Acquisition

#### 2.1.1. DNA Extraction and Sequencing

DNA was extracted from thirty-two vouchered Plantaginaceae specimens from the New York Botanical Gardens (<https://www.nybg.org/plant-research-and-conservation/tour/laboratory-collections> accessed on 17 May 2021) and the University of Rhode Island's Heber W. Youngken Medicinal Garden (<https://web.uri.edu/youngken-garden/> accessed on 17 May 2021) using the QIAGEN DNeasy Plant Mini Kit. Samples were leaf material dried on silica, which were held at 4 C until extraction. The samples included species from the Plantaginaceae family: three *Digitalis*, eleven *Plantago*, fifteen *Veronica*, and three additional species (*Callitriche stagnalis*, *Bacopa monneri*, and *Gratiola ramosa*). Samples were quantified with a broad range Qubit 3 fluorometer assay to ensure successful extraction. Libraries were then prepared for Illumina sequencing, following [38], using a Covaris M220 sonicator to shear to approximately 300 bp, followed by end repair, adapter ligation, amplification, and subsequent AmPure cleanups using the KAPA HyperPrep kit and dual indexed adapters. Libraries were quantified and sized using a Qubit 3 fluorometer and an Agilent 2100 Bioanalyzer, respectively. Nine samples underwent an additional AmPure bead cleanup and quality control process to remove residual primer dimers. Samples were pooled at a 6nM concentration and sequenced on a single lane of the Illumina HiSeq2500 platform (Genewiz, South Plainfield, NJ, USA). The same libraries of *Digitalis grandiflora*, *D. ferruginea*, *D. lutea*, and also *Plantago patagonica* and *P. rhodosperma* were also sequenced on an in-house Illumina MiSeq with a 600-cycle kit, along with an additional independent sample of *D. ferruginea*. Five *Digitalis* samples were additionally sequenced on an in-house Illumina NextSeq500, including the three *Digitalis* libraries from the original HiSeq2500 run, and two additional vouchered samples provided by the University of Mississippi. These last two samples were extracted and prepared for sequencing using the same methods as above. Raw reads from these sequencing runs were deposited under the BioProject PRJNA325670, and their SRA accessions and associated BioSample IDs are available in the Supplementary Tables S2 and S5).

#### 2.1.2. Companion Data from Public Archives

In addition to specimens sequenced as part of this study, we also acquired publicly available sequence data for an additional *Digitalis purpurea* sample, as well as several Plantaginaceae species (*Littorella uniflora*, *Plantago lagopus*, *Plantago ovata*, and *Veronica agrestis*). In addition to these data, we also acquired WGS data for two specimens each from *Aloe vera* and *Linum usitatissimum* (flaxseed). These species were the third and fourth most common botanical ingredients found to co-occur with *Plantago* (plantain) based on a screen of twenty products marketed for digestive health on Amazon. No WGS data were available for the first and second most common co-occurring ingredients, *Rhamnus purshiana* (cascara sagrada) and *Cassia angustifolia* (senna). NCBI identification numbers for all companion data can be found in the appropriate Supplemental Tables (Tables S2, S5, and S7).

### 2.2. Nuclear Enrichment of *Digitalis* WGS Data

For one specimen from each of the five *Digitalis* species, we used *bbmerge.sh* and *bbduk.sh* from the BBTools suite (<https://jgi.doe.gov/data-and-tools/bbtools/>, Last accessed on 21 May 2021) to perform read merging and automated adapter removal. We used *getOrganelle* [39] to assemble circularized chloroplast genomes from these adapter-trimmed reads using kmer values of 21, 45, 65, 85, and 105 and a maximum of 50 extension rounds. Except for the two *Digitalis* specimens used in the multispecies mixes, all *Digitalis* read sets were screened to remove chloroplast-derived data. We first quality-trimmed raw *Digitalis* reads using *bbduk.sh* from the BBTools suite (*qtrim = w*, *ktrim = r*, *trimq = 10*, *maq = 15*, *minlength = 50*), then assessed quality with *FastQC* [40]. After pooling the chloroplast assemblies generated above, we used *bbmap.sh* from the BBTools suite to map and screen these

quality-trimmed reads against this multispecies chloroplast reference dataset, separating reads that mapped (presumably chloroplast-derived reads) from nuclear-derived reads.

### 2.3. Composite Genome Assembly

In order to identify nuclear loci that would be useful for *Digitalis* detection in the absence of a reference genome, we used the SISRS bioinformatics pipeline [37] to generate conserved, orthologous sequence data directly from raw sequencing reads. SISRS generates a so-called ‘composite genome’ (i.e., a *de novo* pan genome) by pooling reads proportionally across species to a final depth of ~10X genomic coverage, and then performing a single *de novo* genome assembly. Using a genome size estimate of 2 Gb for the genus [41], we used *reformat.sh* from the BBTools suite to subsample reads equivalently among *Digitalis* species and evenly among specimens therein, resulting in a final genome assembly dataset containing 20 Gb of primarily nuclear *Digitalis* data. We used Ray v.2.3.2-devel [42] to assemble the composite genome using the subsampled nuclear reads, default parameters, and a k-value of 31. This composite genome represents a ‘taxonomically-averaged’, common set of nuclear loci against which read data from all *Digitalis* species can be mapped and compared.

### 2.4. Nuclear *Digitalis* Read Mapping and SNP Dataset Development

SISRS uses read data from individual specimens, species datasets containing multiple individuals, or specimens pooled at a higher taxonomic level to profile allelic variation at each site in the composite genome. In this study, we profiled *Digitalis* data at the species level (i.e., specimen data pooled together by species) and at the genus level (i.e., pooling all nuclear *Digitalis* datasets together). In combination, these datasets allowed us to isolate (1) all nuclear sites that had a single, fixed allele that was present in all five *Digitalis* species (i.e., putative genus-informative DNA markers) as well as sites that were variable among the *Digitalis* species (i.e., putative species-informative single-nucleotide polymorphisms or SNPs).

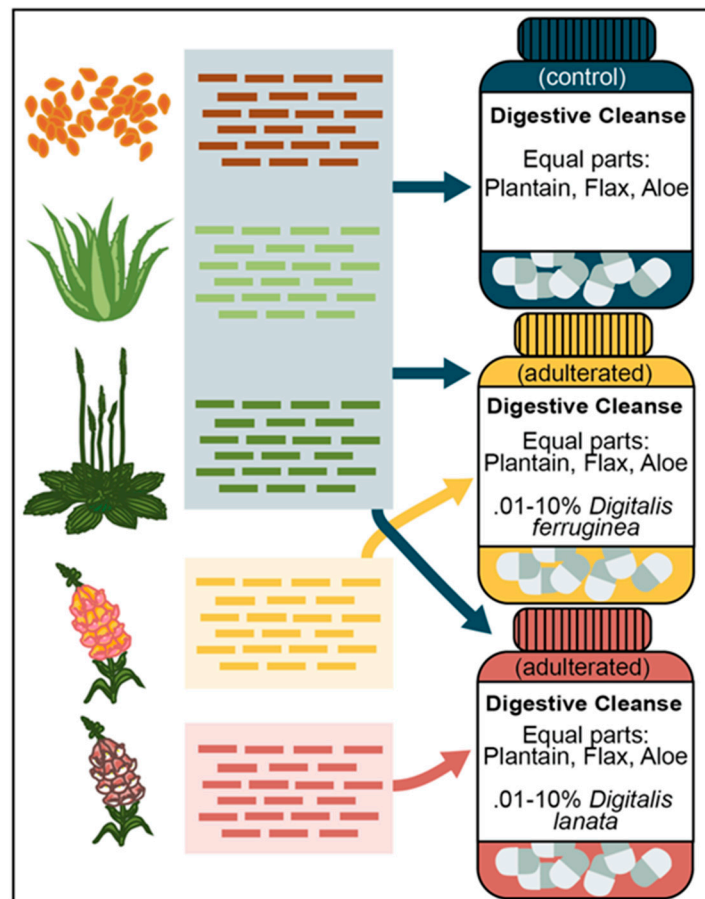
### 2.5. Background Species Mapping and Data Filtration

For both species- and genus-level identification from multi-species mixtures, discernibility will be restricted if identifying alleles are also present in more distantly related taxa (i.e., if they are non-specific). To enrich our dataset for *Digitalis*-specific SNPs, we used SISRS to map WGS data from 34 non-*Digitalis* Plantaginaceae species onto the composite genome, along with one specimen of aloe and flax. We quality-trimmed all read sets using *bbduk.sh* from the BBTools suite as before, but, contrary to the *Digitalis* samples, we did not separate these datasets into nuclear and chloroplast fractions and mapped all quality-trimmed reads. Based on these mappings, we pruned both the genus- and species-level SNP datasets of sites that had read coverage from any of the non-*Digitalis* species.

The SNPs generated above represent sites with informative variation, pre-screened against close evolutionary relatives and other species known to co-occur in mixtures. These sites would represent a good starting dataset for confirmatory tests of single specimens; however, in mixed-species samples, that could contain any number of co-occurring species, sites can be further filtered down to reduce the impact of cross-species read mapping. Using the genus-level SNP set, consisting of sites with *Digitalis*-specific variation among Plantaginaceae, aloe, and flaxseed, we tallied the number of screened SNPs that occurred on each composite genome contig and calculated the proportion of SNPs relative to the total contig length. We then filtered both the genus- and species-level SNP sets down to just those deriving from contigs with at least the median number and proportion of genus-level SNPs. This filtering removes contigs (and their SNPs) that contain predominantly non-specific data and enriches the dataset for regions where *Digitalis* has experienced potentially informative mutations.

## 2.6. Generating Mixed Samples

In order to assess the utility and sensitivity of the species- and genus-level SNPs for identifying *Digitalis* in multispecies mixtures, we spiked increasing amounts of WGS data from *Digitalis ferruginea* and *D. lanata* into a series of simulated mixtures containing equal amounts of WGS data from *Plantago major*, *Aloe vera*, and *Linum usitatissimum* (Figure 1). All read samples were quality-trimmed as before, and no chloroplast read separation was performed. For each mixture, we targeted a practically relevant base depth of ~1.88 Gb, equivalent to an Illumina MiSeq sequencing run with 8 multiplexed samples. For both *Digitalis* species, we generated ten sets of ‘adulterated’ data by adding *Digitalis* WGS data to final proportions ranging from 0.01% (187,500 bases) to 10% (187.5 Mb). Each mixture, including a control dataset that contained no *Digitalis* data, had ten pseudoreplicates generated by randomly subsampling reads from the quality-trimmed datasets using *reformat.sh* from the BBTools suite. For each pseudoreplicate, data from *D. ferruginea* and *D. lanata* were added to identical background mixtures (i.e., Replicate B1 was mapped twice using the same data for *Plantago*, *Aloe*, and *Linum*, while varying the *Digitalis* species), but background data varied among replicate sets (i.e., Replicates B1 and B2 contain different random subsets of *Plantago*, *Aloe*, and *Linum* data).



**Figure 1.** Overview of mock dietary supplement sample design. All mock supplement mixes were created in silico from sequencing data. The control sample consisted of equal parts *Plantago major*, *Aloe vera*, and *Linum usitatissimum* reads. The adulterated supplement mixes contained the same background as the control but also included 0.01%, 0.025%, 0.05%, 0.1%, 0.25%, 0.5%, 1%, 2.5%, 5%, and 10% of either *Digitalis lanata* or *Digitalis ferruginea* reads.

## 2.7. Screening Mixed Samples for *Digitalis* SNPs

We used SISRS to map each of the mix pseudoreplicates against the composite genome, and we surveyed the alleles present at each of the *Digitalis* genus- and species-informative sites that survived upstream screening and filtering. For each pseudoreplicate, we performed both a genus- and species-level inquiry; in both cases we queried the relevant SNPs and tallied (1) sites that had read coverage and (2) sites that contained the *Digitalis* genus- or species-informative allele.

### 2.7.1. Genus-Level *Digitalis* Detection

For genus-level detection, we first used linear models in R [43] to assess the relationship between the number of *Digitalis* bases added to the mixture and the number of recovered *Digitalis* alleles. We then compared the number of *Digitalis* genus-informative alleles detected in the negative control group (where no *Digitalis* data were added) to the count from each pseudoreplicate group using *t*-tests in R, assessing significance at  $\alpha = 0.05$  (i.e., Does this set of pseudoreplicated samples contain more matching *Digitalis* genus SNPs than a sample with no *Digitalis* added?).

### 2.7.2. Species-Level *Digitalis* Detection

For all species-informative SNP positions with coverage in each pseudoreplicate, we calculated the ratio of SNPs supporting and refuting each species assignment and assigned species based on the highest ratio (i.e., matching species-informative alleles/all species-informative SNP positions with coverage) and used a modified Z-score test [44] to statistically assess whether that SNP proportion was statistically distinct from the other species. As a median-based test, the modified Z-score test is robust when comparing small samples sizes (i.e., 5 species), and significance of all p-values was assessed after Bonferroni correction ( $\alpha = 0.05/5 = 0.01$ ). To reduce species assignments involving too few data, we only performed statistical species identification on pseudoreplicates where (1) there was read data in the mixture that covered 5 species-informative SNP positions for each *Digitalis* species, and (2) where the species with the highest ratio of matching species-informative SNPs was supported by at least 5 matching alleles.

## 3. Results

### 3.1. Assembly of the Nuclear *Digitalis* Composite Genome

Chloroplast genomes for each of the five survey *Digitalis* species were assembled with getOrganelle using between 662 Mb and 3.8 Gb of adapter-trimmed reads, and all assemblies resulted in a circular genome (Table S1). Trimmed base counts for the *Digitalis* samples ranged from 629 Mb to 24 Gb per species, and 90–98% of reads across datasets failed to map to the chloroplast dataset and were considered putatively nuclear (Table S2). Using 2 Gb as a genome size estimate for the genus, to achieve 10X genome coverage we targeted a subsampling depth of 4 Gb for each of the five species; however, the *D. purpurea* sample only had 572 Mb of nuclear data and we made up the difference equally among taxa. The composite genome assembly generated by Ray resulted in 1.8M largely fragmentary contigs totaling 334 Mb (N50: 666Kb; L50: 172; Table S3).

### 3.2. Nuclear *Digitalis* Mapping and SNP Dataset Generation

SISRS mapping of the nuclear *Digitalis* species datasets onto the composite genome resulted in base calls for 21 M–159 M sites per species (6.5–47.8% of sites; Table S4), while the *Digitalis* genus dataset containing data from all species resulted in base calls for 272 M sites (81.5% of sites; Table S4). Of these, 6.4 M sites had a single, fixed allele found in all five *Digitalis* species (Table S4). Among the five *Digitalis* species, species-informative SNP counts (sites with a fixed base for all species, and a unique allele for one species) ranged from 14 K–223 K sites per species (Table S4). After removing sites with coverage from non-*Digitalis* species (Table S5) we filtered contigs by the proportion and amount of *Digitalis*-specific sites, allowing SNPs from contigs where (1) at least 22.4% of sites in the

contig were *Digitalis*-specific (median proportion) and (2) with at least 53 *Digitalis*-specific sites (median site count). After filtering, the final *Digitalis* genus SNP count was 2.4 M sites, while species-informative SNPs ranged from 6.3 K–101 K sites per species (Table S4).

### 3.3. Screening Mixed Samples

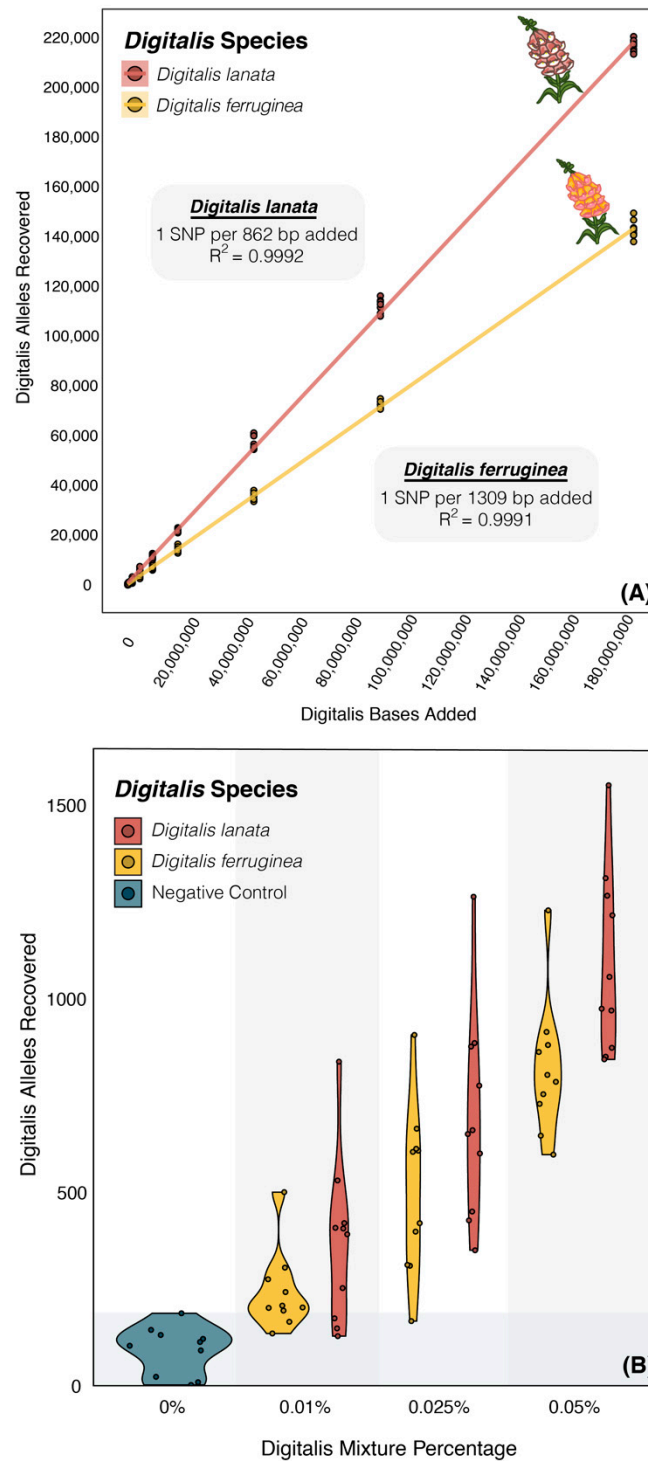
To create the simulated sample mixtures, we pooled equivalent amounts of trimmed WGS data from *Plantago*, *Aloe*, and *Linum* with increasing amounts of spike-in *D. ferruginea* or *D. lanata* data (Table S6) ranging from 0.01% to 10% (Table S7). Each simulated sample contained ~1.9 Gb of trimmed WGS data. We used SISRS to map all reads from each mix pseudoreplicate against the composite genome ( $n = 10$  per spike species, per spike amount), and defined an SNP match as a site where either a genus- or species-informative allele was detected.

#### 3.3.1. Genus-Level Screening

For mix pseudoreplicates spiked with *Digitalis* data, the number of recovered alleles indicating the presence of the genus *Digitalis* was highly correlated with the number of *Digitalis* bases added. *Digitalis ferruginea* resulted in ~1 new SNP match for every 1309 bases added ( $p < 2.2 \times 10^{-16}$ ,  $R^2 = 0.9991$ ), and *D. lanata* resulted in 1 new SNP match for every 862 bases added ( $p < 2 \times 10^{-16}$ ,  $R^2 = 0.9992$ ; Figure 2A; Table S8). The negative control samples with no *Digitalis* data added had positive hits for 8–193 sites (Figure 2B; Table S7), while spiking in the lowest amount of either *D. ferruginea* or *D. lanata* (~187 Kb or ~0.01%) resulted in a significantly higher average of 339 and 360 sites, respectively (both  $p < 2.5 \times 10^{-3}$ ; Figure 2B; Tables S7 and S9). All *Digitalis*-spiked pseudoreplicate groups had a statistically higher average number of *Digitalis* genus matches relative to the negative control group (all  $p < 2.5 \times 10^{-3}$ ; Table S9), but the distribution of matched sites overlapped with the negative control samples at the two lowest spike-in concentrations for both *Digitalis* species (0.01% and 0.025%, 187 Kb and 468 Kb) (Figure 2B).

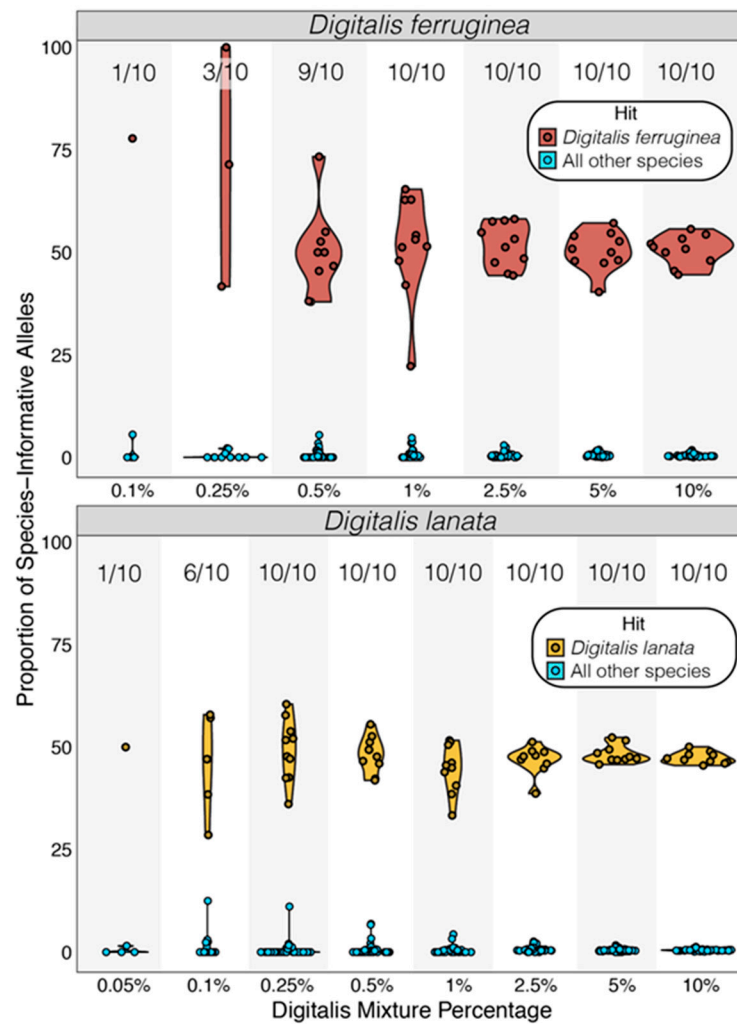
#### 3.3.2. Species-Level Screening

We queried the alleles present at all species-informative SNP positions and only classified samples where (1) each species had 5 SNP positions covered, and (2) at least 5 matching SNPs supported the top species hit. Out of the 100 *Digitalis*-spiked pseudoreplicates for each species (10 pseudoreplicates  $\times$  10 spike-in amounts), *D. ferruginea* had 53 pseudoreplicates that passed this filtering, while 67 *D. lanata* pseudoreplicates contained sufficient data. In each case, the adulterant species in each qualifying pseudoreplicate was correctly identified (Figure 3; all  $p < 9.7 \times 10^{-35}$ ; Tables S10 and S11) and all pseudoreplicates could be classified when 18.75 Mb of *Digitalis* data (1% of the total data) was added (Figure 3; Tables S11 and S12). At lower *Digitalis* concentrations, some *D. ferruginea* could also be identified at 0.1% ( $n = 1/10$ ), 0.25% ( $n = 3/10$ ), and 0.5% ( $n = 9/10$ ) (Figure 3; Tables S11 and S12), while *D. lanata* was correctly identified in all pseudoreplicates containing at least 4.6 Mb (0.25% of the total data; Tables S11 and S12) and also at 0.05% ( $n = 1/10$ ) and 0.1% ( $n = 6/10$ ) (Figure 3; Tables S11 and S12). In all cases, 44–77% of SNPs that associated with the spike-in species had the correct species allele, while the second-best match never reached above 5.6% (Figure 3; Table S12).



**Figure 2.** *Digitalis* detection using SISRS. **(A)** Spiking more WGS data from *Digitalis ferruginea* or *D. lanata* into simulated mixtures containing *Plantago*, *Aloe*, and *Linum* resulted in a strongly correlated increase in the number of *Digitalis* genus-informative alleles recovered from mixed samples ( $p < 2.0 \times 10^{-16}$ ;  $R^2 > 0.999$ ). **(B)** At low concentrations of adulterating *Digitalis* (0.01–0.25%), the distribution of detected *Digitalis* genus-informative alleles overlapped with that of the negative control that had no *Digitalis* data added (grey bar), although group means were still significantly higher (all  $p < 2.5 \times 10^{-3}$ ). When *Digitalis* data made up 0.05% or more of the data (~937 Kb), there was an unambiguously significant increase in detected alleles (all  $p < 3.47 \times 10^{-7}$ ).





**Figure 3.** Accuracy of *Digitalis* species detection. For all mixture pseudoreplicates with sufficient data, the species of *Digitalis* was assigned based on the proportion of SNPs from each sample that contained species-informative alleles. In all cases the highest ratio of matching SNPs derived from the correct species, and all comparisons were significant based on modified Z-score analysis (all  $p \leq 9.72 \times 10^{-35}$ ). Proportions represent pseudoreplicates meeting the minimum data criteria. When *Digitalis* data accounted for 1% or more of the dataset (18.75 Mb), 10/10 pseudoreplicates could be identified, but limited data precluded robust species identification at some lower concentrations.

#### 4. Discussion

Identifying ingredients of complex, mixed samples using DNA methods is an ongoing challenge, and the method that is chosen for a particular question must be carefully considered. Metabarcoding approaches often rely on well-curated reference databases, but due to the limited size of many of these datasets, care must be taken to choose marker regions that both (1) provide appropriate taxonomic discrimination while (2) also accounting for issues such as amplification errors, bias, and artifacts introduced by PCR. Additionally, metabarcoding approaches often preclude quantification beyond relative abundance, and relying on highly targeted techniques can also fail due to fragmented DNA [21], which is often found in highly processed food products and dietary supplements.

On the other end of the spectrum, whole-genome sequencing (WGS) approaches, including low-coverage genome skimming, circumvent many of the limitations associated with traditional metabarcoding or targeted locus methods. WGS data provides an untargeted snapshot of DNA in a sample; it bypasses the need for *a priori* marker selection, reduces negative biases associated with PCR amplification, and provides a more holistic,

semiquantitative (with caveats), representation of sample composition [21]. Especially in the absence of a reference genome, analysis of WGS data typically relies on kmer binning [45–47], BLAST [48], or metagenomic assembly and classification of the resulting contigs using a database [49,50]. The availability of genome annotation data or useful marker sets in botanicals varies widely outside of model clades, and in many cases the taxonomic discrimination of analyses is severely limited by database robustness, or lack thereof. As an attempt to overcome these challenges, many WGS-based identification approaches still begin with a reduction down to smaller, more manageable data subsets, such as extraction of organellar reads [36,51], or specific sets of well-characterized genes [52]; yet, based on both their size (e.g., datasets containing few loci) and the nature of their heredity (e.g., in the case of maternal chloroplast inheritance), these data subsets also present challenges when identifying closely related species or hybrids.

Here we present a pipeline that relies only on WGS data, without the need for a reference genome, annotation data, or any external databases, to generate informative SNPs for the identification of toxic adulterants (*Digitalis*) in mixed botanical samples. This pipeline overcomes both the need for *a priori* marker selection and PCR amplification biases associated with metabarcoding, as well as the typical issues associated with work in non-model clades. While recent plastid-based analyses of *Anemopaegma* and *Veronica* yielded informative SNP counts in the thousands [53,54], here we identify over 2.4 million nuclear genus-informative SNPs for *Digitalis* that were screened against a background of *Plantago* (plantain), *Aloe vera* (aloe), and *Linum usitatissimum* (flax) (Table S4, Figure 1), as well as tens of thousands of species-informative SNPs. Rather than relying on a small subset of plastid reads or gene subsets, working from a large, untargeted nuclear dataset dramatically increased both our resolution and ability to recover orders of magnitude more SNPs when compared to many existing methods.

We simulated mixed samples based on practical data limits (~2 Gb per sample, equivalent to a MiSeq 600-cycle kit multiplexed eight ways) and at this depth unambiguous genus-level *Digitalis* detection was possible even in trace amounts when *Digitalis* data made up only 0.05% of the mixture (~1 Mb of spike-in data) (Figure 2B). Additionally, we see a highly-correlated increase in genus-level SNP recovery with respect to the amount of *Digitalis* data spiked into the mixture (Figure 2A), suggesting that this method is semiquantitative, supporting previous work [21]. Relative to the genus-informative SNP sets, there were fewer species-informative SNPs, and robust identification of species required around 10 times more data (0.5% of the mixture, or just under 10Mb) (Figure 3).

The nuclear loci generated by SISRS [37] were assembled using only *Digitalis* nuclear data, and as such, rescreening this dataset against species other than *Plantago*, *Aloe*, and *Linum* would theoretically allow for detection of *Digitalis* in mixtures containing any primary ingredients. Additionally, these datasets are dynamic, and more data from the existing *Digitalis* species or data from background species not included here could be analyzed without the need to generate a new composite genome from scratch. For instance, we screened our *Digitalis* markers against a single specimen of most outgroup taxa to reduce the amount of cross-species mapping; yet the unadulterated control samples had ~100 SNP matches that should have been *Digitalis*-specific (Figure 2B). In a real-world test case these sites too could be purged from the SNP lists for any downstream tests, and this iterative database refinement would serve to minimize false positives. Notably, the *Digitalis purpurea* sample used in this study resulted in an abnormally high number of species-specific SNPs relative to the other four *Digitalis* species screened (Table S4), which could be due to sample contamination, sequencing errors, or a high degree of evolutionary divergence in this species [55]. If this aberration is not biologically founded (i.e., if due to some systematic error), swapping that sample out for another and regenerating SNPs would also allow for the recovery of more SNPs.

While this pipeline is computational in nature, the SNPs identified in this study can just as easily be used for diagnostic primer design to develop targeted assays for the *Digitalis* genus, as well as specific species. This method does not rely on identifying markers

within well-characterized loci and thus, is considerably less restrictive. Since the number of highly problematic toxic plants commonly found in foods and dietary supplements is finite, the development of additional composite genomes and SNP databases for taxa of concern is a feasible goal. In addition to NGS methods as employed here, these databases can be used to develop suites of targeted primer sets for use with common molecular methods such as PCR, or more specialized quantitative applications like qPCR or ddPCR. The pseudoreplicated, simulated adulteration study described here is largely intended as a proof of concept, yet also provided a clear roadmap for the application of SISRS for identifying species within mixed samples. Continued development of this method, including tests on real-world adulterated samples, may lead to the rapid expansion of both NGS-based and molecular tools for faster identification of toxic plants in foods and consumer products.

## 5. Conclusions

Here, we illustrate how shallow WGS data can be used to detect low concentrations of adulterants of a specific toxic or allergenic contaminant at the genus or species level with no pre-existing reference genome or annotation data. This provides a roadmap for the rapid generation of nuclear markers in non-model groups. Our results indicate that when provided with sufficient sequencing data of background materials, ~2 Gb of data can correctly identify adulteration with simulated contamination of as little as 0.05% of *Digitalis* to both the genus and species level in mixed botanical samples using a modified application of SISRS. In the 1997 case of *Digitalis* adulteration, we believe our method would have detected *D. lanata* in both the dietary supplement and the *Plantago* raw material, as well as provide semiquantitative information regarding the amount of *D. lanata* present.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/article/10.3390/foods10081794/s1>, Table S1: Whole-genome sequence data used to assemble *Digitalis* chloroplast genomes using getOrganelle, Table S2: *Digitalis* whole-genome sequence data used in this study for composite genome assembly and SNP discovery, Table S3: SISRS assembly statistics for *Digitalis* composite genome, Table S4: Results of mapping *Digitalis* whole-genome sequencing data onto composite genome including filtering SNPs to include more *Digitalis*-specific contigs, Table S5: Non-*Digitalis* whole-genome sequence data used in this study for SNP masking, Table S6: Whole-genome sequence data used in this study for creating mixed pseudoreplicates, Table S7: Pseudoreplicated mix components and mapping statistics, Table S8: Results from linear regression of recovered SNPs versus *Digitalis* bases added, Table S9: Summary of genus-level classification results for mixed pseudoreplicates including total matching sites and proportion of sites that matched *Digitalis*, Table S10: Individual species-level classification data for each mix pseudoreplicate, Table S11: Species-level classification results for mix pseudoreplicates meeting or exceeding the minimum classification criteria, Table S12: Comparison of difference between average signal strength for top species match versus second highest species match.

**Author Contributions:** Conceptualization, E.S.H., R.L. and S.M.H.; methodology, E.S.H. and R.L.; software, R.L. formal analysis, R.L.; data curation, R.L.; writing—original draft preparation, E.S.H.; writing—review and editing, E.S.H., R.L. and S.M.H.; supervision, S.M.H.; project administration, E.S.H., R.L. and S.M.H. funding acquisition, S.M.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** Elizabeth Hunter and Robert Literman were supported, in part, by an appointment to the Research Participation Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and the U.S. FDA. This project was also supported through an FDA Chief Scientist Challenge Grant, number 200F07AC04 to Sara M. Handy.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Raw data generated for this project are available under the GenBank BioProject PRJNA325670. Specific SRA accessions for each sample can be found in the Supplementary Tables S2 and S5.

**Acknowledgments:** The authors would like to thank Ikhlas Khan and Amar Chittiboyina from the National Center for Natural Products Research at the University of Mississippi, Riley Kirk and Matthew Bertin from the College of Pharmacy at the University of Rhode Island and especially the New York Botanical Garden for providing botanical tissues. We are grateful to Lindsay Rosen from FDA-CFSAN for DNA extraction, Padmini Ramachandran and Amanda Windsor from FDA-CFSAN for sequencing and technical support, and Christopher Paight from the University of California Santa Barbara for valuable discussions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Withering, W. *An Account of the Foxglove and Some of Its Medical Uses*; Cambridge University Press: Cambridge, UK, 1785.
2. Beller, G.A.; Smith, T.W.; Abelmann, W.H.; Haber, E.; Hood, W.B.J. Digitalis Intoxication—A Prospective Clinical Study with Serum Level Correlations. *N. Engl. J. Med.* **1971**, *284*, 989–997. [[CrossRef](#)] [[PubMed](#)]
3. Lapostolle, F.; Borron, S.W. Chapter 58—Digitalis. In *Haddad and Winchester's Clinical Management of Poisoning and Drug Overdose*, 4th ed.; Shannon, M.W., Borron, S.W., Burns, M.J., Eds.; W.B. Saunders: Philadelphia, PA, USA, 2007; pp. 949–962. ISBN 978-0-7216-0693-4.
4. Haruna, Y.; Kawasaki, T.; Kikkawa, Y.; Mizuno, R.; Matoba, S. Xanthopsia Due to Digoxin Toxicity as a Cause of Traffic Accidents: A Case Report. *Am. J. Case Rep.* **2020**, *21*, e924025-1–e924025-4. [[CrossRef](#)] [[PubMed](#)]
5. Somberg, J.C. Van Gogh and Digitalis. *Am. J. Cardiol.* **2020**, *136*, 164–165. [[CrossRef](#)] [[PubMed](#)]
6. Packer, M. Why Is the Use of Digitalis Withering? Another Reason That We Need Medical Heart Failure Specialists. *Eur. J. Heart Fail.* **2018**, *20*, 851–852. [[CrossRef](#)] [[PubMed](#)]
7. Slifman, N.R.; Obermeyer, W.R.; Aloii, B.K.; Musser, S.M.; Correll, W.A.; Cichowicz, S.M.; Betz, J.M.; Love, L.A. Contamination of Botanical Dietary Supplements by *Digitalis Lanata*. *N. Engl. J. Med.* **1998**, *339*, 806–811. [[CrossRef](#)]
8. Harbaugh Reynaud, D.T.; Mishler, B.D.; Neal-Kababick, J.; Brown, P.N. The Capabilities and Limitations of DNA Barcoding of Dietary Supplements 2015. Available online: [https://www.ahpa.org/portals/0/pdfs/the-capabilities-and-limitations-of-dna-testing-final\\_ahpa.pdf](https://www.ahpa.org/portals/0/pdfs/the-capabilities-and-limitations-of-dna-testing-final_ahpa.pdf) (accessed on 17 May 2021).
9. Parveen, I.; Gafner, S.; Techen, N.; Murch, S.; Khan, I. DNA Barcoding for the Identification of Botanicals in Herbal Medicine and Dietary Supplements: Strengths and Limitations. *Planta Med.* **2016**, *82*, 1225–1235. [[CrossRef](#)]
10. Pawar, R.; Handy, S.; Cheng, R.; Shyong, N.; Grundel, E. Assessment of the Authenticity of Herbal Dietary Supplements: Comparison of Chemical and DNA Barcoding Methods. *Planta Med.* **2017**, *83*, 921–936. [[CrossRef](#)]
11. Mueller, S.; Handy, S.M.; Deeds, J.R.; George, G.O.; Broadhead, W.J.; Pugh, S.E.; Garrett, S.D. Development of a COX1 Based PCR-RFLP Method for Fish Species Identification. *Food Control* **2015**, *55*, 39–42. [[CrossRef](#)]
12. Handy, S.M.; Chizhikov, V.; Yakes, B.J.; Paul, S.Z.; Deeds, J.R.; Mossoba, M.M. Microarray Chip Development Using Infrared Imaging for the Identification of Catfish Species. *Appl. Spectrosc.* **2014**, *68*, 1365–1373. [[CrossRef](#)]
13. Rasmussen Hellberg, R.S.; Naam, A.M.; Handy, S.M.; Hanner, R.H.; Deeds, J.R.; Yancy, H.F.; Morrissey, M.T. Interlaboratory Evaluation of a Real-Time Multiplex Polymerase Chain Reaction Method for Identification of Salmon and Trout Species in Commercial Products. *J. Agric. Food Chem.* **2011**, *59*, 876–884. [[CrossRef](#)]
14. Handy, S.M.; Timme, R.E.; Jacob, S.M.; Deeds, J.R. Development of a Locked Nucleic Acid Real-Time Polymerase Chain Reaction Assay for the Detection of *Pinus Armandii* in Mixed Species Pine Nut Samples Associated with Dysgeusia. *J. Agric. Food Chem.* **2013**, *61*, 1060–1066. [[CrossRef](#)]
15. Shokralla, S.; Hellberg, R.S.; Handy, S.M.; King, I.; Hajibabaei, M. A DNA Mini-Barcoding System for Authentication of Processed Fish Products. *Sci. Rep.* **2015**, *5*, 15894. [[CrossRef](#)]
16. Puente-Lelievre, C.; Eischeid, A.C. Development and Validation of a Duplex Real-Time PCR Assay with Locked Nucleic Acid (LNA) Probes for the Specific Detection of Allergenic Walnut in Complex Food Matrices. *Food Control* **2021**, *121*, 107644. [[CrossRef](#)]
17. Shanmughanandhan, J.; Shanmughanandhan, D.; Ragupathy, S.; Henry, T.A.; Newmaster, S.G. Quantification of *Actaea racemosa* L. (Black Cohosh) from Some of Its Potential Adulterants Using QPCR and DPCR Methods. *Sci. Rep.* **2021**, *11*, 4331. [[CrossRef](#)] [[PubMed](#)]
18. Handy, S.M.; Deeds, J.R.; Ivanova, N.V.; Hebert, P.D.N.; Hanner, R.H.; Ormos, A.; Weigt, L.A.; Moore, M.M.; Yancy, H.F. A Single-Laboratory Validated Method for the Generation of DNA Barcodes for the Identification of Fish for Regulatory Compliance. *J. AOAC Int.* **2011**, *94*, 201–210. [[CrossRef](#)] [[PubMed](#)]
19. Ivanova, N.V.; Kuzmina, M.L.; Braukmann, T.W.A.; Borisenko, A.V.; Zakharov, E.V. Authentication of Herbal Supplements Using Next-Generation Sequencing. *PLoS ONE* **2016**, *11*, e0156426. [[CrossRef](#)]
20. Handy, S.M.; Ott, B.M.; Hunter, E.S.; Zhang, S.; Erickson, D.L.; Wolle, M.M.; Conklin, S.D.; Lane, C.E. Suitability of DNA Sequencing Tools for Identifying Edible Seaweeds Sold in the United States. *J. Agric. Food Chem.* **2020**, *68*, 15516–15525. [[CrossRef](#)] [[PubMed](#)]

21. Handy, S.M.; Pawar, R.S.; Ottesen, A.R.; Ramachandran, P.; Sagi, S.; Zhang, N.; Hsu, E.; Erickson, D.L. HPLC-UV, Metabarcoding and Genome Skims of Botanical Dietary Supplements: A Case Study in Echinacea. *Planta Med.* **2021**, a-1336-1685. [[CrossRef](#)]
22. Zhang, N.; Erickson, D.L.; Ramachandran, P.; Ottesen, A.R.; Timme, R.E.; Funk, V.A.; Luo, Y.; Handy, S.M. An Analysis of Echinacea Chloroplast Genomes: Implications for Future Botanical Identification. *Sci. Rep.* **2017**, *7*, 216. [[CrossRef](#)] [[PubMed](#)]
23. Wirta, H.; Abrego, N.; Miller, K.; Roslin, T.; Vesterinen, E. DNA Traces the Origin of Honey by Identifying Plants, Bacteria and Fungi. *Sci. Rep.* **2021**, *11*, 4798. [[CrossRef](#)] [[PubMed](#)]
24. Mutebi, R.R.; Ario, A.R.; Nabatanzi, M.; Kyamwine, I.B.; Wibabara, Y.; Muwereza, P.; Eurien, D.; Kwesiga, B.; Bulage, L.; Kabwama, S.N.; et al. Large Outbreak of Jimsonweed (*Datura Stramonium*) Poisoning Due to Consumption of Contaminated Humanitarian Relief Food: Uganda, March–April 2019. **2021**; In Review.
25. Literman, R.; Ott, B.M.; Wen, J.; Grauke, L.; Schwartz, R.; Handy, S.M. Reference-Free Discovery of Millions of SNPs Permits Species and Hybrid Identification in *Carya* (Hickory). In prep.
26. Soffritti, G.; Busconi, M.; Sánchez, R.A.; Thiercelin, J.-M.; Polissiou, M.; Roldán, M.; Fernández, J.A. Genetic and Epigenetic Approaches for the Possible Detection of Adulteration and Auto-Adulteration in Saffron (*Crocus Sativus* L.) Spice. *Molecules* **2016**, *21*, 343. [[CrossRef](#)]
27. Böhme, K.; Calo-Mata, P.; Barros-Velázquez, J.; Ortea, I. Recent Applications of Omics-Based Technologies to Main Topics in Food Authentication. *TrAC Trends Anal. Chem.* **2019**, *110*, 221–232. [[CrossRef](#)]
28. Sakamoto, W.; Takami, T. Chloroplast DNA Dynamics: Copy Number, Quality Control and Degradation. *Plant Cell Physiol.* **2018**, *59*, 1120–1127. [[CrossRef](#)] [[PubMed](#)]
29. Golczyk, H.; Greiner, S.; Wanner, G.; Weihe, A.; Bock, R.; Börner, T.; Herrmann, R.G. Chloroplast DNA in Mature and Senescing Leaves: A Reappraisal. *Plant Cell* **2014**, *26*, 847–854. [[CrossRef](#)] [[PubMed](#)]
30. Dodsworth, S. Genome Skimming for Next-Generation Biodiversity Analysis. *Trends. Plant Sci.* **2015**, *20*, 525–527. [[CrossRef](#)]
31. Straub, S.C.K.; Parks, M.; Weitemier, K.; Fishbein, M.; Cronn, R.C.; Liston, A. Navigating the Tip of the Genomic Iceberg: Next-Generation Sequencing for Plant Systematics. *Am. J. Bot.* **2012**, *99*, 349–364. [[CrossRef](#)] [[PubMed](#)]
32. Van der Merwe, M.; McPherson, H.; Siow, J.; Rossetto, M. Next-Gen Phylogeography of Rainforest Trees: Exploring Landscape-Level CpDNA Variation from Whole-Genome Sequencing. *Mol. Ecol. Resour.* **2014**, *14*, 199–208. [[CrossRef](#)] [[PubMed](#)]
33. Chen, F.; Dong, W.; Zhang, J.; Guo, X.; Chen, J.; Wang, Z.; Lin, Z.; Tang, H.; Zhang, L. The Sequenced Angiosperm Genomes and Genome Databases. *Front. Plant Sci.* **2018**, *9*, 418. [[CrossRef](#)] [[PubMed](#)]
34. Breitwieser, F.P.; Lu, J.; Salzberg, S.L. A Review of Methods and Databases for Metagenomic Classification and Assembly. *Brief. Bioinform.* **2019**, *20*, 1125–1136. [[CrossRef](#)]
35. Cook, D.; Lee, S.T.; Gardner, D.R.; Molyneux, R.J.; Johnson, R.L.; Taylor, C.M. Use of Herbarium Voucher Specimens To Investigate Phytochemical Composition in Poisonous Plant Research. *J. Agric. Food Chem.* **2021**, *69*, 4037–4047. [[CrossRef](#)] [[PubMed](#)]
36. Zhang, N.; Ramachandran, P.; Wen, J.; Duke, J.; Metzman, H.; McLaughlin, W.; Ottesen, A.; Timme, R.; Handy, S. Development of a Reference Standard Library of Chloroplast Genome Sequences, GenomeTrakrCP. *Planta Med.* **2017**, *83*, 1420–1430. [[CrossRef](#)] [[PubMed](#)]
37. Schwartz, R.S.; Harkins, K.M.; Stone, A.C.; Cartwright, R.A. A Composite Genome Approach to Identify Phylogenetically Informative Data from Next-Generation Sequencing. *BMC Bioinform.* **2015**, *16*, 193. [[CrossRef](#)] [[PubMed](#)]
38. Windsor, A.M.; Ott, B.M.; Zhang, N.; Wen, J.; Hsu, E.; Handy, S.M. Full Chloroplast Genome Sequence of the Economically Important Dietary Supplement and Spice *Curcuma Longa*. *Microbiol. Resour. Announc.* **2019**, *8*, e00576-19. [[CrossRef](#)]
39. Jin, J.-J.; Yu, W.-B.; Yang, J.-B.; Song, Y.; dePamphilis, C.W.; Yi, T.-S.; Li, D.-Z. GetOrganelle: A Fast and Versatile Toolkit for Accurate de Novo Assembly of Organelle Genomes. *Genome Biol.* **2020**, *21*, 1–31. [[CrossRef](#)] [[PubMed](#)]
40. Babraham Bioinformatics—FastQC A Quality Control Tool for High Throughput Sequence Data. Available online: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed on 30 November 2019).
41. Castro, M.; Castro, S.; Loureiro, J. Genome Size Variation and Incidence of Polyploidy in Scrophulariaceae Senu Lato from the Iberian Peninsula. *AoB Plants* **2012**, *2012*. [[CrossRef](#)]
42. Boisvert, S.; Raymond, F.; Godzaridis, É.; Laviolette, F.; Corbeil, J. Ray Meta: Scalable de Novo Metagenome Assembly and Profiling. *Genome Biol.* **2012**, *13*, R122. [[CrossRef](#)]
43. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
44. Leys, C.; Ley, C.; Klein, O.; Bernard, P.; Licata, L. Detecting Outliers: Do Not Use Standard Deviation around the Mean, Use Absolute Deviation around the Median. *J. Exp. Soc. Psychol.* **2013**, *49*, 764–766. [[CrossRef](#)]
45. Wood, D.E.; Lu, J.; Langmead, B. Improved Metagenomic Analysis with Kraken 2. *Genome Biol.* **2019**, *20*, 257. [[CrossRef](#)]
46. Ounit, R.; Wanamaker, S.; Close, T.J.; Lonardi, S. CLARK: Fast and Accurate Classification of Metagenomic and Genomic Sequences Using Discriminative k-Mers. *BMC Genom.* **2015**, *16*, 236. [[CrossRef](#)]
47. Kim, D.; Song, L.; Breitwieser, F.P.; Salzberg, S.L. Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences. *Genome Res.* **2016**, *26*, 1721–1729. [[CrossRef](#)] [[PubMed](#)]
48. Madden, T. *The BLAST Sequence Analysis Tool*; National Center for Biotechnology Information: Bethesda, MD, USA, 2003.
49. von Meijenfeldt, F.A.B.; Arkhipova, K.; Cambuy, D.D.; Coutinho, F.H.; Dutilh, B.E. Robust Taxonomic Classification of Uncharted Microbial Sequences and Bins with CAT and BAT. *bioRxiv* **2019**. [[CrossRef](#)]

50. Mirdita, M.; Steinegger, M.; Breitwieser, F.; Karin, E.L. Fast and Sensitive Taxonomic Assignment to Metagenomic Contigs. *Bioinformatics* **2021**, btab184. [[CrossRef](#)]
51. Ramachandran, P.; Mammel, M.; Ottesen, A.; Pava-Ripoll, M. MitochonTrakr: A Reference Collection of High-Quality Mitochondrial Genomes for Detecting Insect Species in Food Products. *Mitochondrial DNA Part B* **2019**, *4*, 292–293. [[CrossRef](#)]
52. Kalyuzhnaya, M.G.; Lapidus, A.; Ivanova, N.; Copeland, A.C.; McHardy, A.C.; Szeto, E.; Salamov, A.; Grigoriev, I.V.; Suciú, D.; Levine, S.R.; et al. High-Resolution Metagenomics Targets Specific Functional Types in Complex Microbial Communities. *Nat. Biotechnol.* **2008**, *26*, 1029–1034. [[CrossRef](#)] [[PubMed](#)]
53. Firetti, F.; Zuntini, A.R.; Gaiarsa, J.W.; Oliveira, R.S.; Lohmann, L.G.; Sluys, M.-A.V. Complete Chloroplast Genome Sequences Contribute to Plant Species Delimitation: A Case Study of the *Anemopaegma* Species Complex. *Am. J. Bot.* **2017**, *104*, 1493–1509. [[CrossRef](#)] [[PubMed](#)]
54. Maurya, S.; Darshetkar, A.M.; Yi, D.-K.; Kim, J.; Lee, C.; Ali, M.A.; Choi, S.; Choudhary, R.K.; Kim, S.-Y. Plastome Comparison and Evolution within the Tribes of Plantaginaceae: Insights from an Asian Gypsyweed. *Saudi J. Biol. Sci.* **2020**, *27*, 3489–3498. [[CrossRef](#)] [[PubMed](#)]
55. Kreis, W. The Foxgloves (*Digitalis*) Revisited. *Planta Med.* **2017**, *83*, 962–976. [[CrossRef](#)]