SCIENTIFIC REPORTS

OPEN

An analysis of *Echinacea* chloroplast genomes: Implications for future botanical identification

Received: 30 September 2016
Accepted: 21 February 2017
Published online: 16 March 2017

Ning Zhang[1], David L. Erickson[1], Padmini Ramachandran[1], Andrea R. Ottesen[1], Ruth E. Timme[1], Vicki A. Funk[2], Yan Luo[1] & Sara M. Handy[1]

Echinacea is a common botanical used in dietary supplements, primarily to treat upper respiratory tract infections and to support immune function. There are currently thought to be nine species in the genus _Echinacea_. Due to very low molecular divergence among sister species, traditional DNA barcoding has not been successful for differentiation of _Echinacea_ species. Here, we present the use of full chloroplast genomes to distinguish between all 9 reported species. Total DNA was extracted from specimens stored at the National Museum of Natural History, Smithsonian Institution, which had been collected from the wild with species identification documented by experts in the field. We used Next Generation Sequencing (NGS) and CLC Genomics Workbench to assemble complete chloroplast genomes for all nine species. Full chloroplasts unambiguously differentiated all nine species, compared with the very few single nucleotide polymorphisms (SNPs) available with core DNA barcoding markers. SNPs for any two _Echinacea_ chloroplast genomes ranged from 181 to 910, and provided robust data for unambiguous species delimitation. Implications for DNA-based species identification assays derived from chloroplast genome sequences are discussed in light of product safety, adulteration and quality issues.

Echinacea, i.e., purple coneflower, is one of the most popular botanicals used in dietary supplements. The range of *Echinacea* spans the Atlantic drainage region of the United States and extends into south central Canada[1]. For this genus, the Southern United Stated is an important native area with two species, i.e. *E. tennesseensis* and *E. laevigata* endemic to the southeast United States. Use of *Echinacea* products has dramatically increased: sales in 2013 increased by 94.7% over those in 2012, making it the 8th most commonly sold herb in the United States[2]. By 2014, sales of *Echinacea* had increased by 79% from 2013 and it was the 3rd most commonly sold herb in the United States with the sales surpassing $50 million[3]. Although not approved as a drug by the Food and Drug Administration, *Echinacea* products are often marketed for treatment of upper respiratory infections[4, 5]; other marketed uses include immune system stimulant[6, 7], adjunct therapy for chronic candidiasis in women, and external wound healing[8]. Native Americans have been using *Echinacea* extensively to treat stomach cramps, rabies, toothaches, soremouth, throat, dyspepsia, colds, headache and snake bites[9].

The three species used most commonly in dietary supplements are *E. purpurea*, *E. angustifolia* and *E. pallida*, available as teas, capsules and tablets. Importantly, each species appears to have different pharmacological activities, depending on the particular method of preparation and on which part of a given plant is used[8]. In addition to the three species, there are six other closely-related species in the same genus, i.e., *E. sanguinea*, *E. tennessensis*, *E. paradoxa*, *E. atrorubens*, *E. laevigata*, and *E. speciosa*[10]. Ardjommand-woelkart and Bauer (2016), among others, have noted that both *E. angustifolia* (whole plant) and *E. purpurea* (dry root) have been associated with allergic reactions[11–13]. However, aside from these few instances, there are no known drug interactions or side effects[8] associated with the 9 species.

The increased use of *Echinacea* species has led to concerns about adulterated products[14]. One of the reasons is that a few *Echinacea* species are phenotypically similar so it is easy to misidentify them if not familiar with the morphological variations among them[10]. The most common adulteration of *Echinacea* is the substitution of the root of *Parthenium integrifolium* for *E. purpurea*[15]. The American Herbal Pharmacopoeia Standard of Identity includes additional adulterants for *E. purpurea*: *Helianthus* spp., *Lespedeza capita*, *Eryngium aquaticum*,

[1]Center for Food Safety and Applied Nutrition, Office of Regulatory Science, US Food and Drug Administration, College Park, Maryland, 20740, United States. [2]Department of Botany, National Museum of Natural History, MRC-166, Smithsonian Institution, PO Box 37012, Washington, DC, 20013-0166, USA. Correspondence and requests for materials should be addressed to N.Z. (email: ning.zhang@fda.hhs.gov)

and *Rudbekia nitida* (http://www.herbal-ahp.org/documents/macroscopy/Ech_purpurea_macro.pdf, accessed 09/13/16). Even when *Echinacea* species are being used in products, it is not easy to differentiate among the three most appropriate *Echinacea* species, i.e., *E. purpurea*, *E. angustifolia*, and *E. pallida*; as a result, mislabeling occurs frequently[15, 16]. Given that different species may enact different effects, such adulteration could decrease the safety, efficacy and reliability of commercial *Echinacea* products.

Distinguishing among *Echinacea* species using molecular methods is challenging due to extremely low levels of molecular divergence. This reflects a pattern seen among other members of Asteraceae, which demonstrate substantial morphological variation, but very little molecular differentiation, due to recent and rapid species radiations[17, 18]. Flagel *et al.*[19] used three nuclear markers (*Adh*, *CesA*, and *GPAT*) and two plastid loci (*trnS* and *trnG*) to examine the phylogeny of *Echinacea*; however, no resolved topologies were obtained, suggesting incomplete lineage sorting, as well as the potential for widespread hybridization within the genus[19].

DNA barcoding has been an effective tool for rapidly and accurately identifying many plant species[20–22]. Mitochondrial cytochrome c oxidase (*CO1*) has been successfully used as a barcode for animal species[23]; however, no single universal barcode has been entirely successful for distinguishing all plants to the species level[24]. In 2009, the Plant Working Group of the Consortium for the Barcode of Life (CBOL) proposed a 2-locus combination of *matK* + *rbcL* as a universal plant barcode; however, this approach only provides a discriminatory efficiency of 72%[20]. Many studies have shown that core DNA markers, i.e., *matK* and *rbcL*, cannot resolve closely-related species. For example, the commercially and medicinally important species of turmeric (*Curcuma longa*, Zingiberaceae) cannot be separated from almost a hundred other *Curcuma* species using *matK* and *rbcL*[25]. A similar phenomenon was recently described for Venus slippers (*Paphiopedilum spp*.), where DNA barcodes were only successful 18.86% of the time for this popular family of orchids[26]. A study on DNA differentiation of pine nut samples conducted in our lab also indicated that the core barcoding markers were not effective for this group, so *ycf1* was developed for species level identification[27].

Subsequently, two alternative strategies were proposed to discriminate among plant species: the first was the use of multiple loci[28–30], and the second was the use of whole-chloroplast genomes, termed 'super-barcoding'[31–34]. CBOL demonstrated that the use of seven plastid DNA barcoding markers only improved species discrimination from 72% to 73% when compared with the use of two core markers[20]. The idea of using whole chloroplast genomes to identify plant species was first proposed by Kane and Cronk (2008) and has been highlighted by a few recent review articles[22]. Using complete chloroplast genomes holds promise for efficient differentiation of species compared to a multi-locus approach, especially for closely related species such as *Echinacea*.

Advances in next-generation sequencing platforms have reduced the obstacles of time, effort, and cost, necessary to acquire whole chloroplast genomes. With earlier methods, chloroplast DNA had to be enriched, a time-consuming task requiring substantial fresh leaf tissue[35]. Approaches using polymerase chain reaction (PCR) enrichment, such as long PCR[36] (using 27 primers) or multiple overlapping short-range PCR[37] (using 138 primers), have been used, but these procedures are time-consuming and labor-intensive, and the primers used in such assays do not work equally well across different taxonomic groups. Nonetheless, complete chloroplast genomes have been shown to be highly effective for resolving relationships among species with low molecular divergence[32, 33, 38, 39], and have been successfully employed for species identification[34]. Use of comparative chloroplast genomics has also been useful to identify divergent regions that can be employed for species-specific PCR-based diagnostics. For example, in 2013 Handy *et al.* used a large chloroplast dataset to design a species-specific assay to differentiate *Pinus armandii*, which causes a taste disturbance known as dysgeusia[40], from other species that do not.

Although direct sequencing of genomic DNA is still costly, quickly advancing Next Generation Sequencing (NGS) technologies may ultimately prove to be more cost effective and technically efficient than other (often more time consuming) approaches to full chloroplast sequencing. For example, using the Illumina Miseq and Hiseq (Illumina, San Diego), 2 × 300 and 2 × 250 bp reads (respectively) can be obtained with rapid throughput kits (~27 hours) yielding as much as 12 to15 Gb from a MiSeq and as much as 60 to 120 Gb from a Hiseq. It was estimated that less than 1 GB of whole-DNA short reads can be effectively assembled into a full chloroplast genome with 51x coverage[41]. Therefore, this approach alleviates the need for expensive enrichment methods and fully leverages advances in DNA sequencing and bioinformatics.

In this study we extracted DNA from dried herbarium tissue samples for all 9 *Echinacea* species, sequenced each using the Illumina MiSeq platform, and here present complete chloroplast genomes for each species. Additionally, we highlight how variation within chloroplast regions can be utilized to develop rapid species-specific assays.

Results

The data gathered for each species ranged from 434 MB for *E. tennessensis* to 2,531 MB for *E. purpurea*, with coverage of chloroplast genomes ranging between 20x for *E. tennessensis* and 65x for *E. angustifolia*. Additional information, including GenBank accession numbers, is available in Table 1.

The chloroplast genome of each *Echinacea* species appears to be collinear with the one of *Parthenium argentatum*, the most closely related public cpDNA genome, except for two inversions. These two inversions are specific to *P. argentatum* when compared with the other three Asteraceae species, i.e., *E. purpurea*, *Helianthus annuus*, and *Chrysanthemum indicum* (Figure S1). The first inversion is 891 bp long, located between *trnS* and *psbM*, and the second is 886 bp long, located between *psbM* and *rpoB*, these regions can be used for differentiating *P. argentatum* using PCR. In addition, positions of these two inversions in *Echinacea* species exchange with each other (Figure S1). Based on our alignments, no structural variations were detected among the nine *Echinacea* chloroplast genomes, so *E. purpurea* was used as an example to demonstrate the structure of *Echinacea* spp chloroplasts (Fig. 1).

The length of the chloroplast genome of *E. purpurea* is 151,913 bp. There are two inverted repeats (IRs) of 25,070 bp each, separated by a large single-copy and small single-copy (LSC and SSC) region of 83,602 bp and

Species	Raw data size (MB)	Number of reads	Size of reads (bp)	Coverage of chloroplaste genome	Size of chloroplast genome (bp)	Accession number
E. purpurea	2,531	10,394,828	2 × 300	40	151,913	KX548224
E. sanguinea	2,437	10,966,208	2 × 250	51	151,926	KX548225
E. tennessensis	434	1,814,356	2 × 250	20	151,877	KX548223
E. pallida	832	4,078,614	2 × 250	33	151,883	KX548218
E. paradoxa	1,692	6,202,480	2 × 300	51	151,837	KX548217
E. atrorubens	472	1,923,846	2 × 250	31	151,912	KX548220
E. laevigata	545	2,198,622	2 × 250	28	151,886	KX548219
E. angustifolia	878	3,338,742	2 × 300	65	151,935	KX548221
E. speciosa	483	1,941,430	2 × 250	22	151,860	KX548222

Table 1. The nine species sampled in this study and information on the chloroplast genome assembly.

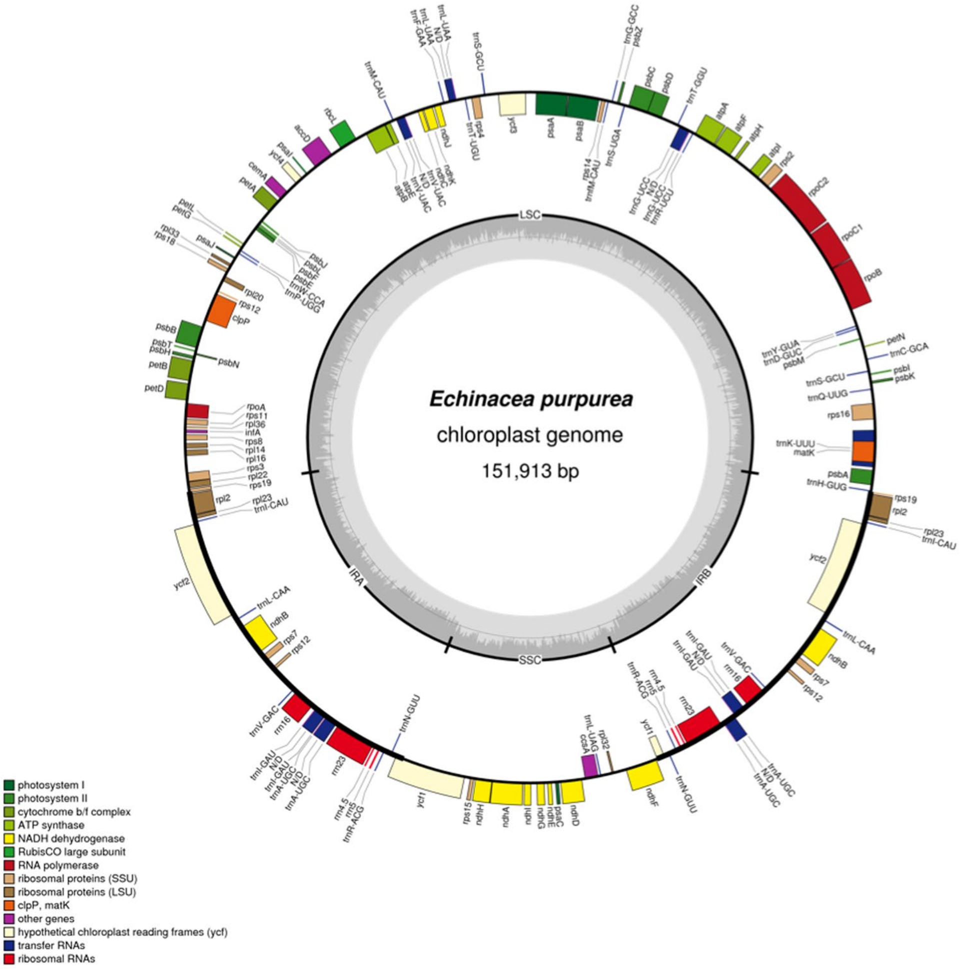

Figure 1. Gene map of the *Echinacea purpurea* chloroplast genome. Genes shown outside the circle are transcribed clockwise and those inside are transcribed counterclockwise. Gene belonging to different functional groups are color-coded as indicated by icons on the lower left corner. Dashed area in the inner circle indicates the GC content of the chloroplast genome. LSC, SSC and IR means large single copy, small single copy and inverted repeat, respectively.

18,171 bp, respectively. The G + C content of *E. purpurea* is 37.6% across the whole chloroplast genome. In total, there are 131 genes with 81 unique protein-coding genes, six of which are duplicated in the IR (Fig. 1). There are 18 unique genes with introns, five of which are duplicated in the IR; two genes have two introns and 16 genes have only one intron. There are 36 tRNA genes, 29 of which are unique and seven of which are duplicated in the

	paradox	*atrorubens*	*sanguinea*	*pallida*	*angustifolia*	*tennesseensis*	*laevigata*	*speciosa*	*purpurea*
paradox		0.12%	0.23%	0.18%	0.44%	0.52%	0.51%	0.50%	0.56%
atrorubens	181		0.20%	0.18%	0.48%	0.55%	0.55%	0.55%	0.60%
sanguinea	345	308		0.16%	0.45%	0.54%	0.53%	0.54%	0.60%
pallida	273	276	247		0.41%	0.50%	0.50%	0.50%	0.55%
angustifolia	672	727	685	629		0.47%	0.45%	0.45%	0.53%
tennesseensis	787	837	827	765	711		0.29%	0.20%	0.31%
laevigata	772	835	813	764	677	445		0.24%	0.31%
speciosa	768	830	827	767	689	309	365		0.23%
purpurea	849	910	908	842	811	469	478	350	

Table 2. Number and percentage of differences among nine *Echinacea* chloroplast genomes.

Genes	Length	Variable sites	Indels	Percentage of identical sites (%)	Timme *et al.*[42]
ycf1	5,049	31	4	99.0	√
rps8	405	3	0	99.3	
rpoA	1,009	4	1	99.3	
rpoB	3,198	7	1	99.3	
petD	483	3	0	99.4	
matK	1,282	6	0	99.4	√
rbcL	1458	7	0	99.5	
ndhF	2,232	11	0	99.5	√
ndhI	501	3	0	99.6	
psbE	252	1	0	99.6	

Table 3. The 10 most-divergent coding regions among nine *Echinacea* species.

IR. There are four unique ribosomal DNA and all of them are duplicated in the IR so there are eight ribosomal DNA in total.

As shown in Table 2, the number of base differences among these nine *Echinacea* species ranges from 181 (0.12%, *E. paradox* vs. *E. atrorubens*) to 910 (0.60%, *E. atrorubens* vs. *E. purpurea*). The number of differences between protein-coding genes is very low: 42 of 81 gene alignments are identical and the most divergent gene is *ycf1*, which has 31 variable sites and 4 indels within the 5059-bp alignment (Table 3). Table 4 lists the twenty-five most variable non-coding regions based on percentage of sequence identities. Eleven of these twenty-five overlap with those identified by Timme *et al.*[42] and three overlap with the ten plastid markers proposed by Shaw *et al.*[43] for low-level phylogenetic inferences[43] (Table 4).

We used both coding and non-coding regions of the chloroplast genomes to effectively separate all *Echinacea* species and infer a phylogeny (Fig. 2). The nine *Echinacea* species separated into two clades with strong support. One clade is comprised of *E. tennesseensis*, *E. speciosa*, *E. purpurea* and *E. laevigata*. *E. tennesseensis* appears to be closely related to *E. speciosa* with a bootstrap value of 63%; and together they are both sister to *E. purpurea* with a bootstrap value of 100%. While *E. laevigata* is closely related to the other three species, i.e., *E. tennesseensis*, *E. speciosa*, and *E. purpurea*. The second clade is comprised of five species and is well-supported with a bootstrap value of 100%. *E. angustifolia* is closely related to the other four species, forming a clade with a bootstrap value of 100%. *E. atrorubens* is sister to *E. paradox* with a bootstrap value of 100%, and *E. pallida* is sister to *E. sanguinea* with a bootstrap value of 57%.

In contrast, using the core barcoding region *matK*, we only identified 5 variable sites and 0 variable sites for *rbcL* within the 943-bp and 599-bp alignments, respectively. Even using both markers, no variations between *E. purpurea* and *E. tennesseensis* or between *E. paradox* and *E. atrorubens* could be identified. As a result, the tree constructed using the two core DNA barcoding markers (*matK* and *rbcL*) provided no resolution at most nodes (Fig. 3). *E. pallida*, *E. sanguinea*, *E. paradox*, and *E. atrorubens* formed a clade with a bootstrap value of 100%, which is congruent with the one reconstructed using chloroplast genomes. *Echinacea paradox* is sister to *E. atrorubens* with a 100% bootstrap value. However, the positions of *E. pallida* and *E. sanguinea* were unresolved and the positions of the other five species could not be resolved using *matK* and *rbcL*. Therefore, these two core DNA markers are too conserved to use in diagnostic identification questions.

Examination of the 727-bp alignment of ITS regions yielded only 7 variable sites. Additionally, no variation was observed among the three species: *E. atrorubens*, *E. purpurea*, and *E. angustifolia*. Thus, differentiation of *Echinacea* species using the ITS region was not robust. In the tree reconstructed using ITS, only 2 bootstrap values of 8 nodes were higher than 50% (Fig. 4a). *E. paradox*, *E. sanguinea*, and *E. speciosa* are highly supported as one clade with a 81% bootstrap value; *E. angustifolia*, *E. purpurea*, *E. atrorubens*, *E. laevigata*, and *E. pallida* group into one clade with a bootstrap value of 58%. Interestingly, the topology reconstructed using ITS is substantially different from the one obtained using chloroplast genomes (Fig. 3).

Genes	Length (bp)	Variable sites	Indels	Percentage of identical sites (%)	Timme *et al.*[42]	Shaw *et al.*[43]
ccsA → trnL-UAG	138	2	3	81.9		
psbI → trnS-GCU	144	4	5	86.8	√	
5 S rRNA → trnR-ACG	312	0	2	86.9		
atpF → atpA	72	0	2	88.9		
rpl32 → ndhF	904	4	7	89.9	√	√
trnT-UGU → trnL-UAA	603	5	8	90.9	√	
petN → psbM	539	3	4	90.9	√	
rps4 → trnT-UGU	392	3	3	91.6		
petD → rpoA	205	3	3	91.7		
ndhI → ndhG	388	3	1	92.5	√	
trnT-GGU → psbD	1270	11	8	92.9		√
ndhD → ccsA	234	2	4	93.2	√	
trnH-GUG → psbA	385	8	4	93.2	√	
trnK-UUU → matK	304	1	3	93.4	√	
psbC → trnS-UGA	246	1	3	93.6		
ndhC → trnV-UAC	998	9	7	93.9	√	√
ycf3 → trnS-GCU	910	8	4	94.0	√	
trnK-UUU → rps16	783	2	5	94.1		
trnR-UCU → trnG-UCC	221	5	2	94.6	√	
rps8 → rpl14	203	1	3	94.6		
psaA → ycf3	747	6	5	94.9		
psaI → ycf4	396	0	2	94.9		
rpoC2 → rps2	259	0	2	95.0		
rbcL → accD	580	3	2	95.0		
rps2 → atpI	233	1	1	95.3		

Table 4. The 25 most-divergent non-coding regions among nine *Echinacea* species.

The alignment of the nine *Echinacea* chloroplast genomes suggests that the intergenic region between *trnH* and *psbA* may be an appropriate gene for DNA barcoding for the majority of *Echinacea* species - especially if used in combination with ITS. However, differentiation relies upon very few SNPs so validation using a greater number of authenticated individuals would be needed. The size of the *trnH-psbA* PCR product ranges from 499 (*E. purpurea*) to 511 bp (*E. laevigata*) and the number of SNPs between any two species ranges from 0 (*E. atrorubens* vs *E. paradox* and *E. speciosa* vs *E. tennesseensis*) to 16 (*E. laevigata* vs E. *purpurea*) (Table S1). According to the chloroplast alignment, universal primers for *trnH-psbA* (trnHf_05[44]/psbA3_f[45]) should successfully amplify all 9 *Echinacea* species. In addition, the alignment indicates that pairs of species that cannot be differentiated using *trnH-psbA* alone, such as (*E. atrorubens* and *E. paradox*a) and (*E. speciosa* vs *E. tennesseensis*) could in theory be differentiated with the addition of the *ITS* marker. However, even with both markers, the number of diagnostic SNPs ranges from only 1 (*E. speciosa* vs *E. tennesseensis*) to 18 (*E. purpurea* vs *E. laevigata*) (Table S2) and bootstrap values for the tree constructed with *trnH* and *psbA and ITS* are extremely low (Fig. 4b).

Discussion

We successfully used direct sequencing of genomic DNA to recover complete chloroplast genomes from all nine reported *Echinacea* species and demonstrated that full chloroplast genomes can effectively differentiate all nine species. In addition to clarifying relationships among species, chloroplast genomes provide valuable data for improved DNA-based identification assays. This is especially true for closely related species, such as *Echinacea* that cannot be currently identified using most core DNA barcoding markers.

Conclusive documentation of indels could identify regions for use with PCR based screening diagnostics[46]. For example if a region that distinguishes important species based on the size of DNA fragments can be identified and validated, this method could be used without sequencing, thus creating a rapid low cost approach to species identification. In the absence of suitable indels, other variable regions in closely related species can be targeted for either PCR, real-time PCR or other sequence based identification methods[40].

There are currently 916 chloroplast genomes of land plants available in GenBank, among them, 456 (49.8%) were sequenced since 2015. With the advancement of NGS technologies and bioinformatics tools, obtaining chloroplast genomes has become quick and relatively inexpensive. Some methods developed for metagenomics, like kSNP[47], Kraken[48] and Pathoscope[49], can be used to identify species using whole-genome sequencing data in conjunction with genome scale references. We are currently investigating these options, and they will be the focus of a future manuscript.

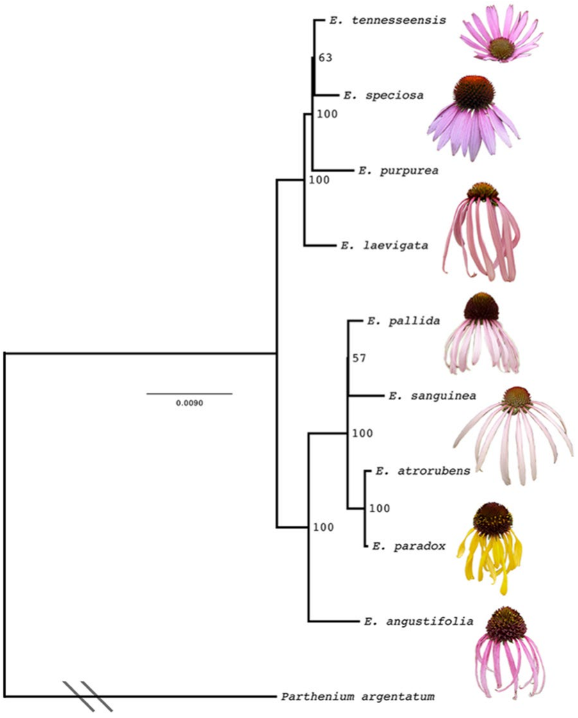

Figure 2. The ML tree of *Echinacea* reconstructed using chloroplast genomes. Numbers on branch nodes are bootstrap values. The branch connecting the outgroup *Parthenium argentatum* and nine *Echinacea* species was collapsed.

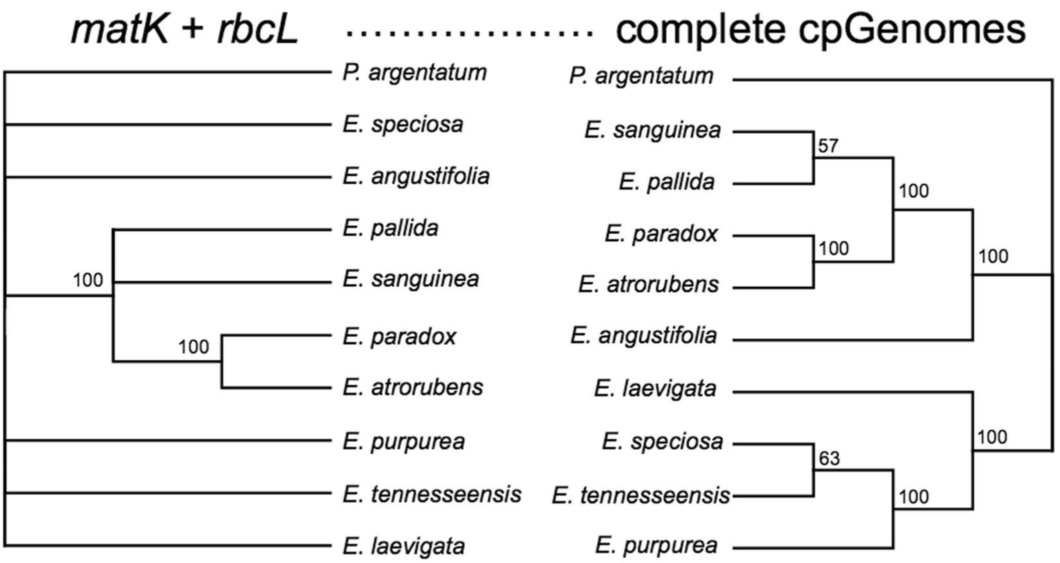

Figure 3. ML trees reconstructed using *matK* + *rbcL* (left) and using chloroplast genomes (right) Numbers are bootstrap values, branches with bootstrap values <50% are collapsed. These two phylogenies show the power of chloroplast genomes for delimitation of *Echinacea* species when compared with core DNA barcodes.

The data generated for this *Echinacea* inquiry will become part of the U.S. Food and Drug Administration's library of chloroplast genomes, the details of which will be discussed in a future publication. Future studies will explore the most useful and efficient way to identify *Echinacea* species using either whole chloroplast genomes or targeted assays developed from the full chloroplast genomes.

Methods

Sampling. We sampled all nine *Echinacea* species available from the U.S. National Herbarium. Voucher information can be found in Table 1 and Table 5.

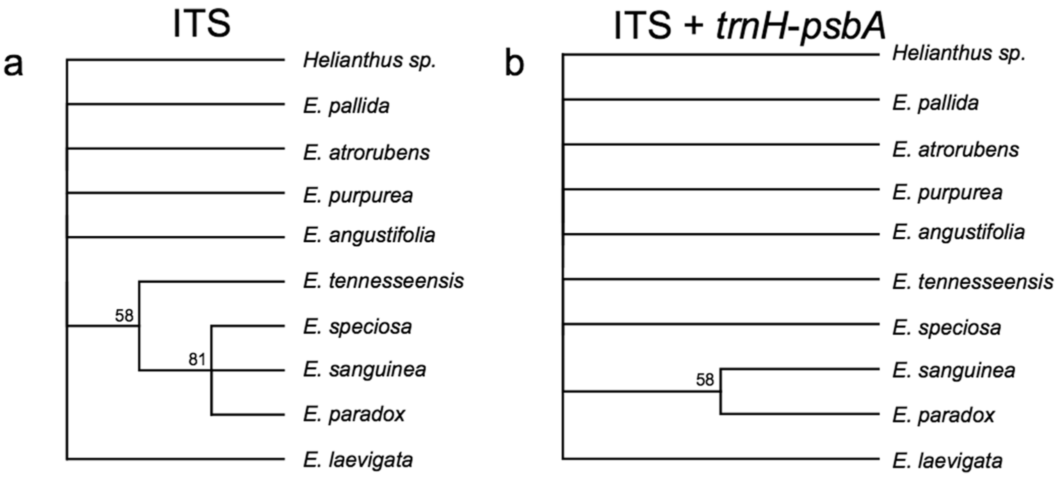

Figure 4. ML trees reconstructed using ITS (**a**) and ITS + *trnH-psbA* (**b**). Numbers are bootstrap values, branches with the bootstrap value <50% are collapsed. Both phylogenies show the lack of resolution among *Echinacea* species using either combination of genes.

Species	Voucher	Year collected
E. purpurea	US 2349097	1958
E. sanguinea	US 1468035	1930
E. tennessensis	US 980416	1916
E. pallida	US 2233063	1948
E. paradoxa	US 1653013	1935
E. atrorubens	US 2235164	1955
E. laevigata	US 3360860	1998
E. angustifolia	US 2802433	1974
E. speciosa	US 2349080	1960

Table 5. Sampling in this *Echinacea* study.

DNA isolation, and sequencing.

Total DNA was extracted from the dry leaves of specimens using the DNeasy Plant Mini Kit (part #69106, Qiagen, Valencia, CA,). For the library construction, 200 ng DNA was taken and sheared into ~550 bp contigs with the Covaris M220 Focused-ultrasonicator. The library was constructed using either the TruSeq DNA HT Sample Prep Kit (Illumina, FC-121-3003) or the TruSeq Nano DNA NeoPrep Kit (Illumina, NP-101-1001). Sequencing was run on the Illumina MiSeq Sequencer with MiSeq Reagent Kit v2 (MS-102-2001) or MiSeq Reagent Kit v3 (MS-102-3001) to obtain 2 × 250 or 2 × 300 reads, respectively.

Genome assembly and annotation.

Before assembly, the reads were trimmed using the Qiagen CLC Genomics Workbench v.8.5.1 (hereafter called CLC) with default settings. Then the trimmed sequences were assembled into contigs using *de novo* assembly, implemented in CLC. In addition, a reference-guided assembly was performed using CLC with the published chloroplast genome of the closest available relative, *Parthenium argentatum* (NC_013553), as the reference genome. After finishing reference-guided assembly, a consensus sequence of *Echinacea* was obtained. Both the consensus sequence from the reference-guided assembly and the contigs from the *de novo* assembly were imported into Geneious Pro 9.0.4, and then those contigs of chloroplast were mapped onto the consensus sequence. The mapped contigs were checked and adjusted manually to align with the consensus sequence obtained using referenced-guided assembly[39]. The final sequence of *Echinacea* chloroplast genome is the ordered sequence of those mapped contigs. We annotated the chloroplast genome using Geneious with the chloroplast genome of *Helianthus annuus* (NC_007977) as the reference since the annotation of *H. annuus* is known to be accurate[42, 50]. All sequence data has been deposited in Genbank (Accession numbers KX548217- KX548225, Table 1).

Retrieving gene sequences of widely-used DNA barcoding markers.

In order to test if core DNA barcode markers can be used for identification here, we obtained gene sequences of *matK*, *rbcL*, and ITS (internal transcribed spacer) for *Echinacea* species and for their closely-related species. In order to be effective, these needed to have variable bases in each of the nine species being investigated.

Based on the alignment of *P. argentatum* with nine *Echinacea* chloroplast genomes, we extracted two core plastid DNA barcoding markers *matK* and *rbcL*. These markers used for DNA barcoding were delimitated by corresponding primers, rbcLa-F (ATGTCACCACAAACAGAGACTAAAGC)[51]/rbcLa-R

(GTAAAATCAAGTCCACCRCG)[28] for *rbcL*, matK-xf (TAATTTACGATCAATTCATTC)[52]/matK-MALP (ACAAGAAAGTCGAAGTAT)[53] for *matK*.

We also obtained the gene sequences of ITS, another commonly used marker, from each *Echinacea* species. To obtain the ITS sequence for each species, the contig containing the ITS was obtained. The contigs of each species obtained using *de novo* assembly mentioned above were built into a BLAST database on the local server, then the ITS sequence of *Echinacea pallida* (EU785938) was used as the seed to search against the database. Usually, the best-hit contig contains the sequence of ETS, 18S, ITS1, 5.8S, ITS2, and 26S. Then we delimitated the region of ITS using the corresponding primers, i.e., ITS1 (TCCGTAGGTGAACCTGCGG)[54]/ITS4 (TCCTCCGCTTATTGATATGC)[54]. Since the ITS sequence of *P. argentatum* is not available, *H. annuus* (JX867644) was used as the outgroup.

Phylogenetic analysis. Whole chloroplast genomes of nine *Echinacea* species and the one of *Parthenium* were aligned using MAFFT v7[55]. As the sequences of IRa and IRb are almost identical, only one of them was included in the phylogenetic analyses. In addition, the sequences of tRNAs and rDNAs of nine *Echinacea* species are almost identical, so those genes were removed for all samples from the alignment. In order to reduce phylogenetic noise, three inverted intergenic regions of *Parthenium* were deleted from the alignment. The program PartitionFinder[56] was used for identifying partitions used in developing model parameters for phylogeny estimation. A maximum likelihood (ML) tree was inferred with RAxML v8.1[57] using the model of GTRGAMMAI, and 1,000 rapid bootstrap replications were performed. The sequences of *matK* + *rbcL* and ITS were aligned with MAFFT v7, then the ML trees were reconstructed using RAxML with the GTRGAMMAI model, and 1,000 rapid bootstrap replications were performed. Since this study mainly focuses on species delimitation rather than phylogeny, these genes were not concatenated for further phylogenetic analyses. These alignments were deposited into the DRYAD with the accession number of XXXX.

References

1. McGregor, R. L. The taxonomy of the genus *Echinacea* (Compositae). *Univ. Kansas Sci. Bul* **48**, 113–142 (1968).
2. Lindstrom, A., Ooyen, C., Lynch, M., Blumenthal, M. & Kawa, K. Sales of herbal dietary supplements increase by 7.9% in 2013, marking a decade of rising sales: turmeric supplements climb to top ranking in natural channel. *HerbalGram* **103**, 52–56 (2014).
3. Smith, T. *et al.* Herbal dietary supplement sales in US increase 6.8% in 2014. *HerbalGram* **107**, 52–59 (2015).
4. Taylor, J. A. *et al.* Efficacy and safety of *Echinacea* in treating upper respiratory tract infections in children: a randomized controlled trial. *Jama* **290**, 2824–2830 (2003).
5. Tierra, M. *Echinacea*: an effective alternative to antibiotics. *J. Herb. Pharmacother* **7**, 79–89 (2008).
6. Speroni, E., Govoni, P., Guizzardi, S., Renzulli, C. & Guerra, M. Anti-inflammatory and cicatrizing activity of *Echinacea pallida* Nutt. root extract. *J. Ethnopharmacol.* **79**, 265–272 (2002).
7. Yu, H.-C. & Kaarlas, M. In *Echinacea: The genus Echinacea* (ed. Yu, H. C., Miller, S. C.) 127–150 (CRC Press, 2004).
8. Barrett, B. Medicinal properties of *Echinacea*: a critical review. *Phytomedicine* **10**, 66–86 (2003).
9. Kindscher, K. Ethnobotany of purple coneflower (*Echinacea angustifolia*, Asteraceae) and other *Echinacea* species. *Econ. Bot.* **43**, 498–507 (1989).
10. McKeown, K. A. A review of the taxonomy of the genus *Echinacea* In *Perspectives on new crops and new uses* (ed. Janick, J.) 482–489 (ASHS press, 1999).
11. Mullins, R. J. *Echinacea*-associated anaphylaxis. *Med. J. Aust.* **168**, 170–171 (1998).
12. Ernst, E. The risk–benefit profile of commonly used herbal therapies: ginkgo, St. John's wort, ginseng, echinacea, saw palmetto, and kava. *Ann. Intern. Med.* **136**, 42–53 (2002).
13. Ardjomand-Woelkart, K. & Bauer, R. Review and assessment of medicinal safety data of orally used *Echinacea* preparations. *Planta Med.* **82**, 17–31 (2016).
14. Laasonen, M. *et al.* Fast identification of *Echinacea purpurea* dried roots using near-infrared spectroscopy. *Anal. Chem.* **74**, 2493–2499 (2002).
15. Bauer, R. & Wagner, H. Neue Ergebnisse zur Analytik von Echinacea—Wurzeln. *Sci. Pharm* **55**, 159–161 (1987).
16. Bauer, R., Remiger, P. & Wagner, H. Alkamides from the roots of *Echinacea angustifolia*. *Phytochemistry* **28**, 505–508 (1989).
17. Wagstaff, S. J. & Breitwieser, I. Phylogeny and classification of *Brachyglottis* (Senecioneae, Asteraceae): an example of a rapid species radiation in New Zealand. *Syst. Bot.* **29**, 1003–1010 (2004).
18. Baldwin, B. G., Kyhos, D. W., Dvorak, J. & Carr, G. D. Chloroplast DNA evidence for a North American origin of the Hawaiian silversword alliance (Asteraceae). *Proc. Natl. Acad. Sci. USA* **88**, 1840–1843 (1991).
19. Flagel, L. E. *et al.* Phylogenetic, morphological, and chemotaxonomic incongruence in the North American endemic genus. *Echinacea. Am. J. Bot.* **95**, 756–765 (2008).
20. Group, C. P. W. *et al.* A DNA barcode for land plants. *Proc. Natl. Acad. Sci. USA* **106**, 12794–12797 (2009).
21. Kress, W. J. *et al.* Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proc. Natl. Acad. Sci. USA* **106**, 18621–18626 (2009).
22. Li, X. *et al.* Plant DNA barcoding: from gene to genome. *Biol. Rev. Camb. Philos. Soc.* **90**, 157–166 (2015).
23. Hebert, P. D., Cywinska, A. & Ball, S. L. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B. Biol. Sci* **270**, 313–321 (2003).
24. Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. & Janzen, D. H. Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. USA* **102**, 8369–8374 (2005).
25. Chen, J., Zhao, J., Erickson, D. L., Xia, N. & Kress, W. J. Testing DNA barcodes in closely related species of *Curcuma* (Zingiberaceae) from Myanmar and China. *Mol. Ecol. Resour.* **15**, 337–348 (2015).
26. Guo, Y.-Y., Huang, L.-Q., Liu, Z.-J. & Wang, X.-Q. Promise and challenge of DNA barcoding in Venus slipper (*Paphiopedilum*). *PLoS One* **11**, e0146880 (2016).
27. Handy, S. M. *et al.* Use of the chloroplast gene *ycf1* for the genetic differentiation of pine nuts obtained from consumers experiencing dysgeusia. *J. Agric. Food Chem.* **59**, 10995–11002 (2011).
28. Kress, W. J. & Erickson, D. L. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS One* **2**, e508 (2007).
29. Erickson, D. L., Spouge, J., Resch, A., Weigt, L. A. & Kress, W. J. DNA barcoding in land plants: developing standards to quantify and maximize success. *Taxon* **57**, 1304 (2008).
30. Kane, N. C. & Cronk, Q. Botany without borders: barcoding in focus. *Mol. Ecol* **17**, 5175–5176 (2008).
31. Sucher, N. J. & Carles, M. C. Genome-based approaches to the authentication of medicinal plants. *Planta Med.* **74**, 603 (2008).
32. Parks, M., Cronn, R. & Liston, A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* **7**, 1 (2009).

33. Yang, J.-B., Tang, M., Li, H.-T., Zhang, Z.-R. & Li, D.-Z. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evol. Biol.* **13**, 84 (2013).
34. Nock, C. J. *et al.* Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol. J* **9**, 328–333 (2011).
35. Jansen, R. K. *et al.* Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol* **395**, 348–384 (2005).
36. Wu, C.-S., Lai, Y.-T., Lin, C.-P., Wang, Y.-N. & Chaw, S.-M. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. *Mol. Phylogenet. Evol.* **52**, 115–124 (2009).
37. Dong, W., Xu, C., Cheng, T., Lin, K. & Zhou, S. Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of Saxifragales. *Genome Biol. Evol* **5**, 989–997 (2013).
38. Jansen, R. K. *et al.* Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. USA* **104**, 19369–19374 (2007).
39. Zhang, N., Wen, J. & Zimmer, E. A. Congruent deep relationships in the grape family (Vitaceae) based on sequences of chloroplast genomes and mitochondrial genes via genome skimming. *PLoS One* **10**, e0144701 (2015).
40. 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.* **61**, 1060–1066 (2013).
41. McPherson, H. *et al.* Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecol.* **13**, 8 (2013).
42. Timme, R. E., Simpson, B. B. & Linder, C. R. High-resolution phylogeny for *Helianthus* (Asteraceae) using the 18S–26S ribosomal DNA external transcribed spacer. *Am. J. Bot.* **94**, 1837–1852 (2007).
43. Shaw, J. *et al.* Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. *Am. J. Bot.* **101**, 1987–2004 (2014).
44. Tate, J. A. & Simpson, B. B. Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. *Syst. Bot.* **28**, 723–737 (2003).
45. Sang, T., Crawford, D. & Stuessy, T. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Am. J. Bot.* **84**, 1120–1120 (1997).
46. Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl. Acad. Sci. USA* **101**, 14812–14817 (2004).
47. Gardner, S. N., Slezak, T. & Hall, B. G. kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* btv271 (2015).
48. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
49. Hong, C. *et al.* PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2**, 33 (2014).
50. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
51. Levin, R. A. *et al.* Family-level relationships of Onagraceae based on chloroplast *rbcL* and *ndhF* data. *Am. J. Bot.* **90**, 107–115 (2003).
52. Ford, C. S. *et al.* Selection of candidate coding DNA barcoding regions for use on land plants. *Bot. J. Linn. Soc* **159**, 1–11 (2009).
53. Dunning, L. T. & Savolainen, V. Broad-scale amplification of *matK* for DNA barcoding plants, a technical note. *Bot. J. Linn. Soc* **164**, 1–9 (2010).
54. White, T. J., Bruns, T., Lee, S. & Taylor, J. *A Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics in PCR Protocols: A Guide to Methods and Applications* (eds Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J.) 315–322 (Academic Press, 1990).
55. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
56. Lanfear, R., Calcott, B., Ho, S. Y. & Guindon, S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701 (2012).
57. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* btu033 (2014).

Acknowledgements

This study was supported by an ORISE fellowship from the Center for Food Safety and Applied Nutrition, Food and Drug Administration. We thank Dr. Lili Fox Vélez for scientific writing support.

Author Contributions

D.L.E., N.Z., A.R.O., V.F. and S.M.H. designed the research. V.F. provided voucher specimen. P.R., N.Z., D.L.E., and A.R.O. extracted DNAs and performed sequencing. N.Z., Y.L. and R.E.T. did computational analysis and deposited sequences. All drafted the manuscript. All authors have read and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-00321-6

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.